TESIS DE DOCTORADO

UNA HERRAMIENTA BASADA EN TERMINOLOGÍAS ESTANDARIZADAS PARA LA ANOTACIÓN SEMÁNTICA DE INFORMACIÓN TEXTUAL

Hadriana Rodríguez Castiñeira

ESCUELA DE DOCTORADO INTERNACIONAL DE LA UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOCTORADO EN INVESTIGACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN

SANTIAGO DE COMPOSTELA

2021

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR/A DE LA TESIS

UNA HERRAMIENTA BASADA EN TERMINOLOGÍAS ESTANDARIZADAS PARA LA ANOTACIÓN SEMÁNTICA DE INFORMACIÓN TEXTUAL

Dña. .Hadriana Rodríguez Castiñeira

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1) *La tesis abarca los resultados de la elaboración de mi trabajo.*
2) *En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.*
3) *La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.*
4) *Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.*

En Santiago de Compostela, 15 de Marzo de 2021

Fdo Hadriana Rodríguez Castiñeira

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

D./Dña. **María Jesús Taboada Iglesias y Diego Martínez Hernández**

En condición de: **Tutora/directora y codirector**

Título de la tesis: **Una herramienta basada en terminologías estandarizadas para la anotación semántica de información textual**

INFORMAN:

Que la presente tesis, se corresponde con el trabajo realizado por Dña. **HADRIANA RODRÍGUEZ CASTIÑEIRA**, bajo nuestra dirección/tutorización, y autorizamos su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director y codirector de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En Santiago de Compostela, 15 de marzo de 2021,

Fdo María Jesús Taboada Iglesias y Diego Martínez Hernández

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A mi familia

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimientos

Antes de nada, agradecer a mi directora de tesis, Chus, por su gran apoyo, dedicación, paciencia y en especial, por transmitirme esa pasión hacia el mundo de la investigación.

También me gustaría agradecer a todas aquellas personas que he conocido durante el transcurso de la tesis por compartir conmigo sus experiencias y conocimientos.

Expresar mi agradecimiento al apoyo económico recibido durante estos años, especialmente a los proyectos que han financiado esta investigación y sin los cuales no hubiese podido realizar este trabajo: el proyecto FIS2012-PI12/00373: OntoNeurophen del Instituto Nacional de Salud Carlos III.

Por último, quería agradecer especialmente a mis padres, a mi hermana Marta y Héctor por su apoyo y, sobre todo, paciencia durante estos años.

Santiago de Compostela, Marzo 2021

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

RESUMEN

Gracias al auge de las nuevas tecnologías, hoy en día es posible disponer de la información en cualquier lugar y en cualquier momento. Sin embargo, dada la inmensa cantidad de información existente, la búsqueda es todavía una tarea que puede llegar a consumir mucho tiempo. En el ámbito de las publicaciones científicas, extraer datos de interés, sin pérdida de información y en el menor tiempo posible, se ha convertido en todo un desafío. En el dominio específico de la medicina, la búsqueda de publicaciones de casos clínicos de pacientes que presentan fenotipos complejos o raros es una tarea necesaria en la rutina clínica. Dicha tarea es especialmente crítica en el caso de las enfermedades raras, donde las publicaciones son escasas, y la búsqueda entre un gran conjunto de documentos, para extraer la información de interés, se convierte en una tarea difícil y complicada. De manera análoga, la investigación en el área de la gestión de proyectos es un área en constante cambio y evolución, no solo por su avance en el campo de estudio, sino también por su carácter multidisciplinar y de expansión hacia nuevos ámbitos de aplicación. Por esta razón, el análisis de tendencias temáticas se ha convertido en una valiosa herramienta para detectar cómo se está conformando la disciplina y poder prever hacia dónde evoluciona. Por ello, los expertos en la materia demandan herramientas que sean capaces de analizar automáticamente las publicaciones científicas y extraer las principales tendencias temáticas del momento. Independientemente del ámbito de aplicación, la naturaleza no estructurada de las publicaciones científicas dificulta el acceso a la información contenida en los textos.

Con el fin de simplificar el acceso a la información textual de las publicaciones y dar solución a la problemática suscitada, este trabajo doctoral se basa en la hipótesis de que el procesamiento automático de la información contenida en las publicaciones científicas se puede mejorar aprovechando los recursos de conocimiento del dominio de aplicación. Sustentado en esta hipótesis, el objetivo de esta tesis es el

diseño e implementación de técnicas léxicas, sintácticas y semánticas que permitan aprovechar al máximo los recursos de conocimiento disponibles para la extracción de la información relevante en el dominio de interés. En primer lugar, las herramientas de anotación basadas en terminologías y ontologías proporcionan una gran capacidad para reconocer conceptos cuando la cobertura de las terminologías es amplia y sus conceptos están enriquecidos con un número suficiente de sinónimos. Para mejorar la capacidad de anotación basada en ontologías, en este trabajo de tesis doctoral se ha diseñado e implementado un método para el aprendizaje de variantes léxicas a partir de la terminología incluida en la propia ontología. El método aprende nuevos términos a partir del conocimiento lexicológico de la propia ontología. En segundo lugar, el uso de las palabras clave de autor facilitan numerosas tareas de recuperación de información y de procesamiento del lenguaje natural, como el resumen de documentos o el análisis de tendencias temáticas. Un problema importante, pero poco abordado hasta ahora, es la unificación de las palabras clave extraídas de un gran conjunto de publicaciones científicas. Para abordar esta carencia, proponemos un enfoque automatizado para la unificación de palabras clave que permita incrementar la calidad de los resultados obtenidos. La metodología propuesta consiste en pre-procesar la terminología, aplicando un conjunto de técnicas léxicas, sintácticas y semánticas, que unifiquen los términos que hacen referencia en el mismo concepto.

En la fase inicial de nuestro trabajo de investigación, se desarrolló una herramienta de anotación semántica, para facilitar la extracción y el análisis de la información contenida en las publicaciones científicas. Inicialmente, este anotador se diseñó e implementó para el ámbito de la biomedicina; en concreto, en el dominio de la enfermedad rara denominada *cerebrotendinous xanthomatosis*, y con la ontología de fenotipos *Human Phenotype Ontology* (HPO). En el estudio y recuperación de casos clínicos de enfermedades raras, se plantean dos problemáticas recurrentes. La primera consiste en recuperar todos los documentos relevantes de la temática y, la segunda, en acceder a la información relevante de los documentos obtenidos. Para ello, en primer lugar, se elaboró un conjunto de patrones lingüísticos basados

en regularidades observadas en los *abstracts* o resúmenes de los textos clínicos. Con este método se pudieron identificar 50 informes clínicos más de los que se habían identificado de forma manual, con una elevada precisión. Este resultado es crucial en el dominio de las enfermedades raras, donde el número de casos clínicos es limitado. Para acceder a la información relevante de los documentos obtenidos, se desarrolló la herramienta OBO Annotator, que aplica un algoritmo de concordancia de cadenas entre toda la terminología de la ontología, previamente indexada para mejorar su rendimiento, y los textos. Esto permite mejorar la velocidad de anotación, sin perder eficiencia, con respecto a las herramientas basadas en el procesamiento del lenguaje natural, como Metamap. Nuestro anotador se puede aplicar para reconocer términos de cualquier ontología OBO, ya que es principalmente un reconocedor de entidades, que compara el texto de entrada con los términos de cualquier ontología OBO, y no depende del conocimiento específico contenido en la ontología. Previo a la anotación, OBO Annotator construye un diccionario a partir de todos los nombres y sinónimos de los conceptos de la ontología. Durante la anotación, utiliza una ventana de rastreo de secuencias de texto para extraer, procesar y generar mediante permutaciones todas las cadenas de caracteres que se compararán con los términos léxicos pre-procesados del diccionario. Por defecto, esta ventana tiene un tamaño de longitud cuatro, esto es, que se seleccionan un máximo de hasta cuatro palabras consecutivas. Este valor por defecto se estableció empíricamente, al analizar la ontología utilizada, en la que se observó que el uso de conceptos formados por más de 5 palabras, que no eran conectores o determinantes, era infrecuente. Además, nuestra herramienta se encarga de filtrar las anotaciones redundantes, es decir, aquellas correspondientes a conceptos generales en presencia de anotaciones de conceptos más específicos. La etapa de filtrado hace uso de las relaciones jerárquicas de la ontología. Nuestros resultados mostraron que es posible, sobre las publicaciones científicas, identificar automáticamente aquellas que corresponden a informes clínicos, con una alta precisión y anotarlas con una calidad satisfactoria (*F-measure* del 74%). La anotación semántica de todas las anomalías fenotípicas encontradas en una enfermedad puede facilitar el diagnóstico temprano

de pacientes con enfermedades raras, al proporcionar los fenotipos relacionados junto con su frecuencia. Además, nuestro estudio experimental sobre anotación de casos clínicos de pacientes publicados en revistas científicas mostró que la herramienta de anotación obtuvo mejores resultados en términos de precisión y exhaustividad (*recall)* que los anotadores existentes en el momento, como el anotador del NCBO o el facilitado por GoPubMed. Posteriormente, surgió el anotador Bio-Lark CR, que en su evaluación, analizó el rendimiento de nuestro anotador. En esta comparativa también se obtuvieron buenos resultados, ya que la precisión obtenida en este caso fue superior al Bio-Lark CR. El *recall* fue ligeramente inferior, pero hay que tener en cuenta que en el momento de la evaluación, el equipo de investigación del Bio-Lark CR utilizó nuestro anotador con una versión antigua de la ontología HPO, por lo que los resultados obtenidos no eran realmente comparables.

A lo largo de los años, se han propuesto diferentes enfoques para ampliar la cobertura de terminologías biomédicas con el objetivo de proporcionar una mayor capacidad para reconocer conceptos. Bajo la hipótesis de que el enriquecimiento terminológico de una ontología mejoraría la capacidad de la anotación semántica, se propuso un nuevo método para el aprendizaje de nuevos sinónimos a partir de las propiedades lexicológicas de una ontología, y se utilizó como caso de uso la ontología HPO. Nuestro enfoque se basa tanto en las propiedades léxicas de los términos como en la estructura jerárquica de la ontología. Al identificar las diferencias léxicas entre un término y sus términos descendentes, el método aprende nuevos términos y modificadores, que permiten generar sinónimos para los términos descendentes. En primer lugar, el método identifica recursivamente todas las superposiciones léxicas en HPO, es decir, todos los pares de términos conectados por una relación jerárquica y donde el término descendiente incluye el término ascendente como sub-cadena propia. En segundo lugar, para cada término descendente en cada superposición léxica, el método genera nuevos sinónimos al reemplazar, en el término descendente, las palabras superpuestas con sinónimos conocidos del término ascendente. En tercer lugar, como la generación sintética de sinónimos puede dar lugar a sinónimos sin sentido, el método filtra estos

candidatos sin sentido, simplemente buscando las frases exactas de los candidatos en MEDLINE y descartando aquellas para las que no se ha recuperado ningún resultado (lo que indica que no son términos utilizados por la comunidad científica). Utilizando las propiedades léxicas y lógicas de la ontología que sirvió de caso de uso (HPO), nuestro método generó 5.964 sinónimos candidatos totales, de los cuales únicamente 745 fueron encontrados en MEDLINE, cuando se buscaron frases exactas. Estos 745 sinónimos cubrieron 488 conceptos únicos de HPO. La evaluación realizada de estos términos identificó una mejora en el desempeño de *F-measure* en las tareas de extracción de información. Además, los nuevos términos permitieron la recuperación de un 6% más del total de artículos de investigación sobre enfermedades hereditarias, y un 33% cuando consideramos únicamente los conceptos altamente informativos (es decir, aquellos cuyo valor del contenido de información de Resnik es más elevado), lo que constata la efectividad del método.

Por otro lado, en el dominio de la investigación en dirección de proyectos también surge la necesidad de extraer la información relevante de un gran conjunto de publicaciones para poder analizar sus tendencias temáticas. Con este fin, hemos utilizado nuestra herramienta de anotación para extraer automáticamente la información relevante y, a partir de ella, realizar un análisis de tendencias temáticas. Debido a la inexistencia de recursos estructurados (como ontologías) en el área, se construyó un diccionario que agrupaba diferentes glosarios de términos revisados y aprobados por instituciones internacionales del área de estudio. Para desarrollar la herramienta de anotación, se partió del núcleo del OBO Annotator y se añadieron dos funcionalidades principales nuevas: la descarga automática de artículos del repositorio seleccionado y la creación de la red de co-ocurrencia de términos. El nuevo producto software desarrollado se denominó PIMAnnot. Para el estudio de las tendencias temáticas, se seleccionaron las publicaciones realizadas entre 2000 y 2018 en la revista *International Journal of Project Management* (IJPM). Seleccionamos esta revista por su relevancia en el área, el alto número de publicaciones que la revista ofrece cada año, y porque era la única revista revisada por pares centrada en investigación en dirección de proyectos a la que teníamos

acceso completo a sus publicaciones. De esta forma, nuestra fuente de datos proporcionó 1.612 artículos de investigación para su análisis. Partiendo de la hipótesis de que los resúmenes de los artículos contienen toda la información relevante del documento, se utilizó esta sección para anotar. Con las anotaciones obtenidas en esta fase se elaboró la red de co-ocurrencia de términos. Esta red de conceptos se construye a partir de la base de que dos conceptos que ocurren en un mismo documento van a estar relacionados. Si esos conceptos aparecen juntos en más de un artículo, la relación tendrá un peso acorde a la frecuencia de ocurrencia de dicho par en cada uno de los documentos donde aparecen, de forma que cuanto más frecuente sea la co-ocurrencia de dos palabras clave, más fuerte será la correlación entre ellas. A partir de la red de co-ocurrencia generada y, aplicando *clustering* sobre ésta, se extraen las principales agrupaciones temáticas. Finalmente, se evalúa la calidad de cada tema mediante la técnica de coherencia temática o *topic coherence* y, mediante análisis de *burst*, se analiza la tendencia, por año, de cada tema de buena calidad obtenido. Para evaluar nuestros resultados se tomaron como referencia estudios previos en el área. Esta evaluación cualitativa permitió comprobar que las cuatro tendencias temáticas identificadas con coherencia elevada mediante nuestra metodología coincidían con los resultados de otros trabajos sobre el mismo periodo de análisis. Esta evaluación confirmó que nuestra herramienta de anotación era capaz de extraer información relevante y de interés utilizando recursos terminológicos diferentes (un diccionario en vez de una ontología), en un contexto totalmente diferente. Teniendo en cuenta estos resultados, podemos concluir que nuestro método puede ser una herramienta útil para la obtención automática de tendencias temáticas en el ámbito de estudio.

De manera análoga, existen numerosos estudios en la investigación de tendencias temáticas que se han basado en el análisis de palabras clave de autor, en vez de la anotación de los textos. Estos enfoques se centran, principalmente, en el análisis de frecuencia o en el análisis de redes de co-ocurrencia para el diseño del mapa conceptual del dominio y la posterior obtención de las tendencias temáticas. Las palabras clave son la sección más importante para numerosas tareas de recuperación de información y procesamiento del lenguaje natural, como el resumen

de documentos o el análisis de tendencias. Sin embargo, los términos clave extraídos de diferentes documentos suelen sufrir de un alto grado de heterogeneidad, ya que el mismo término se puede expresar utilizando diferentes variantes textuales. Suponiendo que las variantes de una misma clave corresponden a diferentes términos clave, la realidad de la unidad de análisis (las palabras) puede distorsionarse. Para garantizar correlaciones correctas, las palabras clave deben estar unificadas antes del análisis de co-ocurrencia. Por unificación de palabras clave, nos referimos al proceso de conciliación de la terminología dispar de términos clave en diferentes documentos. La unificación de palabras clave permite conciliar, por ejemplo, diferentes formas ortográficas y variaciones de un mismo término. Con el objetivo de aportar una solución válida a la problemática suscitada en el análisis de palabras clave de autor, en este trabajo de tesis proponemos un nuevo método de procesamiento de las palabras clave, denominado KeyUnif. Este método se fundamenta en la aplicación de un conjunto de técnicas léxicas, sintácticas y semánticas. Estas técnicas abarcan desde la identificación y aplicación de patrones léxicos hasta el uso de algoritmos más complejos de aprendizaje automático como el Word2Vector. Para probar la eficacia de nuestro procedimiento de unificación, el caso de uso seleccionado ha sido el mismo que para la anotación, es decir, el análisis de tendencias de la investigación de gestión de proyectos a partir de 1.612 artículos publicados durante un período de 19 años entre 2000 y 2018. Nuestro método de unificación de palabras clave identificó y unificó un 37% del total de palabras clave extraídas de las publicaciones, lo que redujo considerablemente el conjunto léxico de partida para nuestro análisis. Del total de términos unificados, las técnicas léxicas identificaron la mayor parte de las variantes (con un 75,38% de unificación), seguidas de las técnicas sintácticas (con un 19,11%) y de las semánticas (con un 5,5%). Para conocer la efectividad de nuestro método, se realizó el mismo análisis de tendencias sobre 1) el conjunto de claves aplicando únicamente *stemming* (modelo pre-procesado), 2) el conjunto de claves aplicando la técnicas de unificación léxica y sintáctica (modelo unificado léxica y sintácticamente) y 3) el conjunto de claves aplicando el método de unificación completo (modelo unificado). Al evaluar la coherencia

temática de los temas extraídos para las tres redes de co-ocurrencia obtenidas (una para cada modelo), se pudo comprobar que el modelo pre-procesado generó 8 temas sin demasiada coherencia, mientras que el modelo unificado léxica y sintácticamente y el modelo totalmente unificado obtuvieron 5 y 6, respectivamente, temáticas de alta calidad frente al total de 9 temáticas generadas por cada aproximación. Aunque las técnicas semánticas identificaron el menor porcentaje de variantes, generaron un gran impacto en el resultado final puesto que añadieron una temática de alta calidad adicional al análisis. Además, en la comparativa realizada para las tres aproximaciones, se pudo observar que en el modelo sin unificación emergieron, sobre la red de co-ocurrencia, términos de significado amplio que se trasladaron a la composición de los temas, mientras que los otros dos modelos están caracterizados por términos más específicos. Los resultados experimentales revelaron, de esta forma, la efectividad de nuestra unificación propuesta en términos de precisión y *recall*, así como la calidad de las tendencias temáticas en términos de coherencia. También demostraron que la unificación de palabras clave de autor reduce la cantidad de términos de significado amplio que surgen en una red de co-ocurrencia.

Por último, este trabajo de tesis doctoral ha comparado los resultados obtenidos en el análisis de tendencias temáticas mediante la herramienta PIMAnnot y mediante la unificación de análisis de palabras clave de autor. Ambos procedimientos proporcionaron resultados complementarios, aunque el método basado en palabras clave detectó un mayor número de temáticas. Esto se puede deber a que el glosario de términos de la comunidad científica es más limitado, al no incorporar la terminología más actual, que sí está presente en las palabras clave de autor.

En resumen, el presente trabajo de investigación se ha centrado en diseñar e implementar diferentes herramientas, técnicas y métodos para extraer conocimiento de grandes conjuntos de publicaciones científicas, de forma precisa, eficaz, rápida y sin pérdida de información. El anotador semántico desarrollado extrae información relevante, con precisión y exhaustividad aceptables, en dos áreas de aplicación diferentes: el área biomédica, haciendo uso de la ontología HPO, y en

la investigación de dirección de proyectos, a partir de un diccionario basado en glosarios validados por el PMI. EL método automático de generación de variantes léxicas (sinónimos) sobre la ontología HPO influyó positivamente sobre el reconocimiento automático de conceptos de las publicaciones científicas, sobre todo al aportar sinónimos altamente informativos, es decir, menos frecuentes en los textos pero que aportan mayor información al análisis. La combinación de técnicas léxicas, sintácticas y semánticas para unificar palabras clave de autor proporciona un método automático y eficiente que mejora el análisis de tendencias temáticas. Se espera que nuestros hallazgos allanen el camino para unificar palabras clave dispares de un *corpus* de documentos y, por lo tanto, aumenten aún más la confiabilidad y validez de los análisis basados en palabras clave. Nuestra propuesta abre la puerta a la viabilidad de desarrollar herramientas orientadas al usuario final que automaticen el proceso de unificación de palabras clave.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ÍNDICE GENERAL

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CAPÍTULO 1
INTRODUCCIÓN

Gracias al auge de las nuevas tecnologías, hoy en día es posible disponer de la información en cualquier lugar y en cualquier momento. Sin embargo, dada la inmensa cantidad de información existente, la búsqueda es aún una tarea que puede llegar a consumir mucho tiempo. Por esta razón, la comunidad científica y las grandes empresas tecnológicas están destinando importantes recursos al desarrollo de herramientas que simplifiquen dicha tarea, con el objetivo de ahorrar tiempo y esfuerzo a los usuarios.

1.1 ANTECEDENTES Y REVISIÓN BIBLIOGRÁFICA

A partir de la revolución tecnológica de la década de los 50, la difusión y el intercambio de la información se convierten en dos pilares fundamentales en la era de la digitalización. Pero es a partir del 2002 cuando se consolida el almacenamiento digital de documentos y, por primera vez, éste supera al almacenamiento tradicional en papel, en cuanto a número de ejemplares. Hoy en día, la proliferación de la documentación crece a pasos agigantados, superando incluso la capacidad de almacenamiento y procesamiento. Esta situación complica, tanto a entidades como a usuarios particulares, las tareas de recopilación, filtrado, análisis y visualización de la información, orientadas a extraer el máximo beneficio de su disponibilidad. Así pues, el verdadero problema no es la gran cantidad de información disponible, sino nuestra incapacidad para filtrarla y analizarla correctamente.

El filtrado correcto de la información relevante (sin pérdida de datos), en el mínimo tiempo posible, se convierte entonces en un aspecto clave. En este contexto surge el concepto de *Text Mining* o Minería de Textos, una rama de la minería de datos que se centra en analizar y derivar información nueva de documentos de texto (Zhu et

al., 2013), tales como páginas web, correos electrónicos o artículos de revistas, por ejemplo. Se trata de un método muy eficiente para generar nueva información y conocimiento, que de otro modo sería difícil de encontrar. Esta práctica permite reducir el tiempo dedicado a la lectura de textos extensos para deducir nueva información (Cohen et al., 2005). La tecnología de minería de textos es actualmente aplicada por una extensa variedad de individuos, ya que sus aplicaciones son muy diversas:

- Investigación. En esta área se invierte mucho tiempo en analizar y obtener información relevante y, a veces, difícilmente accesible. La minería de textos permite a los investigadores encontrar más información y de forma más rápida y eficiente. En particular, en el área de la medicina se valora la rapidez en beneficio del avance en el estudio de ciertas enfermedades (Pletscher-Frankild et al., 2015).

- Negocios. En el mundo empresarial esta tecnología encuentra muchas aplicaciones en la analítica de los datos de sus clientes y ventas, principalmente, para la toma de decisiones de negocio (Fattori et al., 2003). Otra de sus aplicaciones actuales son los algoritmos de respuesta inteligente a los clientes, que suelen estar basados en un conjunto de preguntas y respuestas comunes para dar solución a las dudas más básicas (Guerreiro y Rita, 2020).

- Seguridad. En este dominio destaca el análisis de blogs o páginas web para prevenir delitos en Internet. La minería de texto permite el reconocimiento de palabras clave para identificar posibles fraudes (Holton, 2009).

- Mensajería electrónica. Uno de los usos más cotidianos de la minería de textos es la aplicación de filtros de correo electrónico (Khan y Qamar, 2016). A partir del reconocimiento de una serie de palabras clave o patrones (por ejemplo, el correo que

proviene de un cierto destinatario o que contiene ciertas palabras en el asunto), se puede organizar el buzón de correo.

El campo de la extracción de la información también abarca la anotación de textos. La práctica de la anotación la inician los eruditos medievales como un foro de debate e intercambio de conocimiento, y es una práctica muy habitual hoy en día en diferentes disciplinas (Wolfe, 2002). En el campo de la minería de textos, la anotación se ha centrado fundamentalmente en el reconocimiento de entidades en texto. Entre otras tecnologías, el reconocimiento de entidades puede apoyarse en algún diccionario terminológico que proporcione el vocabulario utilizado en el dominio de estudio. Cuando las entidades que se reconocen en el texto son conceptos de una ontología o terminología estandarizada, se habla de anotación semántica (Uren et al., 2006). Podemos definir la anotación semántica como el proceso de enriquecimiento de los documentos textuales con metadatos, es decir, con referencias que vinculan el contenido con conceptos. La descripción de dichos metadatos viene dada por la ontología, que facilita encontrar, interpretar y reutilizar los contenidos no estructurados en lo que aparecen las anotaciones.

El procedimiento utilizado en la anotación semántica consiste en el reconocimiento de entidades nombradas (NER) (Atkinson y Bull, 2012), realizado mediante procesamiento de lenguaje natural (PLN), seguido de una anotación tradicional (Kiryakov et al., 2005). Un ejemplo de entorno de anotación semántica es GATE (Cunningham et al., 2013), una herramienta que proporciona facilidades (tokenizadores, rotuladores de partes del habla, gramáticas de coincidencia de patrones, etc.) para desarrollar y distribuir módulos de software que procesan el lenguaje natural. Por ejemplo, la plataforma KIM (Huang et al., 2009) (Gestión de conocimiento e información) ofrece un anotador semántico basado en GATE y KIMO, una ontología formal de nivel superior. También se ha desarrollado una plataforma semántica para la anotación de servicios en la nube (Bettembourg et al., 2012) y otra en el dominio de aprendizaje electrónico (Tsatsaronis et al., 2012), utilizando GATE para reconocer entidades nombradas de múltiples ontologías. La primera plataforma aplica enfoques estadísticos basados en las

estructuras sintácticas del texto para eliminar la ambigüedad de las entidades reconocidas por GATE, mientras que la segunda amplía las anotaciones reconocidas en el texto con un gráfico, lo que facilita el acceso y la navegación de los contenidos.

Aunque las herramientas de anotación basadas en PLN pueden lograr resultados de buena calidad, necesitan una gran memoria y recursos computacionales. Existen también alternativas que reemplazan las técnicas de análisis lingüístico por otros procedimientos. SemTag (Dill et al., 2003) anota páginas web a gran escala con términos de una ontología de nivel superior, llamada TAP. Los textos se tokenizan y luego se procesan para encontrar todas las instancias de la ontología. Cada anotación de un término candidato se guarda con 10 palabras a cada lado (contexto). Simultáneamente, se escanea una muestra representativa de las páginas web para determinar los contextos de cada concepto en la ontología. Se utiliza un modelo de espacio vectorial para eliminar la ambigüedad de las anotaciones de los candidatos al comparar sus contextos con los contextos conceptuales en la ontología. Este enfoque presenta una alta precisión porque construye dinámicamente los contextos en la ontología utilizando el mismo texto que se debe anotar. Sin embargo, también necesita gran memoria y recursos computacionales. Otros enfoques dependen principalmente de las expresiones regulares (Wessman et al., 2005) o las gramáticas libres de contexto (Kiyavitskaya et al., 2009) para reconocer las entidades nombradas. En este sentido, no consumen muchos recursos, pero requieren la construcción de la gramática para cada aplicación.

Por otro lado, existen métodos de anotación semántica basados en el aprendizaje automático, como Word2Vector (Ma y Zhang, 2015). Este algoritmo pertenece al conjunto de lenguajes de modelado y técnicas aprendizaje en procesamiento del lenguaje natural (PLN), en dónde las palabras o frases del lenguaje se representan como vectores de números reales. A este conjunto de lenguajes se les denomina *word embedding* (Naili et al., 2017). Word2Vector primero construye un vocabulario a partir del corpus de texto de análisis y aprende las representaciones vectoriales de cada palabra. Además, este algoritmo tiene la capacidad de calcular la distancia en términos de similitud semántica entre cada palabra.

Ya en el dominio específico de la biomedicina, donde existen excelentes recursos terminológicos estandarizados y ontologías, una alternativa es aprovechar dichos recursos para reconocer sus entidades en los textos. Este es el enfoque propuesto por el anotador del NCBO (Musen et al., 2012), que aplica un algoritmo de concordancia de cadenas, llamado Mgrep, para anotar los textos con conceptos de una o varias ontologías seleccionadas. Mgrep proporciona mayor velocidad, precisión, flexibilidad y escalabilidad que los anotadores basados en PLN (Shah et al., 2009).

La anotación o reconocimiento de entidades tiene especial interés en el área de la biomedicina, dónde el análisis de casos clínicos publicados en la literatura existente, de forma rápida y precisa se convierten en factores determinantes en el diagnóstico y tratamiento de pacientes con síntomas raros y poco conocidos. Es por ello que expertos del dominio demanden el uso de herramientas que faciliten este proceso y les ayuden en su trabajo diario. Pero no solo la biomedicina está interesada en este tipo de herramientas. En otros dominios totalmente diferentes como, por ejemplo, la investigación en dirección de proyectos, también se reclaman herramientas que tengan la habilidad para filtrar y organizar la información correctamente, con el fin de avanzar en el conocimiento del dominio.

1.1.1 Extracción de información relevante en el ámbito biomédico

La cantidad de información en el campo de la biomedicina crece a gran velocidad. PubMed (McKenzie, 1996), uno de los grandes repositorios de información biomédica, cuenta con más de 32 millones de documentos. Filtrar la información relevante en el menor tiempo posible es uno de los grandes retos actuales. Este aspecto es relevante dentro del estudio de las enfermedades raras, donde el número de pacientes existente es muy limitado. La extracción de la mayor cantidad posible de casos clínicos disponibles en la bibliografía permite avanzar en el estudio de la patología, y mejorar su diagnóstico y tratamiento.

Los métodos de búsqueda clásicos basados en palabras clave se ven muy limitados dentro de este contexto, debido a las inconsistencias entre la terminología utilizada por cada autor (variaciones ortográficas,

sinónimos, siglas o términos con significado demasiado amplio, por ejemplo). Ante esta situación, en los últimos años han ido surgiendo herramientas online de uso libre que proporcionan la información clínica de PubMed de una forma más eficiente y organizada. Ejemplos de tales herramientas son GoPubMed (Müller et al., 2004), Anne O'Tate (Smalheiser et al., 2006) o PubReMiner (Slater, 2014).

GoPubMed (Müller et al., 2004) fue uno de los primeros buscadores de la Web 2.0. Esta herramienta consistía en un buscador basado en ontologías. Utilizaba la terminología MeSH (Medical Subjects Headings) (Medicine, National Library of Medical Subject Headings, 2003) y la ontología GO (Gene Ontology) (Doms y Schroeder, 2005) como apoyo en la búsqueda de información. Como resultado proporcionaba los artículos encontrados anotados con términos de MeSH y GO.

Anne O'Tate (Smalheiser et al., 2006) es una aplicación gratuita basada en web para realizar búsquedas en PubMed que consiste en realizar minería de texto avanzada sobre la literatura biomédica. La herramienta está diseñada para extraer información al buscar por términos clave relevantes, términos de títulos de temas médicos (MeSH) y datos bibliométricos, para ayudar a los usuarios a perfeccionar y desarrollar sus estrategias de búsqueda.

PubReMiner (Slater, 2014) es otro asistente para la búsqueda en PubMed que permite iniciar la exploración de dicha base de datos a partir de una palabra cualquiera. Gracias a su búsqueda avanzada, posibilita alcanzar un resultado que se ajuste mejor a los intereses bibliográficos de quienes utilizan el sistema.

El objetivo de todas estas herramientas es facilitar las tareas de búsqueda a los investigadores, extrayendo los datos biomédicos más relevantes, es decir, que se ajusten a sus necesidades de información dentro de PubMed. Pese a todo, estas herramientas siguen sin representar un modelo ideal, ya que la información que devuelven en muchos casos depende en gran medida del dominio clínico concreto y de las ontologías utilizadas. En esta tesis se propone un nuevo método para extraer automáticamente publicaciones sobre casos clínicos de pacientes con enfermedades raras de PubMed y, posteriormente, anotar la información recuperada con sus anomalías fenotípicas, en nuestro

caso conceptos de la ontología HPO (Robinson et al., 2008). El objetivo es mejorar la extracción de los casos clínicos relevantes, así como mejorar su anotación semántica, con el fin último de facilitar su búsqueda.

1.1.2 Ampliando el léxico del ámbito biomédico mediante aprendizaje

Las técnicas de reconocimiento de entidades han demostrado ser muy útiles en la minería de textos biomédicos. Recientemente, se han aplicado con éxito para identificar entidades en la investigación del cáncer (Zhu et al., 2013), factores de riesgo de enfermedades cardíacas en pacientes diabéticos (Urbain et al., 2015) o información fenotípica (Köler et al., 2014), entre otras. Los reconocedores de entidades biomédicas se incluyen principalmente en las categorías generales de enfoques basados en terminología, basados en reglas y en patrones de aprendizaje estadístico (Cohen et al., 2005). Además, las ontologías han desempeñado un papel clave como recursos terminológicos para extraer información de textos biomédicos (Funk et al., 2014). Sin embargo, los conceptos de una ontología son difíciles de reconocer en el texto, ya que comúnmente su descripción en la ontología es diferente de su descripción en el texto (Schulz et al., 2013). Para tratar de solventar este problema, en esta tesis doctoral se propone un método automático para ampliar el léxico de la ontología estudiada, la HPO (*Human Phenotype Ontology*) (Köhler et al., 2014). Nuestro aportación principal es el diseño e implementación de un nuevo método de generación automática de variantes léxicas de los términos incluidos en la ontología, combinando procedimientos léxicos y lógicos.

1.1.3 Extracción de información en la literatura sobre investigación en dirección de proyectos

La investigación en dirección de proyectos (PM) es un campo muy interdisciplinar debido a su alta aplicabilidad en diferentes entornos, desde proyectos de construcción e ingeniería del software hasta proyectos del día a día. Actualmente, PM aún no es una disciplina bien establecida, ya que está en constante evolución, no solo por su avance en el campo de estudio sino también por su carácter multidisciplinar y

de expansión hacia nuevos campos de aplicación (Pollack & Adler, 2015). Por esta razón, el análisis de las tendencias temáticas se ha convertido en una valiosa herramienta en la investigación del PM en las últimas dos décadas (Urli, 2000; Söderlund, 2004; Smyth & Morris, 2007; Carden & Egan, 2008; Kwak & Anbari, 2009; Artto, 2009; Polack & Adler, 2015; Padalkar & Gopinath, 2016). De ahí, la necesidad de proveer una herramienta que permita detectar automáticamente las tendencias temáticas, mediante la aplicación de técnicas de extracción de conocimiento. En este trabajo se propone una herramienta de detección de tendencias temáticas en PM, que facilita a los investigadores y académicos la toma de decisiones, al proporcionarles información sobre la investigación que se está realizando y las oportunidades de innovación en el campo.

1.1.4 Unificación de palabras clave de autor en la investigación en dirección de proyectos

Una alternativa a la anotación semántica de textos para la extracción de información es el uso de las palabras clave de autor de los artículos de investigación. Las palabras clave de autor se consideran la unidad mínima de resumen de un documento, cubriendo las temáticas principales sobre las que versa el artículo (Liu et al, 2009). Por este motivo, son muy utilizadas para tareas como el resumen de documentos (Deng et al., 2020), la agrupación de documentos (Kim y Cho, 2020), la indexación automática (Vega-Oliveros et al., 2019), la clasificación (Rinaldi et al., 2020) o el análisis de tendencias temáticas (Mao et al., 2010; Pollack&Adler, 2015; Kim et al., 2020), entre otras. Sin embargo, estos términos clave de autor suelen adolecer de un alto grado de heterogeneidad (a menudo hay muchas formas diferentes de representar el mismo tipo de información), que procede fundamentalmente de 1) el uso de términos demasiado específicos que y, a la vez, poco frecuentes o únicos en el conjunto de documentos a analizar; 2) la utilización de claves muy generales, es decir, poco descriptivas; 3) el empleo indistinto de acrónimos y sus términos equivalentes completos; 4) el uso de claves que no son nombres o pronombres (por ej., frases nominales con preposiciones o conjunciones). Por esta razón, surge la necesidad de seleccionar y

unificar la terminología considerada en el conjunto de palabras clave (Ding, 2001). Con la unificación se reduce considerablemente el número de términos diferentes para designar el mismo concepto, lo que conduce a una representación más clara y precisa del mapa conceptual del dominio. Actualmente, esta etapa se suele realizar de manera manual, normalmente con la ayuda de expertos en el dominio. Siguiendo este procedimiento, se requiere dedicar mucho tiempo y esfuerzo a revisar y unificar manualmente las claves, antes de proceder a su análisis (Zhang et al., 2016; Kahssed et al., 2017; Leung et al., 2017). En los campos en los que existen recursos terminológicos, tales como biomedicina (MESH), física (PhySH) o economía (STW), éstos se pueden utilizar para clasificar las palabras clave en temáticas de investigación. Sin embargo, esta solución no es viable en todas las áreas, como la de investigación en dirección de proyectos (PM), porque no se dispone de un glosario terminológico aceptado y reconocido por la comunidad de investigación, ni de buenos recursos terminológicos que faciliten el proceso de normalización del vocabulario. En este trabajo proponemos un conjunto de técnicas léxicas, sintácticas y semánticas para automatizar, en la medida de lo posible, la unificación de palabras clave de autor, con el fin de mejorar la calidad de nuestros resultados.

1.2 HIPÓTESIS Y OBJETIVOS

La hipótesis fundamental de este trabajo doctoral es que el procesamiento automático de información textual se puede mejorar haciendo uso de los recursos de conocimiento del dominio de aplicación, junto con las técnicas léxicas, sintácticas y semánticas que permiten aprovechar al máximo dichos recursos. Teniendo en cuenta esto, las hipótesis de partida de nuestro trabajo son:

1. Es posible incrementar la eficacia de los anotadores basados en ontologías, mejorando los algoritmos de anotación y el procesamiento de los recursos terminológicos existentes.

2. Es posible aprender nuevas variantes léxicas a partir de los términos existentes en la ontología y las relaciones jerárquicas entre conceptos.

3. La unificación de las palabras clave de autor que indexan las publicaciones científicas permiten mejorar los resultados del análisis de tendencias temáticas de un dominio.

El objetivo principal de este trabajo es el diseño e implementación de técnicas para la extracción de información relevante y útil en el dominio de interés. Este objetivo global se desglosa en los siguientes sub-objetivos:

- Mejorar las técnicas de anotación semántica de información.

- Proponer nuevas técnicas para la extracción de las secciones relevantes en las descripciones textuales de casos clínicos en las publicaciones científicas.

- Proponer nuevas técnicas lexicológicas para enriquecer las terminologías/ontologías con nuevas variantes léxicas, que mejoren las prestaciones de los anotadores semánticos.

- Proponer una metodología automática para la detección de tendencias en el campo de la dirección de proyectos.

- Proponer un método automático de unificación de palabras clave de autor, que mejore el estudio de las tendencias temáticas, en un dominio concreto de aplicación.

1.3 ESTRUCTURA DE LA MEMORIA

Nuestro trabajo de investigación comienza introduciendo las principales técnicas y herramientas sobre anotación semántica en el capítulo 2. En esta sección se define qué es una ontología y se realiza un breve repaso de los principales anotadores semánticos sobre ontologías biomédicas. Además, introduciremos el análisis de co-

ocurrencia de términos, como un método ampliamente usado para extraer conocimiento de las publicaciones, e introduciremos brevemente el análisis de *burst*, otro de los principales métodos utilizados en nuestros estudios experimentales.

A continuación, en el capítulo 3, describiremos el diseño e implementación de nuestro anotador semántico, denominado OBO Annotator. Al igual que Mgrep (Shah et al., 2009), OBO Annotator aplica un algoritmo que compara directamente los términos de la ontología con los textos a anotar. Esto permite mejorar la velocidad de anotación, sin perder eficiencia, con respecto a las herramientas basadas en el procesamiento del lenguaje natural, como Metamap (Aronson, 2006). Previo a la anotación, OBOAnnotator construye un diccionario a partir de todos los nombres y sinónimos de los conceptos de la ontología. Durante la anotación, utiliza una ventana de rastreo de secuencias de texto (por defecto, 4 palabras) para extraer, procesar y generar mediante permuaciones todas las cadenas de caracteres que se compararán con los términos léxicos pre-procesados del diccionario. El anotador utiliza las relaciones is-a contenidas en la ontología para filtrar anotaciones redundantes en un mismo texto. En concreto, aquellas anotaciones correspondientes a conceptos relacionados a través de relaciones padre-hijo se consideran redundantes, y solamente las anotaciones correspondientes a los conceptos más específicos se mantienen. EL OBOAnnotator se ha probado sobre la *Human Phenotype Ontology* (HPO) para extraer el conocimiento relevante sobre casos clínicos de pacientes de CTX. El proceso de validación basado en un entorno de comparación con otros anotadores disponibles en el momento confirmó que el OBOAnnotator fue la herramienta que anotó de forma más precisa y eficiente. Posterioremente, para mejorar los resultados alcanzados con el OBO Annotator, se diseñaron e implementaron nuevas técnicas de generación de sinónimos (principalmente, variantes léxicas). Dichas técnicas han permitido ampliar la cobertura de la terminología de la ontología HPO. La principal novedad de estas técnicas, explicadas en el capítulo 4, es el uso de las propiedades léxicas y lógicas de la ontología para inferir nuevos términos.

Una vez mejorada la anotación semántica sobre textos de investigación biomédicos, en el capítulo 5 probamos nuestra herramienta en un nuevo entorno de investigación, la dirección de proyectos (PM). La anotación se utiliza en este escenario para extraer las principales tendencias temáticas de la disciplina, a lo largo de un período dado de tiempo. Para este entorno de investigación, no existen recursos terminológicos, como ontologías, por lo que se partirá de un diccionario construido a partir de glosarios ya existentes en el área y aceptados por la comunidad científica.

En el capítulo 6, como resultado de las limitaciones encontradas en el diccionario propuesto, se propone comparar el análisis de las tendencias temáticas basado en la anotación textual frente al basado en las palabras clave de autor. Durante el desarrollo de dicho entorno, detectamos un problema importante pero poco abordado hasta el momento: la unificación de un gran conjunto de palabras clave heterogéneas, como resultado de la extracción de documentos dentro de un dominio. Para abordar esta carencia, en esta tesis proponemos un enfoque automatizado que se basa en la combinación de un conjunto de técnicas léxicas, sintácticas y semánticas.

En resumen, en esta tesis se facilitan diferentes técnicas léxicas, sintácticas y semánticas para mejorar y facilitar el procesamiento de grandes conjuntos de datos y extraer la información relevante. Las etapas en las que se divide nuestro trabajo de investigación son:

I. **Diseño e implementación de una herramienta informática para la anotación semántica de informes clínicos de pacientes (OBO Annotator).**

   El objetivo de este proceso es la recuperación, anotación semántica e indexado de los artículos de investigación relacionados con casos clínicos de pacientes con enfermedades raras. Usando los resúmenes disponibles de PubMed, nuestro método identifica los artículos asociados a informes clínicos de pacientes y extrae automáticamente los fragmentos relativos a la descripción del fenotipo clínico. Para ello, el método usa una gramática general desarrollada para tal fin. Para anotar los segmentos textuales extraídos, el algoritmo compara secuencias

de texto pre-procesadas con términos de un diccionario construido offline a partir de todos los nombres y sinónimos de la ontología. Usando las relaciones jerárquicas definidas en la ontología, las anotaciones de conceptos generales son filtradas, en presencia de anotaciones correspondientes a conceptos más específicos. Finalmente, a partir del conjunto completo de anotaciones, el algoritmo extrae una sub-ontología mínima para el contexto de la enfermedad de estudio (CTX).

II. **Diseño e implementación de un nuevo método de aprendizaje de sinónimos para la ontología HPO.**

En esta tesis presentamos un nuevo procedimiento automatizado de sustitución de sinónimos orientado a enriquecer la ontología HPO con nuevos sinónimos. Sobre la base de que la estructura de HPO es altamente compositiva (Gkoutos et al., 2009; Groza et al., 2013; Oellrich et al., 2013), planteamos la hipótesis de que HPO podría enriquecerse mediante técnicas lexicológicas. Nuestro método identifica subs-ecuencias comunes de palabras compartidas entre un término y sus términos descendientes, para generar los nuevos sinónimos. Esto hace posible aplicar la técnica a toda la ontología y no restringirlo a partes específicas como en (Groza et al., 2013; Oellrich et al., 2013). Además, debido a que PubMed es un excelente recurso que proporciona evidencia actualizada sobre el uso de la terminología por parte de la comunidad, también suponemos que validar la existencia de los sinónimos generados mediante la búsqueda de estas frases exactas en MEDLINE puede ayudar a descartar automáticamente los sinónimos sin sentido. El método propuesto incluye dos etapas principales. En la primera, comienza identificando recursivamente todas las superposiciones léxicas que se dan en la ontología HPO, es decir, todos los pares de términos conectados por una relación jerárquica y donde el término descendiente incluye el término ascendente como sub-cadena propia. Posteriormente, reconoce

las frases exactas de los sinónimos generados en MEDLINE y descarta aquellas para las que no se recuperó ningún resultado.

III. **Desarrollo de un método novedoso y automatizado para la detección de tendencias en la dirección de proyectos.**

En esta etapa, se propondrá una metodología para analizar automáticamente las publicaciones científicas y académicas relevantes en la investigación de dirección de proyectos, con la finalidad de identificar tendencias temáticas. La principal aportación de esta etapa es el diseño de un método automático para la detección de tendencias en PM, basado en la anotación semántica de documentos y el análisis de co-ocurrencia.

La metodología propuesta abarca cuatro etapas principales. En un primer momento, se diseña la consulta de búsqueda para la adquisición de la fuente de datos utilizada para el estudio. A continuación, los datos recuperados, se procesan automáticamente con nuestra herramienta de anotación, para lo cual se utiliza un diccionario hecho a medida con la terminología propia del dominio. Para identificar temáticas sobre la dirección de proyectos, se elaboran redes de co-ocurrencia de términos, que permiten identificar los términos con mayor frecuencia de asociación con el resto de los términos. Finalmente, se realiza el análisis y visualización de la red de co-ocurrencia para extraer los tópicos y sus tendencias en el periodo analizado.

IV. **Unificación de términos clave de autor para la identificación de tendencias temáticas en la investigación en dirección de proyectos.**

El objetivo de este apartado es proporcionar un método capaz de mejorar la calidad de las palabras clave de autor que tradicionalmente se utilizan para el análisis del mapa conceptual de un dominio concreto de estudio. Se trata de pre-procesar el conjunto terminológico inicial utilizado en los análisis. La metodología propuesta aúna un conjunto de técnicas léxicas, sintácticas y semánticas. Con ellas, se pretende mejorar la

calidad de los resultados obtenidos. En contraposición a la herramienta de anotación utilizada para la extracción de tendencias en PM, en este estudio pretendemos analizar la calidad y precisión de los resultados únicamente pre-procesando el léxico utilizado por los autores en sus publicaciones. Se trata de una alternativa ante la inexistencia de recursos terminológicos propios del dominio.PUBLICACIONES

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

1.3.1 Publicaciones en revistas JCR

- Taboada, M., Rodríguez, M., Martínez, D., Pardo, M. y Sobrido, M.J., (2014). *Automated semantic annotation of rare disease cases: a case study.* Database Oxford Journals.

- Maarouf, H., Taboada, M., Rodríguez, H., Arias, M., Sesar, Á. y Sobrido, M.J., (2017). *An ontology-aware integration of clinical models, terminologies and guidelines: an exploratory study of the Scale for the Assessment and Rating of Ataxia (SARA).* BMC Med Inform Decis Mak. Volumen 17.

- Taboada, M., Rodriguez, H., Gudivada, R.C. y Martinez, D., (2017). *A new synonym-substitution method to enrich the Human Phenotype Ontology*. BMC Bioinformatics. Volumen 18.

1.3.2 Publicaciones en congresos

- Rodríguez, H., (2015). *Detección de tendencias temáticas en la investigación reciente sobre dirección y gestión y gestión de proyectos.* XIX Congreso Internacional de Dirección e Ingeniería de Proyectos. AEIPRO 2015.

- Rodríguez, H., (2016). *Una ontología para el ámbito de la dirección de proyectos.* XX Congreso Internacional de Ingeniería y Dirección de Proyectos AEIPRO 2016.

1.3.3 Publicaciones en revisión en revistas JCR

- Rodríguez, H., Díaz, E., Martínez, D. y Taboada, M., (2021). An automated approach to unifying keywords: A study on thematic trends in project management. Expert Systems with Applications.

CAPÍTULO 2
ESTADO DEL ARTE

En esta sección se documentan los recursos y metodologías utilizados durante el presente trabajo de investigación. De esta forma, se detalla el contexto de actuación para un mejor entendimiento del trabajo de análisis e investigación. Por un lado, comenzaremos introduciendo las ontologías y las herramientas de anotación semántica. Continuaremos hablando de técnicas de *machine learning* y, finalmente, presentaremos los principales métodos de representación y análisis utilizados en este estudio.

2.1 ONTOLOGÍAS

Una ontología consiste en una especificación del vocabulario compartido en un dominio concreto, como la definición de clases, relaciones, funciones y otros objetos (Gruber, 1993). Así, una ontología representa la visión de un dominio de aplicación común, reusable y compartible y provee de significado a las estructuras de información que intercambian los sistemas de información (Müller et al., 2003).

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

El uso de ontologías para representar conocimiento biomédico no es nuevo, ya que han sido usadas ampliamente en este campo con propósitos diferentes. SNOMED CT, *Gene Ontology* y *Human Phenotype Ontology* son alguno de los ejemplos más importantes de ontologías en este dominio de aplicación. En los siguientes sub-

apartados introduciremos estas tres ontologías y hablaremos de repositorios, formatos y software para la consulta de ontologías.

2.1.1 Human Phenotype Ontology (HPO)

La Ontología del Fenotipo Humano (HPO) proporciona un vocabulario estandarizado de las anormalidades fenotípicas que se encuentran en las enfermedades humanas (Robinson et al., 2008). Cada término de la HPO describe una anomalía fenotípica, como la comunicación interauricular. El HPO se está desarrollando actualmente utilizando la literatura médica, Orphanet, DECIPHER y OMIM. La HPO contiene actualmente más de 13.000 términos y más de 156.000 anotaciones a enfermedades hereditarias. El proyecto HPO y otros han desarrollado software para el diagnóstico diferencial basado en el fenotipo, el diagnóstico genómico y la investigación traslacional. La HPO es un producto emblemático de la Iniciativa *Monarch*, un consorcio internacional apoyado por los NIH dedicado a la integración semántica de datos biomédicos y de organismos modelo con el objetivo final de mejorar la investigación biomédica. La HPO, como parte de la Iniciativa Monarch, es un componente central de uno de los 13 proyectos impulsores de la hoja de ruta estratégica de la Alianza Global para la Genómica y la Salud (GA4GH).

HPO puede usarse para diagnósticos clínicos en genética humana (*Phenomizer*), investigación bioinformática centrada en las relaciones entre las anormalidades del fenotipo humano y las redes bioquímicas y celulares, y como vocabulario estándar de bases de datos clínicas, entre muchas otras posibilidades.

2.1.2 Gene Ontology

Gene Ontology (GO) (Vanteru et al., 2008) es un proyecto cuyo objetivo es producir un vocabulario dinámico y controlado que se pueda aplicar a todas las células a pesar de lo reciente o cambiante que pueda ser el conocimiento celular génico y proteico. El proyecto de GO es un esfuerzo colaborativo que trata de ofrecer descripciones consistentes de productos génicos a lo largo de distintas bases de datos. El proyecto comienza en 1998 como una colaboración entre tres modelos de bases de datos de organismos: *Fly Base* (Drosophila), la base de datos

Saccharomyces Genome (SGD) y la *Mouse Genome* (MGD). Desde entonces, el consorcio GO ha crecido para incluir más bases de datos, incluyendo varios de los mayores repositorios de genomas de plantas, animales y microorganismos.

En la ontología GO cada término tiene definidas relaciones con uno o más términos pertenecientes al mismo dominio y, a veces, con otros dominios. El vocabulario está diseñado para que sea independiente de la especie e incluye términos aplicables a células procariotas y eucariotas, organismos unicelulares y multicelulares.

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

2.1.3 SNOMED CT

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) (Dhombres y Bodenreider, 2016) es una terminología multilíngüe gestionada por la *International Health Terminology Standards Development Organisation* (IHTSDO). Se trata de un estándar para la información clínica que se usa en más de 50 países y está alineada con otros estándares internacionales. Hablamos de un vocabulario clínicamente validado, semánticamente rico y controlado que permite una gran expresividad. SNOMED CT es una terminología global pero que permite ser adaptada a las necesidades de un país o una región, mediante mecanismos como los *reference sets*. Esta terminología está centrada en los conceptos. También, está traducido a cinco idiomas y en vías de traducción a otras lenguas

SNOMED CT ofrece una forma estandarizada de representar la información clínica creada por el médico permitiendo así una interpretación automatizada. SNOMED CT contribuye a la mejora del

cuidado del paciente mediante el desarrollo de HCE que permite la recuperación de información clínica de una forma basada en significado, sirve de ayuda en procesos como el diagnóstico, el desarrollo de informes estadísticos consistentes, en el análisis de costes, vigilancia de la salud pública, etc. Los pacientes se benefician del uso de SNOMED CT porque mejora la información de la historia clínica y facilita la comunicación. Por lo tanto, SNOMED CT se usa en la recolección de gran variedad de información clínica, la conexión de bases de conocimiento, recuperación de la información, y agregación, análisis e intercambio de información entre otras funciones.

Actualmente, SNOMED CT contiene más de 310.000 conceptos activos unívocos, cada uno descrito por un término preferido y uno o más términos adicionales llamados sinónimos. Cada concepto, con significado único, está descrito lógicamente a través de sus relaciones con otros conceptos y organizado jerárquicamente.

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

- **Conceptos:** representan ideas clínicas. Cada concepto tiene un identificador único llamado *concept identifier*. Estos componentes están organizados en jerarquías que van desde lo general a lo específico, permitiendo que se almacene información referente a detalles clínicos para posteriormente agregarla en un nivel más general. Es decir, las relaciones **is a** posibilitan representar la correspondencia lógica de inclusión jerárquica.

- **Descripciones:** son las encargadas de asociar textos comprensibles con los conceptos, proporcionan la parte inteligible especificando el significado de los conceptos de SNOMED CT y así, permite diferenciar unos conceptos de otros dentro de la jerarquía. Existen principalmente dos tipos de descripciones: el *Fully Specified Name* (FSN) que

intenta ser un término diferenciador del concepto; y los sinónimos de un concepto que proporcionan descripciones alternativas para referirnos a este. Cada descripción tiene un identificador numérico.

- **Relaciones:** son vínculos que poseen un significado y que enlazan cada concepto a otros conceptos con significado vinculado. Estas relaciones denotan definiciones formales y otras características del concepto, es decir, lo definen lógicamente. Cada relación tiene un identificador único. La relación más relevante es la **is a** (es un) que relaciona un concepto con un concepto más general a este. Por ejemplo, *neumonía viral* tiene una relación de **is a** con el concepto general de *neumonía* (concepto padre). Esta relación **is a** define la jerarquía de los conceptos de SNOMED CT. También hay que tener en cuenta, que un concepto puede tener más de un concepto padre. Otros tipos de relaciones representan otras características de un concepto. Por ejemplo, el concepto neumonía viral tiene una relación de causative *agent* (agente causal) con el concepto virus y está vinculado al concepto pulmón mediante la relación *finding site* (localización). El conjunto de conceptos que pueden formar parte del origen de este tipo de relaciones se denomina *dominio* y aquellos que pueden formar parte del destino de la relación se llaman *rango*.

2.1.4 Portales de ontologías

2.1.4.1 OBO Foundry

La iniciativa *Open Biomedical Ontologies* (OBO) Foundry (Schober et al., 2009) es un entorno colaborativo entre los desarrolladores de ontologías basadas en la ciencia, que están estableciendo una serie de principios para el desarrollo de las ontologías con el objetivo de crear una colección de referencia e interoperable en el ámbito biomédico. De este modo, OBO Foundry proporciona un repositorio de ontologías biológicas y biomédicas, incluyendo la

Human Phenotype Ontology (HPO), ontología importante dentro del marco de desarrollo del presente proyecto.

2.1.4.2 BioPortal

El BioPortal es una herramienta web presentada por el NCBO que representa un portal de acceso a todas las ontologías y recursos terminológicos del ámbito biomédico. Incluye más de 300 ontologías, proporcionando un amplio y abundante léxico, lo que permite expandir las anotaciones con términos más generales a los identificados en el texto considerado. El núcleo es Mgrep, un motor de mapeo de conceptos basado en un eficiente algoritmo de coincidencia de cadenas.

El BioPortal permite buscar términos en ontologías, navegar por la lista de ontologías incluidas y buscar recursos biomédicos. También proporciona una de las principales herramientas de anotación que existen, en la que se nos permite seleccionar la ontología a utilizar para realizar el procesamiento sobre un texto concreto y exportar los términos encontrados en el texto introducido teniendo en cuenta la ontología seleccionada (anotaciones) en diferentes formatos, entre otras opciones.

2.1.5 Lenguajes de ontologías

OWL y OBO son lenguajes de representación formal de conocimiento, que facilitan formas de expresar un significado desde un punto de vista computacional, siendo este significado el que complementa y da estructura a la información ya presente. Sin embargo, para aplicar realmente estas tecnologías sobre volúmenes de información considerables se requiere la construcción de mapas detallados sobre dominios de conocimiento particulares. A continuación, introduciremos las principales puntos fuertes de cada uno de estos lenguajes.

2.1.5.1 OWL

Como su propio nombre indica, el Lenguaje de Ontologías Web[1] (OWL) está diseñado para ser usado en aplicaciones que

[1] https://www.w3.org/2007/09/OWL-Overview-es.html

necesitan procesar el contenido de la información en lugar de únicamente representar información para los humanos. Si bien existían lenguajes ontológicos previos, especialmente destinados a temas científicos y aplicaciones de comercio electrónico, no fueron definidos para ser compatibles con la arquitectura de la *World Wide Web* en general, y de la Web Semántica en particular. OWL (McGuiness y van Harmelen, 2004)facilita un mejor mecanismo de interpretación de contenido Web que los mecanismos admitidos por XML, RDF, y esquema RDF (RDF-S) proporcionando vocabulario adicional junto con una semántica formal. OWL tiene tres sub-lenguajes, con un nivel de expresividad creciente: *OWL Lite*, *OWL DL*, y *OWL Full*.

2.1.5.2 OBO

El formato de archivo plano OBO[2] es una forma sencilla de representar una ontología cuyos modelos representan un subconjunto de los conceptos en el lenguaje lógico de descripción OWL (McGuiness y van Harmelen, 2004). El formato OBO fue diseñado para maximizar la legibilidad humana y ha sido ampliamente utilizado en el campo de las bio-ontologías y probablemente ha sido uno de los muchos factores responsables del éxito de la Ontología Genética. La facilidad de análisis, la extensibilidad y la redundancia mínima son sus características básicas.

2.1.6 Manipulación de ontologías: Protégé

Protégé es un software desarrollado por el *Standford Center for Biomedical Informatics research* de la escuela de medicina de la universidad de Standford, que pretende proporcionar una interfaz de creación y manipulación de ontologías en diferentes lenguajes.

Este programa soporta la creación y edición de una o más ontologías en un único espacio de trabajo. Además se caracteriza por:

- Soporte para editar ontologías en formato OWL y OBO.

[2] http://owlcollab.github.io/oboformat/doc/GO.format.obo-1_2.html

- Una interfaz simple e intuitiva para la edición, la cual provee acceso a los constructores más comúnmente utilizados por OWL.

- Historial de cambios y de revisiones realizadas.

- Herramientas colaborativas como compartición, permisos y notificaciones vía email.

- Interfaz de usuario customizable.

- Formularios web para su edición.

2.2 TÉCNICAS DE EXTRACCIÓN DE CONOCIMIENTO

En la siguiente sección introduciremos la minería de textos y profundizaremos en la anotación semántica. Ambos procesos se basan en la extracción de conocimiento de documentos con el objetivo de profundizar en el análisis de textos.

2.2.1 Text Mining

La minería de textos, también conocida como *text analytics* o *text mining*, se define como el proceso de examinar grandes colecciones de documentos con el objetivo de generar nueva información (Cohen et al., 2005). Se utiliza para descubrir información relevante transformando el texto en datos que pueden ser utilizados para posteriores análisis. La minería de textos es el proceso encargado del descubrimiento de información que no existía explícitamente en ningún texto de la colección, pero que surge de relacionar el contenido de varios de ellos.

No debemos confundir la minería de textos con la típica búsqueda web o de palabras clave, la cual ofrece información que ya existía o que coincide con lo solicitado. El objetivo de la minería de texto es descubrir nueva información que puede ser desconocida u estar oculta en el contexto relacionado. Para conseguir extraer esta nueva información, la minería de textos utiliza variedad de metodologías de análisis, como el Procesamiento Natural de Lenguaje (PNL). Este

último permite el tratamiento computacional del lenguaje humano. Esto requiere un proceso de modelización matemática, para que un ordenador, que solo entiende bytes y dígitos, comprenda el lenguaje humano. Es, por tanto, un conjunto de herramientas lingüísticas, que cubren las capas de procesamiento léxico, morfosintáctico y semántico, modelos de *machine learning* y *deep learning*, así como arquitecturas software que permiten combinar en tiempo real los componentes anteriores.

El uso combinado de PNL y *text mining* es una metodología importante para la ciencia de datos. Dado que es imposible leer toda la información existente e identificar lo más importante, la aplicación del *text mining* (usando PNL), lo hace por nosotros. Más que una herramienta de búsqueda, la minería de texto con PNL permite ir mucho más allá, proporcionando información detallada sobre el texto en sí mismo y revelando patrones de entre millones de documentos.

2.2.2 Anotación semántica

En los últimos años, se han desarrollado un gran número de herramientas de anotación semántica para el dominio biomédico, muchas de las cuales han sido resultado de proyectos de investigación. En los siguientes apartados introduciremos las principales herramientas de anotación semántica en el ámbito de la biomedicina.

2.2.2.1 MetaMap

MetaMap (Aronson, 2006) es probablemente el anotador biomédico más conocido y utilizado. Desarrollado por la Biblioteca Nacional de Medicina de EE.UU, se encarga de alinear las menciones de entidades biomédicas del texto de entrada a los conceptos correspondientes en el meta *tesauro* de UMLS (NLM, 2009). Cada anotación incluye una puntuación que refleja el grado de coincidencia con el término o frase biomédica del texto de entrada. El proceso de anotación se puede adaptar de varias formas mediante la configuración de varios elementos del proceso de anotación, como el vocabulario utilizado, los filtros sintácticos aplicados al texto de entrada y la coincidencia entre el texto y los conceptos. Además de la flexibilidad que ofrecen estas opciones de configuración, otro aspecto importante

de MetaMap es su enfoque exhaustivo y basado en principios lingüísticos para los análisis léxicos y sintácticos del texto de entrada. Sin embargo, esta minuciosidad también es la causa de una de las principales debilidades de MetaMap: su largo tiempo de procesamiento y, por lo tanto, su insuficiencia para anotar grandes corpus de documentos. Su otra debilidad radica en que no es capaz de abordar eficazmente términos ambiguos.

2.2.2.2 Mgrep

Mgrep (Shah et al., 2009) consiste en un reconocedor de conceptos con un alto grado de precisión (> 95%) en el reconocimiento de nombres de enfermedades desarrollado por el Centro Nacional de Informática Biomédica Integrativa (NCIBI) de la Universidad de Michigan. Mgrep implementa una estructura de datos novedosa basada en árbol de *radix* que permite una comparación rápida y eficiente de textos con un conjunto de términos del diccionario.

2.2.2.3 El anotador del National Center for Biomedical Ontology (NCBO)

El anotador NCBO del Centro Nacional de Ontología Biomédica de EE.UU. (NCBO) es un servicio web disponible gratuitamente. Su proceso de anotación se realiza en dos fases. La primera fase se basa en una herramienta de reconocimiento de conceptos que utiliza un diccionario para identificar menciones de conceptos biomédicos en el texto de entrada. En particular, el anotador NCBO hace uso de la herramienta MGrep, que fue elegida sobre MetaMap debido a su mejor desempeño en varias dimensiones. El diccionario para esta etapa de anotación se construye extrayendo nombres de conceptos y descripciones de ontologías biomédicas y/o tesauros relevantes para el dominio del corpus que se anotará. En la segunda etapa, el conjunto inicial de conceptos, denominado anotaciones directas, se amplía utilizando la estructura y la semántica de las ontologías biomédicas relevantes. Por ejemplo, las medidas de distancia semántica se utilizan para ampliar las anotaciones directas con conceptos relacionados semánticamente; el cálculo de la distancia semántica es configurable y puede basarse, por ejemplo, en la distancia

entre los conceptos en la ontología. Las relaciones semánticas entre conceptos de diferentes ontologías, establecidas a través de alineamientos, sirven como otra fuente para encontrar conceptos relacionados semánticamente que pueden usarse para ampliar el alcance de las anotaciones directas.

2.2.2.4 SIFR Annotator

El SIFR Annotator (Tchechmedjiev et al., 2018) consiste en un servicio web de anotación basado en ontologías de acceso público para procesar datos de textos biomédicos en francés. El servicio, desarrollado durante el proyecto de indexación semántica de recursos de datos biomédicos franceses (2013-2019), se incluye en el SIFR BioPortal, una plataforma abierta para alojar ontologías y terminologías biomédicas francesas basadas en la tecnología desarrollada por el NCBO. El portal facilita el uso y fomento de ontologías ofreciendo un conjunto de servicios (búsqueda, alineamientos, metadatos, versionado, visualización, etc.) incluso con fines de anotación.

2.2.2.5 GoPubMed

GoPubMed (Doms y Schroeder, 2005) fue un buscador basado en conocimiento para textos biomédicos. Fue uno de los primeros buscadores de la Web 2.0, desarrollado por la Universidad Técnica de Dresde en Alemania. Esta herramienta consistía en un buscador basado en ontologías, por lo que las palabras clave de búsqueda nos devolvían los resultados clasificados según MeSH (Medicine, National Library of Medical Subject Headings, 2003) y Gene Ontology (Doms y Schroeder, 2005). GoPubMed utilizaba los modelos de datos MeSH y GO, ya que representaban gran parte del conocimiento estructurado en el ámbito médico, por lo que nuestras búsquedas proporcionaban el concepto asociado y que se correspondía al definido en uno de estos dos modelos.

En cada búsqueda realizada en GoPubMed, además del listado de artículos encontrados, se proporcionaban características del proceso de búsqueda como la localización geográfica de los autores (aspecto interesante para los expertos clínicos a la hora de contactar con ellos y poder obtener más detalles), el número de artículos por año, etc. Así,

esta potente herramienta proporcionaba un buen ejemplo de la inteligencia de los mecanismos de búsqueda.

A diferencia del BioPortal, GoPubMed era un servicio que no permitía la descarga de su contenido ni de las anotaciones que realizaba. Actualmente ya no se encuentra disponible, pero fue un recurso muy importante a la hora de realizar la evaluación de nuestro trabajo de investigación.

2.2.2.6 BioLark-CR

El reconocedor de conceptos Bio-LarK (Groza et al., 2015) se ha desarrollado como parte del proyecto SKELETOME, con el objetivo inicial de realizar la anotación automática de fenotipos esqueléticos en los resúmenes clínicos de los pacientes. Posteriormente, se amplió para habilitar el fenotipo CR utilizando HPO. Bio-LarK CR utiliza un enfoque de recuperación de información para indexar y recuperar conceptos de HPO, combinado con una serie de técnicas de lenguaje para permitir la normalización y descomposición de términos (por ejemplo, variación léxica de token). Además del CR estándar, el sistema puede descomponer y alinear términos conjuntivos, así como reconocer y procesar fenotipos no canónicos, que se alinearían con los mismos términos que en el ejemplo anterior. Esto se logra a través de un enfoque de coincidencia de patrones eficiente que utiliza reglas elaboradas manualmente sobre la estructura superficial de la oración. El reconocimiento de fenotipos no canónicos es una característica opcional de Bio-LarK CR y puede activarse o desactivarse sujeto al uso intencionado del sistema.

2.3 MACHINE LEARNING

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial (IA) que crea sistemas que aprenden automáticamente (Hoffman, 2001). Esto quiere decir que identifica patrones complejos en un gran conjunto de datos y que el sistema mejoran de forma autónoma con el tiempo, sin intervención humana. Se dice que un agente aprende cuando su desempeño mejora con la experiencia, es decir, cuando la habilidad no estaba presente entre sus rasgos de nacimiento. Los modelos o programas resultantes deben ser

capaces de generalizar comportamientos e inferencias para un conjunto más amplio (potencialmente infinito) de datos.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos o reconocimiento del habla y lenguaje escrito, por ejemplo. Algunos sistemas de aprendizaje automático intentan eliminar toda necesidad de intuición o conocimiento experto de los procesos de análisis de datos, mientras otros tratan de establecer un marco de colaboración entre el experto y el ordenador. De todas formas, la intuición humana no puede ser reemplazada en su totalidad, ya que el diseñador del sistema ha de especificar la forma de representación de los datos y los métodos de manipulación y caracterización de los mismos.

Aunque el aprendizaje automático no es un concepto nuevo, en los últimos años tuvo un auge exponencial en su uso y en su aplicación. Una de las razones principales de este auge es el aumento de la capacidad de procesamiento de los ordenadores y la disminución de los costes del mismo, permitiendo así que pueda estar al alcance de todos.

A continuación introduciremos el *clustering* y el Word2Vec, dos métodos englobados dentro del aprendizaje automático.

2.3.1 Clustering

El *clustering* o clusterización es un método de aprendizaje no supervisado dentro del *machine learning* (Li y Wu, 2012). Este método consiste en el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como clústeres. Cada clúster dentro de un grafo está formado por una colección de objetos o datos que a términos de análisis resultan similares entre si, pero que poseen elementos diferenciales con respecto a otros objetos pertenecientes al conjunto de datos y que pueden conformar un clúster independiente. De esta forma, el *clustering* sirve para segmentar datos en grupos de dimensiones similares en base a determinadas.

Los procesos de *clustering* poseen un nivel de dificultad importante, debido a que dependiendo de los criterios y las reglas que diseñemos para generar el clúster, este será eficiente o no para el objetivo que deseemos lograr. En primer lugar, para realizar un proceso de *clustering* debemos definir el número de agrupaciones que debemos

hacer en el conjunto de datos. Posteriormente debemos definir las formas de los grupos de similitudes y asignar un *centroide* de donde partirá el recorte o el agrupamiento. Al definir estos parámetros debemos establecer un margen de error para empezar a definir los clústeres de nuestro conjunto de datos.

Definiendo una métrica o un nivel de error dentro del modelo podemos delimitar niveles aceptables de fallo. Al determinar el error general del modelo este margen de error debe ser incorporado al algoritmo de entrenamiento de aprendizaje automático. Desde este momento debemos procurar generar una especie de bucle que repita el proceso miles de veces en poco tiempo para encontrar todas las combinaciones de errores que pueden existir en el modelo.

Este proceso se debe repetir de forma continua hasta que el algoritmo pueda entender por completo los posibles errores que se generan dentro de los conjuntos de datos.

A continuación se define uno de los modelos de *clustering* más utilizados: el algoritmo *k-means*.

2.3.1.1 Algoritmo de k-means

K-means es tal vez el método clásico para aplicar y entender el proceso de creación de clústeres. Se establece un número de grupos previamente determinado. En este caso el algoritmo buscará los mejores *centroides* para realizar el agrupamiento, de manera que los miembros de cada grupo estén lo más cerca posible de sus *centroides*. El algoritmo funciona de forma iterativa, actualizando el centro de los clústeres de manera que se reduzcan las distancias entre los miembros de cada clúster y su centro.

2.3.2 Word2Vector

Un modo de representar las relaciones entre términos son los *Vector Space models* (VSMs). Los VSM representan las palabras como un vector en un espacio multidimensional, de forma que las palabras similares o relacionadas se encuentren representadas por puntos cercanos. De esta forma, capturamos información semántica, puesto que, por ejemplo, palabras como “rojo”, “negro” y “blanco” se encontrarán en una misma zona de ese espacio multidimensional y lo

mismo pasaría con palabras como “león”, “tigre”, “leopardo”. Un modelo particularmente eficiente desde el punto de vista computacional es Word2Vector (Hu et al., 2019). Este modelo se encuentra disponible de dos formas: *Continuous Bag-of-Words* (CBOW) o el modelo Skip-Gram. Skip-Gram suele funcionar mejor que *CBOW* (Henry et al., 2018).

Dado un conjunto de frases o *corpus*, el modelo Skip-Gram analiza las palabras de cada sentencia y trata de usar cada palabra para predecir qué palabras serán vecinas. Por ejemplo, a la palabra “caperucita” le seguirá “roja” con más probabilidad que cualquier otra palabra. Una vez obtenido el vocabulario generado con las palabras del corpus, el objetivo es entrenar la red neuronal con las sentencias del corpus para obtener, para cada palabra, la probabilidad de que otra palabra del vocabulario sea su vecina. Una vez entrenada la red, el algoritmo devuelve los pesos de cada término en forma de vector que está intentando aprender (de ahí el nombre “*word 2 vector*”, convertir una palabra en un vector). “*Window size*” o tamaño de la ventana es el parámetro que determina el número de palabras de cada contexto. Para entrenar la red neuronal se utilizan pares de palabras encontradas en los datos de entrenamiento, aunque la selección de los pares de palabras dependerá del tamaño de la ventana.

2.4 TÉCNICAS DE ANÁLISIS Y REPRESENTACIÓN DE LOS RESULTADOS

En esta sección introducimos el método de análisis de co-ocurrencia de palabras que utilizamos en esta tesis. Además, hablaremos del análisis de *burst*, muy útil en el estudio de tendencias.

2.4.1 Análisis de redes de co-ocurrencia

El análisis de redes sociales o de co-ocurrencia se inició en los años setenta del siglo pasado, con el objetivo de examinar la estructura de las relaciones entre entidades sociales utilizando diferentes herramientas gráficas, matemáticas y estadísticas, que se habían empleado originalmente. en sociología. Si bien el análisis de redes sociales se centró inicialmente en individuos, grupos, empresas u organizaciones, se extendió rápidamente a la capacidad de diseminación de ideas, enfermedades o influencias, procesos de

circulación en mercados o redes de tráfico. La alta aplicabilidad del análisis de redes sociales se debe a la simplicidad de sus elementos básicos, facilitando casi cualquier modelo relacional. Este modelo, aunque aparentemente simple, permite la aplicación de análisis complejos que nos ayudan a comprender la ejecución interna de la red en su conjunto, así como el comportamiento individual de las entidades de la red.

El uso del análisis de redes sociales para detectar tendencias en la investigación de dirección de proyectos, por ejemplo, se basa en una revisión de las palabras clave elegidas por los autores para fines de indización o los presentes en los resúmenes o referencias bibliográficas. Por lo tanto, el elemento básico de estos enfoques es la palabra clave, y cada artículo se caracteriza por unos pocos de estos vinculados entre sí. Muchas palabras clave se duplicarán en diferentes documentos sobre PM, y es esta reiteración la que permite representar el dominio de PM mediante redes de palabras clave. El conjunto de palabras clave que están más fuertemente relacionadas entre sí que con las otras palabras clave en la red describe un tema. Cada tema se caracteriza por una palabra clave central y las relaciones estadísticas con las otras palabras clave de la red.

Para visualizar y trabajar con la red se dispone de software específico a tal fin, como Gephi que introduciremos en el siguiente apartado.

2.4.1.1 Una herramienta para el análisis de redes: Gephi

Gephi es un software de código abierto de análisis y visualización de redes desarrollado por estudiantes de la *University of Technology of Compiègne* (UTC) en Francia. Este programa ha sido utilizado en proyectos académicos de investigación, periodismo y, por ejemplo, para visualizar la conectividad global del contenido del *New York Times* y para examinar el tráfico de red de *Twitter*. Es ampliamente utilizado dentro del campo de las humanidades digitales, una comunidad donde muchos de sus desarrolladores están involucrados.

A continuación, se indican los principales conceptos utilizados en la aplicación del análisis de redes con Gephi:

- **Grado**. Número de arcos u aristas conectados al nodo.

- **Grado ponderado**. Número de arcos u aristas conectados al nodo calculado sobre mayor grado identificado sobre la red.

- **Excentricidad**. La excentricidad de un nodo V es el máximo de los costes de todos los caminos de coste mínimo con destino V.

- **Centralidad**. La centralidad mide la influencia de un nodo dentro de la red.

- **Centralidad de vector intermedio**. Es la medida que cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos.

- **Page Rank** (rango de página). Se trata del famoso algoritmo utilizado por Google para asignar un ranking de relevancia a los documentos indexados por un motor de búsqueda. Este algoritmo permite dar un valor numérico (ranking) a cada nodo de un grafo para medir su conectividad.

- **Modularidad**. Define la capacidad de la red para descomponerse en comunidades o subredes. Toma valores entre 0 y 1, siendo el 1 el valor indicativo de que la red se puede dividir en grupos de forma clara y sencilla.

- **Coeficiente de agrupamiento**. Medida que cuantifica qué tanto está de interconectado un nodo con sus vecinos. Si el coeficiente tiene un valor pequeño indica un vértice poco agrupado en la red.

Una de las características clave de este software es la capacidad de mostrar el proceso de especialización, con el objetivo de transformar la red en un mapa, y *ForceAtlas2* (Jacomy et al., 2014) es su algoritmo de diseño predeterminado. Este algoritmo ha sido desarrollado por el equipo de Gephi como una solución integral para las redes típicas de los usuarios (sin escala, de 10 a 10.000 nodos). *ForceAtlas2* intenta integrar diferentes técnicas como la simulación de *Barnes Hut*, la fuerza de repulsión dependiente del grado y las temperaturas adaptativas locales y globales. El diseño y las características de este algoritmo proporcionan una forma genérica e intuitiva de representar las redes, y su implementación de velocidades locales y globales adaptables ofrece buenos rendimientos para redes de menos de 100.000 nodos, al tiempo que mantiene un diseño continuo (sin fases ni parada automática).

2.4.2 Análisis de "burst"

Este tipo de análisis consiste en analizar un flujo de documentos y encontrar características cuyo comportamiento sea "explosivo" ("*bursty*"), es decir, ocurren con alta intensidad durante un período de tiempo limitado. El análisis utiliza un autómata probabilístico cuyos estados corresponden a las frecuencias de palabras individuales (Kleinberg, 2003). Para cada palabra individual se calcula su secuencia de estado más probable a lo largo del curso de una secuencia, observando simplemente su frecuencia. Las transiciones de estado corresponden a puntos en el tiempo alrededor de los cuales la frecuencia de la palabra cambia significativamente, es decir, al principio o al final de una "ráfaga" en el uso de la palabra.

El resultado de este análisis es una lista clasificada de las ráfagas de palabras más significativas en el flujo de documentos, junto con los intervalos de tiempo en que ocurrieron. Esto puede servir como un medio para identificar temas o conceptos que se destacaron a lo largo del tiempo, se discutieron activamente durante un período de tiempo y, más tarde, se desvanecieron nuevamente.

CAPÍTULO 3
ANOTACIÓN SEMÁNTICA DE INFORMES CLÍNICOS DE PACIENTES

Hoy en día la gran mayoría de la población busca información sobre salud en internet y los expertos en el ámbito biomédico ven esta vía también como una gran fuente de actualización de sus conocimientos (Tao et al., 2013). Sin embargo, mucha información de atención médica consiste en su mayoría en narraciones no estructuradas, por lo que extraer la información relevante de forma rápida, eficiente y precisa es todo un desafío. Los motores de búsqueda tradicionales como, por ejemplo, los buscadores web, utilizan palabras clave para indexar los documentos y devolver a los usuarios aquéllos de interés acorde a sus búsquedas. Actualmente, los motores de búsqueda han mejorado mucho sus algoritmos con el objetivo de incrementar la precisión de sus búsquedas y la rapidez con la que devuelven sus resultados. Uno de los algoritmos más conocidos es PageRank[3], producto de Google, que ha sentado las bases del indexado actual. El buscador de Google devuelve sus resultados en forma de pequeños fragmentos de la información contenida para cada documento resultante de la búsqueda, lo cual es de gran ayuda al usuario. Sin embargo, los motores de búsqueda suelen devolver cientos de miles de enlaces, muchos de los cuales no son relevantes para la búsqueda concreta de los usuarios. Además, muchas veces esta información aún no es lo suficientemente precisa ya que los usuarios aún deben revisar los resultados obtenidos hasta que localizan la información de interés

En las últimas décadas, la comunidad biomédica ha recurrido a ontologías (y terminologías) para convertir la información disponible en formatos semi-estructurados y no estructurados en conocimiento

[3] https://patents.google.com/patent/US6285999

formalizado como mecanismo para abordar el problema de la integración de datos. El uso de ontologías simplifica la creación de anotaciones semánticas, que se pueden utilizar como índices semánticos de datos para la extracción y recuperación de la información médica en formato textual. Una anotación no es más que un enlace entre un concepto de una ontología y una pieza de texto, que puede ser parte de un artículo, un experimento, un ensayo clínico o una historia clínica, por ejemplo. Un enfoque para facilitar la búsqueda y la consulta de texto biomédico es convertir el texto sin formato en una red de datos anotada, es decir, convertir los datos que originalmente se encontraban en texto libre en formatos estructurados. En el dominio biomédico, las anotaciones deben ser precisas, lo que requiere conocimiento e identificación del contexto, y automáticas, ya que la anotación manual lleva mucho tiempo debido al gran volumen de texto que debe procesarse. En este campo se han realizado muchos esfuerzos, desde el reconocimiento de entidades (NER) hasta la extracción de información, en los que la mayoría de los casos se centran en un pequeño conjunto de categorías para ser anotadas, como persona, ubicación, organización, etc. Por ello, a menudo se requiere de un vocabulario especial que generalmente se beneficia de un conjunto limitado de plantillas lingüísticas predefinidas.

Los esfuerzos para formalizar la información biomédica han dado lugar a muchas ontologías biomédicas interesantes, como *Gene Ontology (GO)* (Harris, 2004) y *SNOMED CT* (Dhombres y Bodenreider, 2016). La anotación semántica basada en ontologías para documentos biomédicos es necesaria para captar información semántica importante, mejorar la interoperabilidad entre sistemas y permitir la búsqueda semántica en lugar de la búsqueda de texto sin formato (Vanteru et al., 2008). Además, proporciona una plataforma para la comprobación de coherencia, soporte de decisiones, etc. Existen algunos sistemas de anotación basados en ontologías como SemTag (Dill et al., 2003), DBpediaSpotlight o Wiki Machine. Cada uno de ellos está dedicado a una ontología particular, por ejemplo, el catálogo de entidades TAP de Stanford para SemTag y DBpedia Lexicalization Dataset2 para DBpediaSpotlight.

En este capítulo de la tesis doctoral proponemos un método para realizar la anotación semántica de casos clínicos de pacientes documentados y publicados en revistas sobre una enfermedad rara en la que la existencia de documentación está más acotada. La enfermedad de estudio es la "*cerebrotendinous xanthomatosis*" o CTX. Lo que pretendemos es proporcionar un método automático que reconozca anomalías fenotípicas a partir de la información que describe casos clínicos de pacientes. Para cumplir con este objetivo, hemos utilizado la ontología HPO, ya que posee una fuente de conocimiento sobre anomalías fenotípicas conocida y fiable en enfermedades humanas. Además, pretendemos crear una sub-ontología para la enfermedad de estudio a partir de la ontología dada, como una nueva fuente de conocimiento precisa acerca de la enfermedad en concreto.

En el momento de desarrollo de este trabajo de investigación se tomó la decisión de construir un nuevo anotador debido a las carencias de las herramientas de anotación existentes en aquel momento. La única opción online disponible era el anotador NCBO. El NCBO está basado en el *Mgrep* (Shah et al., 2009), un reconocedor de conceptos con un alto grado de precisión (>95%) en la identificación de los términos para describir las enfermedades. Desarrollado por el *National Center for Integrative Biomedical Informatics* (NCIBI), *el anotador NCBO* implementa una estructura de datos basada en árbol que permite un reconocimiento rápido y eficiente de un conjunto de términos del diccionario sobre el texto. A pesar del uso de *Mgrep*, el anotador del NCBO sufría de varios problemas: no pre-procesaba los términos ni tenía en cuenta los sinónimos; por lo que su precisión y *recall* eran bajos para nuestros casos de uso.

Teniendo en cuenta estas limitaciones, necesitábamos implementar un nuevo anotador, y su implementación se planteó como un reconocedor de conceptos basado en los metadatos textuales de ontologías expresadas en el lenguaje OBO. Dicho anotador lo denominamos OBO Annotator. El objetivo era mejorar la precisión y *recall* de los anotadores disponibles en el momento de implementación, aprovechando al máximo la terminología y relaciones existentes en la ontología de uso. Para probar nuestra propuesta, evaluamos su funcionamiento en el dominio clínico de la enfermedad rara CTX para

anotar resúmenes de casos clínicos disponibles en PubMed, con el objetivo de mejorar el acceso a las descripciones fenotípicas disponibles públicamente sobre pacientes con este enfermedad rara.

A continuación, en este capítulo, comenzaremos introduciendo el diseño del OBO Annotator y su relevancia en el trabajo de investigación. Abordaremos también los retos principales que afrontamos en la anotación semántica en los casos clínicos de pacientes. Seguidamente introduciremos los aspectos metodológicos utilizados, junto a los resultados obtenidos con nuestro método y la discusión de los resultados, comparando los resultados obtenidos con los trabajos previos en el área. Finalmente, mostraremos las principales conclusiones obtenidas de este capítulo.

3.1 OBO ANNOTATOR

El OBO Annotator es un reconocedor de conceptos diseñado e implementado para anotar anomalías fenotípicas de la literatura biomédica. Dada la importancia actual de la ontología HPO en la descripción de anomalías fenotípicas, utilizamos esta ontología para probar la validez de nuestro anotador. Aún así, nuestro anotador se puede aplicar para reconocer términos de cualquier ontología OBO, ya que es principalmente un reconocedor de entidades, que compara el texto de entrada con los términos de cualquier ontología OBO, y no depende del conocimiento específico contenido en la ontología.

Hoy en día es muy frecuente el uso de la *Human Phenotype Ontology* (HPO) para codificar fenotipos. Esta ontología se ha desarrollado principalmente para ofrecer un recopilatorio de manifestaciones de enfermedades humanas para el análisis computacional. Además, sus desarrolladores la mantienen actualizada y sus actualizaciones se distribuyen de forma regular. El reconocimiento de conceptos que utiliza HPO tiene un inmenso potencial para extraer automáticamente información de grandes cantidades de registros de pacientes existentes o ensayos controlados. Sin embargo, reconocer fenotipos representa un desafío, en gran parte debido a la gran variabilidad léxica y sintáctica con la que aparecen en los textos biomédicos. Para mitigar el problema, nuestro grupo desarrolló el anotador bautizado con el nombre de OBO Annotator. Al

mismo tiempo que trabajábamos en nuestro anotador, otro grupo de investigación ajeno al nuestro desarrolló el anotador Bio-LarK CR[4] (Groza et al., 2015). Ambos anotadores hacen uso de la HPO para reconocer fenotipos clínicos.

Para optimizar el rendimiento del anotador, el tiempo de ejecución y el espacio de memoria requerido, en una primera fase de pre-procesado, nuestro anotador indexa todos los conceptos de la ontología OBO considerada, para poder reconocerlos después, durante la anotación, más rápidamente en el texto. En concreto, la ontología considerada se pre-procesa automáticamente para obtener un diccionario de conceptos en el que se recopila el término tal y como aparece en la ontología, la raíz ("*stem*") de la que procede el término, y el identificador en la ontología. De esta forma, se reducen los tiempos de identificación de términos al tener precargada la ontología en un formato más eficiente.

Por otro lado, durante la anotación, OBO Annotator permite la equiparación rápida y eficiente de texto frente al conjunto de términos del diccionario pre-procesado de la ontología. Para ello, hace coincidir secuencias de texto de hasta un número específico de palabras con los índices léxicos pre-procesados de la ontología. En el caso de coincidencia de la secuencia de texto con algún término del diccionario, se almacena la pieza de texto reconocida, su posición en el texto, así como la anotación al concepto correspondiente en la ontología. En la figura 3.1 se resume el proceso de anotación realizado por el OBO Annotator. Inicialmente, se eliminan las palabras "comunes" del texto: conectores, artículos o determinantes, ya recopiladas en un fichero bajo el nombre de "*stopwords*". En la figura 3.1 vemos cómo se eliminan las "*stopwords*" del texto considerado. A continuación, se divide el texto de entrada, ya pre-procesado, en fragmentos más pequeños. Para ello, se utiliza una "ventana" que se va deslizando en el texto para extraer secuencias de palabras, que se ajustan a los índices léxicos. Por defecto, esta ventana tiene un tamaño de longitud cuatro, esto es, que se seleccionan un máximo de hasta cuatro palabras consecutivas. Este valor por defecto se estableció empíricamente, al analizar la ontología utilizada, en la que se observó que el uso de conceptos formados por

[4] http://www.bio-lark.org/cr_restapi.html

más de 5 palabras, que no eran conectores o determinantes, era infrecuente. Hay que tener en cuenta que el uso de esta "ventana" viene limitado por los signos de puntuación. Es decir, si el anotador encuentra un punto o una coma, solo selecciona las palabras hasta el signo de puntuación, aunque ello suponga coger un menor número de palabras, dado que no tiene sentido intentar buscar términos que pertenecen a sentencias diferentes. En la figura 3.1 ejemplificamos el uso de la "ventana" de 4 unidades léxicas o tokens.

En una segunda fase, los términos seleccionados con la "ventana" se procesan, es decir, se les aplica el algoritmo de *Stemming* para extraer la raíz o lexema del término, y luego se buscan en el diccionario de la ontología OBO (ya comentado en el paso previo). Para ello, además, se aplican las permutaciones de las palabras consideradas en la "ventana", lo que genera las diferentes variantes de términos que se comprueban en el diccionario de la ontología. Se excluyen de las permutaciones construcciones imposibles, como que el adjetivo aparezca después del sustantivo, ya que la construcción no existe en inglés. En el caso de que no haya una coincidencia exacta, la secuencia se divide en subsecuencias más pequeñas, que se comparan. Por ejemplo, la secuencia "*brain [and] cerebellar atrophy*" no se puede combinar con ningún término en el diccionario. Por lo tanto, el algoritmo lo corta en subsecuencias más pequeñas como "*brain atrophy*" o "*cerebellar atrophy*", las cuales sí existen en el diccionario. Además, puede darse el caso de reconocer cadenas más largas o más cortas de palabras, más o menos específicas. En el caso de que se produzcan estas anotaciones superpuestas, siempre nos quedaremos con la más específica. Por ejemplo, en la figura 3.1, el anotador OBO nunca genera anotaciones como "*seizures*" cuando también reconoce "*febrile seizures*" en el mismo texto. En este caso anotamos el concepto más específico.

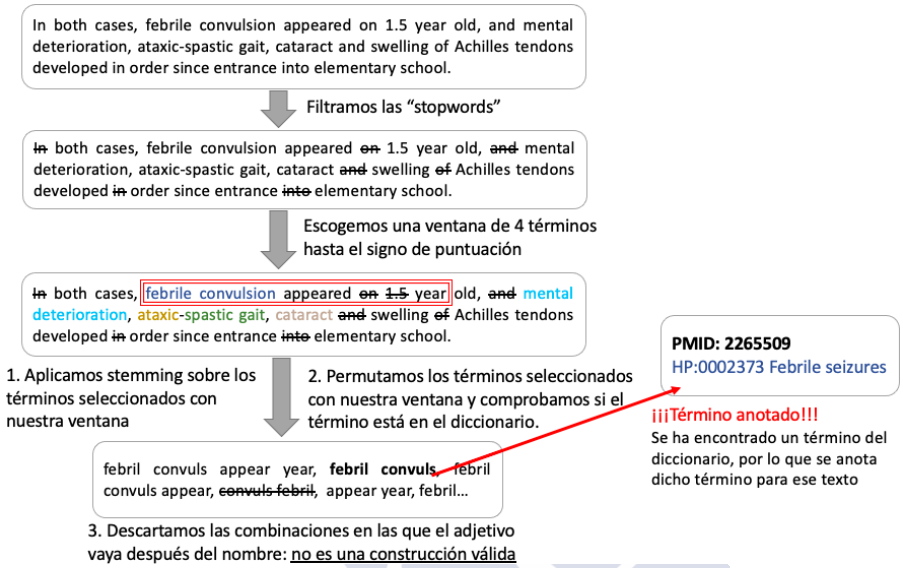

Figura 3.1: Funcionamiento del OBO Annotator.

3.2 RETOS PRINCIPALES

En este apartado se introducen los retos que abordaremos en este trabajo de investigación acerca de las limitaciones existentes en las herramientas de anotación semántica existentes en el momento de realización de este estudio.

Los dominios clínicos han acumulado una amplia experiencia y conocimiento en las últimas décadas. Uno de los principales repositorios de literatura científica es PubMed (McKenzie, 1996). Desafortunadamente, se trata de un repositorio de textos en lenguaje natural, lo que dificulta la búsqueda automatizada, el análisis y la integración de datos de pacientes. Los principales problemas en el uso de PubMed son:

1. *Recuperar todos los documentos relevantes de una temática*. Un gran desafío en el uso de la información de PubMed es la recuperación automática de resúmenes o documentos relevantes para la consulta. Para aliviar este problema, PubMed indexa los artículos utilizando la

terminología estándar *Medical Subject Headings* (MeSH) (Lu et al., 2009), lo que facilita su búsqueda en temas específicos. Además, PubMed proporciona filtros para limitar la búsqueda al seleccionar diferentes criterios, como el tipo de artículo y las fechas de publicación. En nuestro caso, podíamos haber limitado nuestra búsqueda filtrando por "*case reports*". Sin embargo, no todas las publicaciones que contienen casos clínicos relevantes han sido clasificadas por sus autores como "*case reports*" en PubMed y, por lo tanto, este procedimiento no era adecuado. Este es un gran inconveniente en enfermedades raras, donde el número de publicaciones es limitado y, por lo tanto, es crucial recuperar el mayor número posible de casos clínicos .

2. *Acceder a la información relevante de los casos recuperados*. Una vez que se han recuperado los informes de los pacientes, es fundamental contar con buenas herramientas para realizar búsquedas y hacer las preguntas pertinentes de manera automática y eficiente. El motor de búsqueda utilizado por PubMed presenta los resultados ordenados por orden descendente del número de identificación de PubMed, que es un modo primitivo de recibir información. Algunos enfoques en el momento de realizar esta investigación mitigaban el problema al organizar la información recuperada con el apoyo de alguna ontología.

3.3 METODOLOGÍA DEL PROCESO DE LA ANOTACIÓN SEMÁNTICA

En la presente sección detallaremos la metodología seguida durante el proceso de anotación semántica de los casos clínicos de pacientes diagnosticados con CTX, abarcando las diferentes fases realizadas, así como los métodos aplicados en cada una de ellas para cumplir con los objetivos marcados.

En la figura 3.2 se representan las etapas seguidas en el presente trabajo de investigación sobre la anotación semántica de casos clínicos

de pacientes. En primer lugar, se seleccionaron los casos clínicos de pacientes para la enfermedad considerada, la "*cerebrotendinous xanthomatosis*" (CTX), realizando una búsqueda avanzada por palabras clave en el repositorio de información seleccionado (PubMed). Dichos casos clínicos fueron indexados para facilitar su recuperación. En segundo lugar, de los resúmenes o *abstracts* obtenidos en el primer paso, se extrajeron automáticamente los fragmentos relevantes, es decir, aquellos párrafos que hacían alusión a anormalidades fenotípicas. A continuación, estos fragmentos de texto fueron anotados usando el reconocedor de conceptos desarrollado, el OBO Annotator. Finalmente, se analizaron los resultados obtenidos, calculando la precisión y la exhaustividad (*recall)* de nuestro método y se definió un marco comparativo (*benchmark*) para medir el rendimiento del OBO Annotator frente a otras alternativas existentes en aquel momento, el anotador del NCBO y GOPubMed. Adicionalmente, se extrajo la sub-ontología mínima a partir de HPO para generar el conocimiento jerarquizado para nuestra enfermedad rara de estudio.

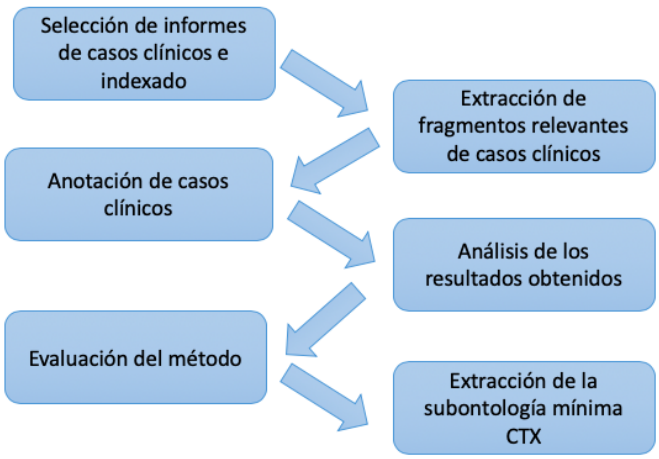

Figura 3.2: Etapas del desarrollo del proyecto de anotación semántica de informes clínicos de pacientes.

3.3.1 Selección de informes de casos clínicos

Dado que PubMed es uno de los principales referentes de documentación biomédica disponible, fue la fuente de datos

seleccionada para la obtención de toda la información relevante sobre la enfermedad rara de nuestro trabajo, la CTX. Para ello, se utilizaron los filtros avanzados disponibles en la aplicación web, que generaron la siguiente consulta:

((((((((((((Cerebral Cholesterinosis[Title/Abstract]) OR Cerebral Cholesterinoses[Title/Abstract]) OR Cerebrotendinous Xanthomatoses[Title/Abstract]) OR Cerebrotendinous Xanthomatosis[Title/Abstract]) OR Xanthomatoses, Cerebrotendinous[Title/Abstract]) OR Xanthomatosis, Cerebrotendinous[Title/Abstract]) OR Cerebrotendinous cholestanolosis[Title/Abstract]) OR Cerebrotendinous cholestanoloses[Title/Abstract]) OR Van Bogaert-Scherer-Epstein Disease[Title/Abstract]) OR Bogaert-Scherer-Epstein Disease, Van[Title/Abstract]) OR Disease, Van Bogaert-Scherer-Epstein[Title/Abstract]) OR Van Bogaert Scherer Epstein Disease[Title/Abstract]) OR Sterol 27-hydroxylase deficiency[Title/Abstract]) OR Cerberotendinous xanthomatosis[Title/Abstract]

3.3.2 Extracción de la información relevante

Una vez extraídos los resúmenes o *abstracts* de todos los artículos disponibles en PubMed sobre CTX (figura 3.3), obtuvimos los fragmentos relevantes de los casos clínicos de pacientes, con la ayuda de un conjunto de patrones lingüísticos que detallamos más adelante. Acto seguido, los anotamos con la ontología HPO. Después de filtrar algunas anotaciones incorrectas, se generó una sub-ontología mínima específica de CTX del conjunto completo de anotaciones, que facilita y agiliza la búsqueda de la información basada en fenotipos clínicos.

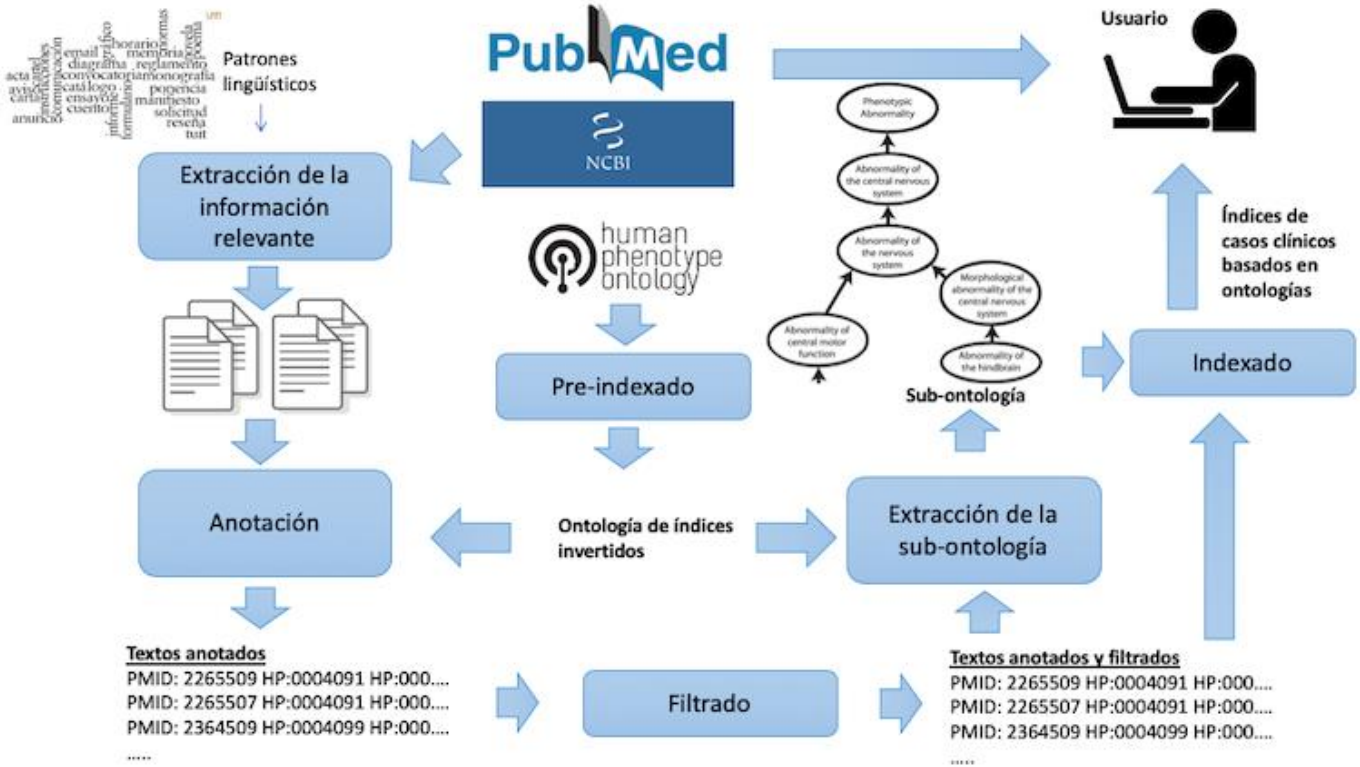

Figura 3.3: Anotación semántica e indexación de casos clínicos procedentes de PubMed.

Como es habitual en los textos técnicos, la literatura sobre informes de casos clínicos utiliza un conjunto limitado de estructuras lingüísticas para organizar y fortalecer el dominio léxico, lo que reduce la ambigüedad en la comunicación. Bajo la premisa de que la descripción del paciente en los informes clínicos generalmente tiene una configuración modular, fácilmente identificable, diseñamos una metodología simple para encontrar estas estructuras que denominamos *patrones lingüísticos*. Primero, se seleccionó al azar un conjunto reducido de *abstracts* del conjunto completo de casos sobre CTX. Luego, estos *abstracts* se analizaron para identificar las estructuras utilizadas en la descripción de los casos clínicos. Ejemplos de estos patrones son los siguientes:

In the present [study\report] we [reviewed\examined ...] [<an age>] [patients\male ...].
A case [study\report] on a [[<an age>] [patient\male ...] is [described \ presented ...]

Con los patrones lingüísticos identificados, elaboramos una propuesta de patrones base que nos permitiesen buscar frases que los cumpliesen en otro subconjunto diferente de informes sobre CTX. El resultado de esta segunda búsqueda nos permitió ajustar las estructuras base para obtener los patrones que finalmente se utilizaron. Únicamente las estructuras con una elevada tasa de éxito y bajo ruido fueron las seleccionadas como patrones válidos.

A continuación, implementamos este conjunto de patrones para extraer los fragmentos relevantes de los *abstracts*. Para ello, el algoritmo busca la primera aparición de cualquier patrón de los definidos dentro del *abstract*, analizando las oraciones que lo componen de forma secuencial. La figura 3.4 muestra un ejemplo de extracción de un fragmento relevante de un *abstract* de PubMed.

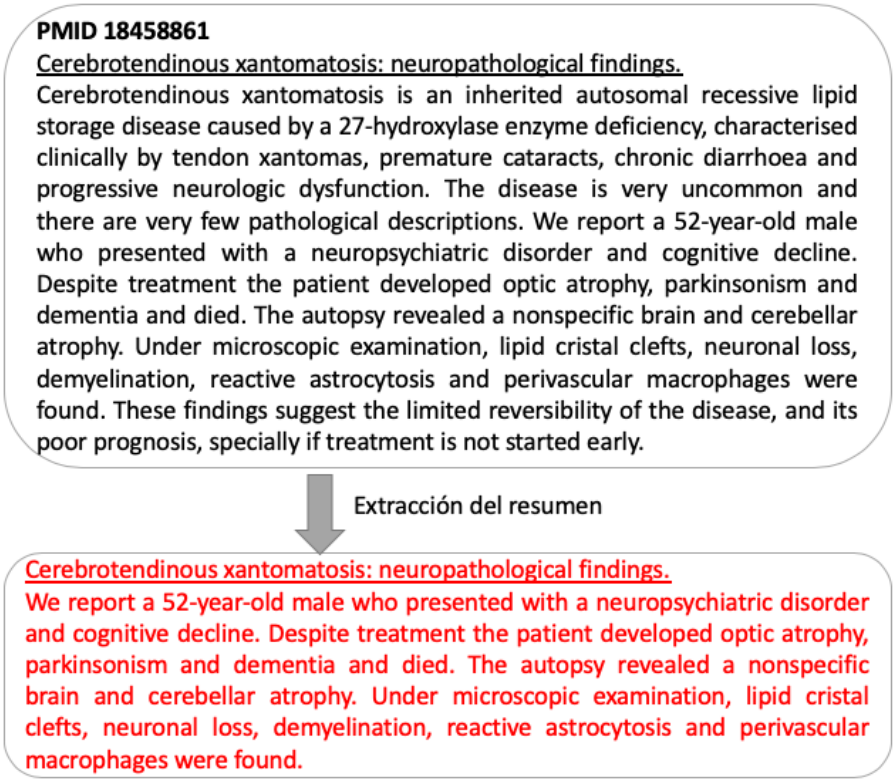

Figura 3.4: Ejemplo de cómo obtener el extracto de información relevante de un *abstract*.

3.3.3 Anotación de casos clínicos de pacientes

Las anotaciones de HPO sobre los *abstracts* considerados para nuestro estudio se crearon utilizando nuestro propio anotador llamado OBO Annotator. También usamos el anotador proporcionado por el Bioportal de NCBO (Whetzel et al., 2013) y GoPubMed (Doms y Schroeder, 2005) para comparar el funcionamiento de nuestra herramienta.

3.3.4 Filtrado de anotaciones incorrectas

Con el objetivo de verificar la viabilidad del OBO Annotator, antes de implementar el sistema completo, se realizó una prueba de concepto con una versión inicial. La evaluación de esta prueba fue realizada por

dos neurólogas del equipo de trabajo: M. Pardo y M.J. Sobrido. Durante la evaluación, identificaron un error recurrente: la palabra "*xanthomatosis*" siempre estaba erróneamente anotada con el término HPO "*cerebrotendinous xanthomatosis*" (id: 0000991). Por lo tanto, decidimos agregar un nuevo paso a nuestro método para eliminar estas anotaciones incorrectas y recurrentes. Este paso se desarrolló como un filtro, que elimina la anotación HP: 0000991 cuando está vinculada a esa cadena. Este filtro en particular es específico para el dominio CTX. Además, se programó un filtro explícito para el anotador NCBO que consistió en eliminar las anotaciones superpuestas generadas por el NCBO. Este tipo de filtro, sin embargo, es independiente del dominio clínico. Finalmente, el algoritmo eliminaba tanto las anotaciones repetidas como las anotaciones generales en presencia de otras específicas.

Como ejemplo de anotación del OBO Annotator, la parte superior de la figura 3.5 muestra las anotaciones directas reconocidas por el anotador en el resumen de PubMed con referencia PMID 2265509, que se reducen al conjunto mínimo de fenotipos mostrado en la parte inferior. Las anotaciones directas "*cerebellar ataxia*" o "*tremor*" (en la parte superior) se eliminaron del conjunto relevante de anotaciones (en la parte inferior), ya que existen otras anotaciones más específicas ("*gait ataxia*" y "*resting tremor*", respectivamente).

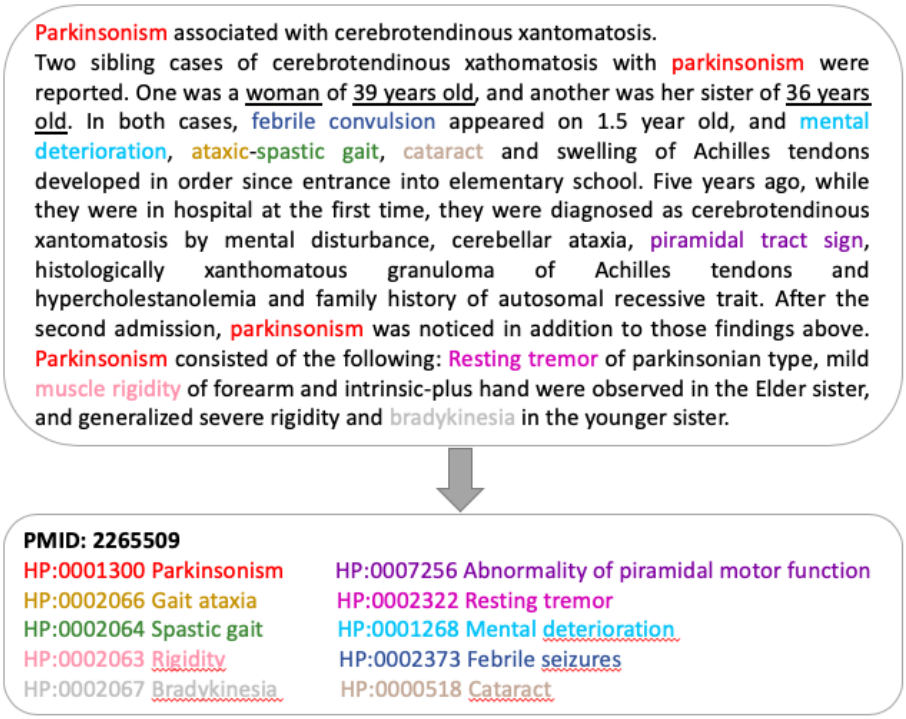

Figura 3.5: Ejemplo de anotaciones generadas para un abstract utilizando el OBO Annotator y la ontología HPO.

3.3.5 Extracción de la sub-ontología mínima

Con el objetivo de filtrar conceptos de la ontología que no son relevantes para el conjunto de informes de casos clínicos analizados, nos propusimos construir la sub-ontología mínima. La sub-ontología resultante es una ontología reducida en la que se incluyeron todos los términos extraídos de los resúmenes de PubMed considerados junto con los conceptos padre de éstos (es decir, todos los conceptos de la ontología desde los términos extraídos hasta el concepto raíz).

3.4 Evaluación

Evaluar la calidad o el rendimiento de nuestro método es realmente difícil debido a la falta de un estándar contra el que comparar nuestros resultados (nuestro "*gold standard*"). Además, la realización de revisiones manuales requiere mucho tiempo y es propensa a errores. En

vista de esta situación, sugerimos una evaluación centrada en uno de los posibles escenarios de aplicación, como es la anotación basada en los rasgos fenotípicos de nuestra enfermedad rara. Por otro lado, también proponemos realizar una evaluación que pueda destacar los beneficios de usar un proceso de anotación automático frente a las versiones actuales de anotaciones validadas por expertos.

3.4.1 Evaluación frente a documentos etiquetados como "*case reports*"

Con el objetivo de reducir la carga de trabajo vinculada al procedimiento de revisión pero sin disminuir su calidad, se utilizaron dos conjuntos de datos: 1) el conjunto completo de artículos sobre CTX disponibles en PubMed y, 2) el conjunto reducido de artículos CTX etiquetados como "case reports". Teniendo en cuenta estos dos conjuntos de artículos, por un lado identificamos el porcentaje de resultados comunes y no comunes obtenidos por nuestro método automatizado sobre el primer conjunto y, por otro, el mismo porcentaje obtenido manualmente por las neurólogas sobre el conjunto de "case reports" (al que llamaremos “método manual”).

Las dos neurólogas del equipo de trabajo verificaron manualmente el título y los *abstracts* de los documentos CTX etiquetados como "*case reports*", clasificándolos como correctos o incorrectos (cuando el resumen no describía los casos de pacientes con CTX). También verificaron manualmente los documentos propuestos por nuestro método que no estaban etiquetados como "*case reports*", y los clasificaron como relevantes para describir un caso clínico de CTX.

Dentro de este contexto, definimos la precisión como la fracción entre el número de documentos correctos y el número total de documentos propuestos por cada método (manual y automatizado); y el *recall* como la fracción entre el número de documentos correctos propuestos por cada método y el número total de documentos relevantes. Con el objetivo de comparar sistemas, una forma estándar de combinar estas dos medidas en la recuperación de información es la *F-measure*, que es una media armónica ponderada de la precisión y el *recall*.

$$F - measure = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

3.4.2 Evaluación de la relevancia de la anotación

Realizamos la evaluación de nuestro método de anotación de dos formas. Primero, evaluamos manualmente la precisión y el *recall* de la anotación automática, revisando 50 *abstracts* elegidos al azar de los fragmentos extraídos por nuestro método. En segundo lugar, comparamos automáticamente las anotaciones obtenidas con el conjunto de anotaciones de fenotipos ya existentes sobre CTX proporcionadas por el consorcio de HPO (Köler et al., 2014).

3.5 RESULTADOS

En los siguientes apartados se analizarán los resultados extraídos con nuestra metodología.

3.5.1 Selección de informes de casos clínicos y extracción de los fragmentos relevantes

Nuestro conjunto de datos incluyó 515 resúmenes seleccionados de PubMed correspondientes a artículos con la palabra clave "*cerebrotendinous xanthomatosis*" en el título y en el *abstract*, y un subconjunto de 223 resúmenes limitados a informes de casos clínicos. Para extraer estos informes de casos clínicos con una búsqueda en PubMed basta con seleccionar en los filtros de búsqueda el tipo de documento que se requiere, en nuestro caso, "*case report*".

Durante el proceso de anotación sólo se tuvieron en cuenta el título y los fragmentos de cada resumen de los resultados devueltos por PubMed, que se asociaron directamente al caso del paciente. Al restringirnos a la ontología específica de HPO, la anotación se centró en el dominio de los fenotipos humanos en enfermedades neurogenéticas.

De un conjunto de 50 *abstracts* seleccionados al azar del total de los 223 *abstracts* de informes de casos clínicos, se utilizó una lista provisional de estructuras base o patrones para buscar frases de muestra en otro subconjunto diferente a los 50 resúmenes. Se diseñó un conjunto de cinco patrones lingüísticos (ver Apéndice A) a partir de las

estructuras base con la tasa de éxito más alta y el ruido más bajo. Este conjunto de patrones diseñados se usó para extraer los fragmentos relevantes del conjunto completo de datos de 515 *abstracts*. El resumen del proceso completo se puede ver en la figura 3.6.

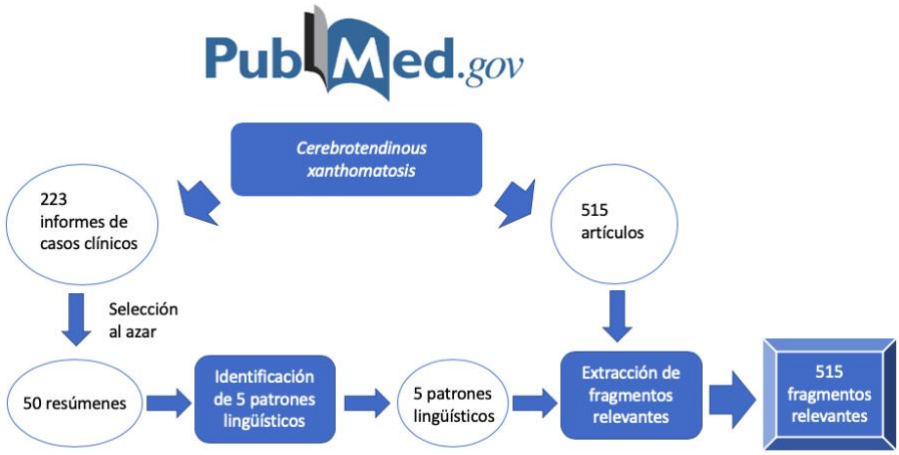

Figura 3.6: Extracción de fragmentos relevantes de los artículos sobre CTX.

La tabla 3.1 muestra los resultados de la evaluación al aplicar el método de revisión manual y el automático. En la tabla 3.1, podemos ver que nuestro enfoque automático logró una precisión del 99% en comparación con el 97% del método manual. Los valores de *recall* fueron significativamente más bajos: 65 frente al 81% del método manual.

Medida	Método manual	Método automático
Número de artículos seleccionados	223	174
Precisión (%)	97	99
Recall (%)	81	65
F-measure (%)	88	78

Tabla 3.1: Evaluación del rendimiento en la identificación de casos clínicos de pacientes.

Con un *recall* significativamente menor que el método manual, el rendimiento de nuestro método nos parecía insuficiente, por lo que analizamos más en detalle los resultados. Al aplicar el método

automatizado a los 223 documentos seleccionados mediante el método manual, nuestro método solamente seleccionó 124 *abtracts* (55%) etiquetados como "*case reports*" (figura 3.7). Además, identificó otros 50 documentos que no se habían etiquetado como "case reports". Revisando los 99 artículos seleccionados por el método manual, pero no por nuestro método, identificamos 56 artículos sin *abstract* disponible (solo estaba el título disponible), 8 artículos que no describían los casos de CTX y otros 35 artículos correctamente etiquetados como "*case reports*". Debido a que nuestro método se basaba en el procesamiento de los *abstracts* completos, los 56 artículos sin resumen disponible no se pudieron identificar automáticamente. Por lo tanto, decidimos agregar automáticamente estos "*case reports*" sin resumen disponible en PubMed a los resultados del método automatizado, y lo llamamos "método combinado". De esta manera, el número total de documentos seleccionados fue de 230 (frente a 223 del método manual).

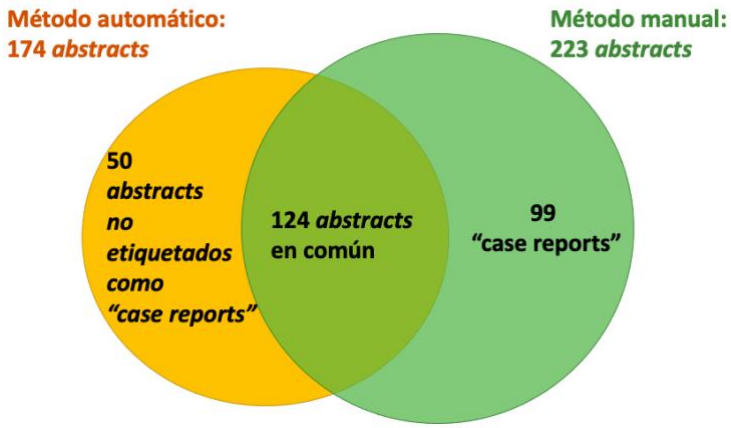

Figura 3.7: Diagrama de Venn para mostrar el solapamiento de los dos métodos.

Una vez más, las dos neurólogas del equipo de trabajo evaluaron manualmente los resultados. El método combinado logró una precisión del 99% en comparación con el 97% para el método manual y un *recall* del 87% frente al 81% del método manual (tabla 3.2).

Medida	Método manual	Método automático	Método combinado
Número de artículos seleccionados	223	174	230
Precisión (%)	97	99	99
Recall (%)	81	65	87
F-measure (%)	88	78	93

Tabla 3.2: Evaluación del rendimiento del proceso de identificación de casos clínicos de pacientes para los tres métodos comentados.

3.5.2 Relevancia de la anotación: Calidad de las anotaciones

Se realizaron tres pruebas diferentes utilizando (i) el OBO Annotator con HPO, (ii) el anotador NCBO con HPO y (iii) el servicio proporcionado por GoPubMed. Las primeras dos pruebas se ejecutaron en 230 fragmentos de resúmenes, que incluyeron 174 extraídos por nuestro método más 56 fragmentos etiquetados como "*report cases*" que carecían de resumen (y por ello, sólo se procesó el texto del título del artículo). Estas pruebas nos permitieron comparar los anotadores OBO y NCBO, ya que se utilizaron con la misma ontología (HPO). Como no pudimos asegurarnos de que los dos anotadores estuvieran usando la misma versión de HPO, cada vez que se identificaba una disparidad entre los dos anotadores, verificamos manualmente si el concepto anotado por un anotador podía ser anotado por el otro usando el mismo sinónimo. De esta forma, nos aseguramos que los dos anotadores estaban usando la misma versión de HPO, al menos, para el dominio de la CTX.

De los 230 fragmentos de artículos (véase tabla 3.3), el método reconoció como mínimo un concepto en el *abstract* o en el título. El anotador OBO reconoció 145 términos (63%) y el anotador NCBO 126 (55%). En promedio, el OBO Annotator anotó 3,3 conceptos por resumen con una desviación estándar de 2,56, mientras que el anotador de NCBO anotó 2,9 conceptos con una desviación estándar de 2,05. El número máximo de conceptos por resumen que reconoció el anotador OBO fue de 11, mientras que en el anotador NCBO fue de 8. Finalmente, el número total de anotaciones reconocidas por cada anotador se puede ver en la última fila de la tabla 3.3: 456 (el anotador OBO) frente a 344 (el anotador NCBO). En total, hubo 326 (71%) superposiciones, 18 (4%) diferencias y 112 (25%) anotaciones OBO

adicionales, en comparación con las anotaciones de NCBO. El 25% de anotaciones adicionales y el 4% de anotaciones específicas reconocidas por el anotador OBO se deben a las siguientes mejoras: nuestro anotador reemplazó tokens para los lexemas correspondientes (13,4%), dividió secuencias de texto en sub-secuencias (10,2%), utilizó sinónimos de HPO relacionados (en la ontología, como "*related*") además de sinónimos exactos (2,9%) y eliminó palabras comunes y signos de puntuación (2,6%).

También estudiamos la tasa de falsos positivos vinculada a las características incorporadas en el OBO Annotator. Únicamente el 17% de las anotaciones OBO adicionales y más específicas fueron incorrectas debido a sinónimos incorrectos incluidos en HPO (7,5%), el uso de raíces en lugar de tokens (4,3%), el uso de variaciones de términos (3,2%) y la división de secuencias de texto en sub-secuencias (2,1%).

Resultado de anotar	OBO Annotator	Anotador de NCBO (BioPortal)
Número de *abstracts* anotados	145	126
Porcentaje de *abstracts* anotados (%)	63	55
Media de número de conceptos por *abstract*	3,30	2,90
Desviación estándar	2,56	2,05
Número máximo de conceptos por *abstract*	11	8
Número total de anotaciones	456	344

Tabla 3.3: Resultados de anotación para el OBO Annotator y el anotador de NCBO.

Además, no todos los resúmenes fueron anotados. Esto se debe, en parte, al hecho de que algunos de ellos no tenían *abstracts* disponibles, es decir, solo el título estaba disponible, y otros no describían los fenotipos del paciente en el *abstract*. Sin embargo, aunque ambos anotadores son capaces de reconocer muchos de los signos sistémicos y neurológicos, principalmente en el caso del OBO Annotator, no son capaces de reconocer las características fisiológicas y neurofisiológicas, así como algunas anomalías morfológicas y bioquímicas. En general, las anomalías obtenidas de las pruebas realizadas en el laboratorio no se describen utilizando un solo término

estándar, sino con construcciones terminológicas que incluyen diferentes aspectos, tales como el lugar de ocurrencia, el tipo o grado de la lesión y la técnica utilizada. Así, un algoritmo basado en el reconocimiento del nombre por sí mismo no es suficiente para anotar el conjunto completo de anomalías fenotípicas.

Para la tercera prueba, como no teníamos acceso al anotador utilizado por GoPubMed, anotamos manualmente 50 resúmenes aleatorios de los obtenidos en el portal web de GoPubMed. Cabe señalar que GoPubMed tenía en cuenta tanto las anotaciones hechas automáticamente a través de los resúmenes, como las anotaciones manuales hechas por los usuarios de PubMed. En nuestro experimento, solamente tuvimos en cuenta las anotaciones automáticas para garantizar las mismas condiciones en los otros anotadores. La comparación de los resultados con esta tercera prueba fue útil para mostrar el impacto del uso de diferentes ontologías.

En la tabla 3.4 mostramos una pequeña comparación de los resultados, con tres mediciones para realizar la evaluación: el promedio de conceptos por resumen (cobertura), precisión y *recall* en un total de 50 *abstracts* elegidos aleatoriamente. En promedio, 3,86 conceptos por resumen fueron reconocidos por el anotador OBO, 3,14 por el anotador NCBO y 2,54 por el anotador GoPubMed. Además, el OBO Annotator alcanzó una precisión del 94%, un valor ligeramente inferior al de los otros anotadores (97%). Pero la principal diferencia entre nuestro anotador y los otros fue el *recall* y la *F-measure*. El *recall* de nuestro método fue considerablemente mayor: 61% en comparación con el 49 % (el anotador de NCBO) y el 41% (el anotador de GoPubMed); así como la *F-measure*: 74% frente a 65% y 58%, respectivamente. Como hemos comentado anteriormente, las dos razones principales para las anotaciones OBO adicionales y, por lo tanto, mayor *recall*, fueron la sustitución de tokens para sus lexemas correspondientes y la consideración de partes de las secuencias de texto (subsecuencias). Aunque no sabíamos exactamente cómo funcionaba el anotador de GoPubMed, los resultados mostraban la misma tendencia (lexemas y subsecuencias) que los de la comparación entre el anotador NCBO y el OBO Annotator. Además, la cobertura y *recall* de GoPubMed fue inferior a las de los anotadores OBO y NCBO, ya que se basaba en una

terminología diferente, que no era específica para el dominio del fenotipo humano.

Medida	OBO Annotator	Anotador de NCBO (BioPortal)	PubMed
Cobertura	3,86	3,14	2,54
Precisión (%)	94	97	97
Recall (%)	61	49	41
F-measure (%)	74	65	58

Tabla 3.4: Evaluación de los resultados obtenidos mediante nuestro método, el anotador de NCBO y el servicio de PubMed.

3.5.3 Relevancia de la anotación: superposición con las anotaciones "validadas"

Se denominan anotaciones validadas (o "*curated*") a aquellas anotaciones que han sido validadas y aprobadas por expertos en el dominio. Para analizar la relevancia de las anotaciones, comparamos las anotaciones "*curated*" y las obtenidas automáticamente a través de las ontologías obtenidas para estos dos casos. La ontología generada por anotaciones automáticas incluyó 324 conceptos de HPO, mientras que la ontología validada sólo 137 conceptos. En total, ambas ontologías comparten 121 conceptos HPO. Ambas ontologías están disponibles en http://www.usc.es/keam/PhenotypeAnnotation/ .

3.6 Discusión

Convencionalmente, la investigación clínica se ha centrado en enfermedades relacionadas con la mayor parte de la población de pacientes. La comprensión científica actual de la biología humana a nivel molecular ha dado la bienvenida al estudio de las enfermedades a un nivel más individual. Con el objetivo de desarrollar tratamientos más específicos e individualizados, establecer grupos más pequeños de enfermedades que compartan características comunes es todo un desafío. Las enfermedades raras pueden jugar un papel importante como herramientas para descubrir los mecanismos fundamentales de las enfermedades[5].

[5] http://www.findacure.org.uk

Proporcionar herramientas computacionales orientadas a extraer automáticamente fenotipos de grupos de pacientes que compartan características comunes puede facilitar en gran medida el estudio de la enfermedad. En particular, la indexación semántica facilita la síntesis y el filtrado de la información de múltiples fuentes de rápido crecimiento. Sin embargo, hoy en día la anotación semántica se logra principalmente de forma manual. Por lo tanto, automatizarlo para administrar el enorme volumen de información que se encuentra disponible diariamente es todo un desafío (Tsatsaronis et al., 2012).

Nuestro trabajo de investigación se centró en la indexación semántica de un dominio en particular: los informes de casos clínicos de la literatura. Nuestros resultados confirman que es posible extraer fragmentos relevantes de información de *abstracts* de casos clínicos de pacientes revisados por pares y extraídos de la literatura médica. Nuestras técnicas generan un índice semántico de datos de pacientes no identificados, que podría migrarse y analizarse con métodos más específicos si fuera necesario.

3.6.1 Hallazgos y significado de la selección de informes de casos

Nuestros resultados mostraron que al utilizar el conjunto de patrones lingüísticos basados en las regularidades observadas en los *abstracts* clínicos, pudimos identificar 50 informes de casos más que no se habían etiquetado como tales (figura 3.7). Este resultado es crucial en las enfermedades raras, donde el número de casos es limitado, y es importante recuperar la mayor cantidad posible. Una de las principales ventajas de nuestro método es la alta precisión para identificar automáticamente nuevos casos relevantes, con la posibilidad de utilizarlo como un complemento a la identificación manual de los informes de casos. Sin embargo, la contribución más significativa es el reconocimiento de los fragmentos de código relevantes para la anotación, ya que este hecho causa un aumento directo de la precisión de la anotación.

Aunque hemos demostrado en este trabajo que el conjunto logrado de patrones lingüísticos es un recurso valioso para reconocer los fragmentos relevantes en el dominio de CTX, no podemos decir que

estos mismos patrones podrían ser apropiados en otros dominios de enfermedades raras. Como paso de validación preliminar, probamos estos patrones en 50 extractos seleccionados al azar de la enfermedad de Huntington y en 50 extractos de la ataxia de Friedreich. En el primer caso, se logró una precisión del 95% y un *recall* del 67%, mientras que en el segundo caso, se logró una precisión del 99% y un *recall* del 25%. Como conclusión, nuestro conjunto de patrones lingüísticos no se podría usar directamente en otros dominios de enfermedades raras, aunque podrían ser válidos como patrones.

3.6.2 Calidad de la anotación

A diferencia de la mayoría de los otros trabajos sobre búsqueda semántica, nos centramos en evaluar el proceso de anotación, ya que es el núcleo de las herramientas que utilizan las ontologías para la exploración de la literatura. En este contexto, nos gustaría resaltar que cuando realizamos este trabajo no había estudios definitivos que mostrasen la calidad de la anotación basada en ontologías, excepto para la ontología GO (Skunca et al., 2012).

Obviamente, nuestros resultados mostraron que la calidad de los resultados de búsqueda depende tanto de la eficacia de extraer fragmentos de información relevantes, como del mecanismo de anotación. Los servicios suministrados por NCBO y GoPubMed nos ofrecieron, en el momento de realización de la evaluación, un punto de referencia para medir la mejora proporcionada por el rendimiento de nuestro método. Los tres anotadores utilizados en nuestro estudio se basaban en el reconocimiento de conceptos. Una evaluación comparativa previa entre MetaMap y Mgrep (Shah et al., 2009) mostró que los reconocedores de conceptos tienen claras ventajas en cuanto a la velocidad, la flexibilidad y la escalabilidad, en comparación con las herramientas de procesamiento de lenguaje natural (PLN).

Una de las principales contribuciones de nuestro trabajo fue la evaluación detallada de los resultados de la anotación, proporcionando medidas precisas de rendimiento además de permitir saber más sobre los límites de los enfoques y cómo podrían mejorarse. La evaluación indicó que la calidad de la anotación fue satisfactoria (74% de *F-measure* cuando se utilizó nuestro anotador OBO; tabla 3.4). Aunque

los detalles de cómo funciona Mgrep no son claros en las publicaciones, y no estamos seguros de si el anotador utilizado por GoPubMed es exactamente el mismo seguido en Doms y Schroeder (2005), las principales diferencias de nuestro anotador son:

- El enriquecimiento del pre-procesamiento léxico de los términos de ontología OBO (*offline*) y texto (*online*).

- La extracción de secuencias y subsecuencias de palabras con la ventana que se va deslizando en el texto preprocesado.

El pre-procesamiento léxico se implementó ajustando algunas piezas de software ya implementadas, como el algoritmo *stemming* de Porter, y algunos recursos disponibles (como las "*stopwords"* o palabras comunes que se eliminaron antes de la etapa de procesamiento, y los adjetivos en inglés). Debido a que este pre-procesamiento no hacía uso de las técnicas de PLN, se conservaron características como la velocidad, la flexibilidad y la escalabilidad, al tiempo que se producía un aumento del 12 y 20% en el *recall* (tabla 3.4), en comparación con los servicios de NCBO y GoPubMed, respectivamente. En total, el 74% de la *F-measure* puede considerarse un buen resultado, ya que el método se aplicó sin utilizar otras técnicas. Hay que tener en cuenta que el pre-procesamiento se diseñó para aplicarse a cualquier ontología OBO y no a las características específicas de la ontología HPO. El pre-procesamiento específico a la ontología HPO aumentaría el *recall*. Por ejemplo, la sustitución de palabras frecuentes de los términos de HPO por sinónimos, como "*abnormality*" por "*lesions*", o el enriquecimiento de los índices léxicos con términos de otras ontologías, mediante el uso de la propiedad OBO 'xref', cuya función principal es establecer asignaciones entre conceptos de diferentes ontologías. Los experimentos preliminares que realizamos con la propiedad xref (para UMLS y MeSH) revelaron que es importante comenzar desde mapeos seleccionados. El primer caso (con xref UMLS sin revisar) dio lugar a numerosos errores, mientras que en el segundo caso (con xref MeSH revisado), el diccionario se enriqueció con algunos nuevos sinónimos.

Uno de los conocimientos adquiridos en nuestro trabajo es que los nombres de los fenotipos no suelen ser más largos que cuatro palabras, y son manejables computacionalmente para realizar una búsqueda completa de todas las posibles permutaciones de cuatro o menos palabras en el texto. Esto es especialmente cierto cuando tratamos de anotar la mayoría de los signos sistémicos y neurológicos. Sin embargo, una de las limitaciones es que el método no es capaz de reconocer las características fisiológicas y neurofisiológicas, así como algunas anomalías morfológicas y bioquímicas. En tales casos, parece más apropiado usar técnicas basadas en co-ocurrencia tras la anotación. Además, este paso posterior también permitiría actualizar la ontología.

3.6.3 Comparación con anotaciones revisadas

Como se mencionó anteriormente (ver sección 3.1.2), la cobertura total de la ontología inducida por las anotaciones automáticas (324 conceptos) es mayor que la de la ontología revisada (137 conceptos). La tabla 3.5 enumera el conjunto de conceptos de HPO generados a partir de las anotaciones reconocidas en la literatura y no presentes en la ontología generada (aunque ésta incluye algún concepto más general). En este caso, solo los fenotipos mencionados en al menos cuatro resúmenes participaron en el estudio comparativo. Por lo tanto, no tuvimos en cuenta los fenotipos poco comunes.

Todas las anotaciones que conducen a los conceptos en la tabla 3.5 se revisaron manualmente, y solo dos casos fueron erróneos (filas 3 y 10). Corresponden a los conceptos de "*congenital cataract*" y "*peripheral demyelination*". El primero tiene erróneamente "*bilateral cataract*" como sinónimo de HPO y el segundo, "*demyelination*". Como resultado, el fenotipo "*bilateral cataract*" siempre se anota como "*congenital cataract*", y "*central demyelination*" se anota como "*peripheral demyelination*", anotaciones que son claramente erróneas. El resto de los 11 conceptos fueron correctos y podrían considerarse como fenotipos candidatos para agregarse en versiones más recientes.

Concepto HPO	¿Anotaciones correctas?	Recomendación
abnormal emotion / affect behaviour	Sí	Sí
chronic diarrhea	Sí	Sí
congenital cataract	No	Revisar sinónimos
gait disturbance	Sí	Sí
global development delay	Sí	Sí
juvenile cataract	Sí	Sí
lower limb spasticity	Sí	Sí
paraplegia/paraparesis	Sí	Sí
parkinsonism	Sí	Sí
peripheral demyelination	Sí	Revisar sinónimos
polyneuropathy	Sí	Sí
progressive neurologic deterioration	Sí	Sí
spastic gait	Sí	Sí

Tabla 3.5: Extracto del listado de conceptos más específicos extraídos de la literatura que no están en la ontología.

La tabla 3.6 muestra el conjunto de conceptos de HPO presentes en las anotaciones revisadas y no inducidas a partir de anotaciones en la literatura. Revisamos manualmente el conjunto de *abstracts*, buscando cualquier término que describiese estos conceptos. No pudimos encontrar nueve fenotipos en los *abstracts*. Tal vez si hubiéramos analizado los artículos completos, hubiéramos anotado estos fenotipos, ya que a menudo se caracterizan en las descripciones clínicas de CTX (Federico et al., 1993). Del resto de conceptos, la razón principal para omitirlos es que los conceptos tienen nombres diferentes en los resúmenes de HPO y PubMed. En algunos casos, hay múltiples formas de expresar el concepto. Por ejemplo, "*abnormality of the dentate nucleus*" también se puede describir como "*lesion*" o "*hyperintensity of the dentate nucleus*". Una vez más, reemplazar las palabras recurrentes de los términos de HPO por sinónimos, como "*abnormality*" por "*lesions*" permitiría que se reconociesen estos términos. En otras casuísticas hay formas más adecuadas de expresar el concepto. Por ejemplo, una forma más sencilla de expresar "*electromyography (EMG): axonal abnormality*" es mediante "*axonal abnormality*". De manera similar, el término HPO "*abnormality of central somatosensory evoked potentials*" es una larga serie de palabras que difícilmente va a

aparecer en los textos. Como se mencionó anteriormente, en los casos en que las anomalías provienen de las pruebas realizadas en el laboratorio, un algoritmo basado únicamente en el reconocimiento de nombres no es suficiente para anotar el conjunto completo de anomalías fenotípicas. En el futuro, se plantea el uso de técnicas basadas en co-ocurrencia para poder anotar estos tipos de fenotipos.

Concepto HPO	¿Está en los *abstracts*?	Razón para la omisión
abnormality of central somatosensory evoked potentials	Sí	Demasiadas palabras
abnormality of the dentate nucleus	Sí	Nombre diferente
abnormality of the periventricular white matter	Sí	Nombre diferente
angina pectoris	No	
cerebral calcification	Sí	Nombre diferente
delusions	No	
developmental regression	No	Concepto diferente
electroencephalography with generalized slow activity	Sí	Demasiadas palabras
EMG: axonal abnormality	Sí	Nombre diferente
hallucinations	No	
limitation of joint mobility	No	
lipomatous tumor	Sí	Nombre diferente
malabsorption	No	
myocardial infarction	No	
respiratory insufficiency	No	
xanthelasma	No	

Tabla 3.6: Extracto de la lista de conceptos que están en la ontología, pero no en la literatura.

3.6.4 Retos pendientes

En nuestro estudio, decidimos utilizar *abstracts* en lugar de artículos completos, ya que más de la mitad de estos últimos no son gratuitos. Además de los disponibles, la mayoría de los documentos necesitan ser transformados a formato PDF. Como consecuencia, 9 de

los 137 conceptos (6,5%) en la ontología generada no se pudieron encontrar en la literatura de casos de CTX. Esta prueba sugiere que se pueden reconocer un gran número de fenotipos relevantes para CTX basándose solamente en los *abstracts*.

3.6.5 Estudios posteriores

Tras la publicación del trabajo realizado con el OBO Annotator surgió otro estudio comparativo en el que se analizó la precisión de nuestra herramienta (Groza et al., 2015). En dicho estudio se presenta otra herramienta de anotación conocida con el nombre de Bio-Lark CR.

Bio-Lark CR utiliza un enfoque de recuperación de información basado en indexar y recuperar conceptos de HPO, permitiendo la normalización y descomposición de términos (por ejemplo, variación léxica de tokens). Además, el sistema puede descomponer y alinear términos conjuntivos (por ejemplo, "*short and broad fingers*" alineados a HP: 0009381 – "*short fingers*" y HP: 0001500 – "*broad fingers*"), así como reconocer y procesar fenotipos canónicos, como "*fingers are short and broad*", que se alinearían con los mismos términos que en el ejemplo anterior. Esto se logra a través de un enfoque eficiente de coincidencia de patrones que utiliza reglas diseñadas manualmente sobre la estructura de la oración. El reconocimiento de fenotipos no canónicos es una característica opcional de Bio-Lark CR y se puede habilitar o deshabilitar en base al uso del sistema.

En el estudio realizado se compararon tres anotadores: el NCBO Annotator, el OBO Annotator (nuestro anotador) y su anotador, el Bio-Lark CR. En las tablas 3.7 y 3.8 se ven los resultados obtenidos para el *corpus gold standard* de HPO y un corpus diseñado específicamente por lo autores para hacer pruebas para este trabajo, respectivamente. El corpus de prueba consiste en 2164 entradas, cada una de las cuales corresponde a un concepto de HPO. Mientras que el *gold standard* de HPO es un conjunto de 228 *abstracts* manualmente anotados y registrados en la base de datos de OMIM.

En la tabla 3.7, se muestran los resultados para el *corpus gold standar*d, en el que el OBO Annotator y el Bio-LarK CR tienen una eficiencia similar, la diferencia en F-Score es sólo de 0,02, a pesar de que el OBO Annotator utiliza una versión antigua de la ontología HPO

con respecto al período de publicación del software Bio-Lark.. La eficiencia, en este caso, del NCBO Annotator es un 10% menor que los otros dos sistemas, algo similar a nuestro estudio.

	Precisión	*Recall*	F1
NCBO Annotator	0,54	0,39	0,45
OBO Annotator	0,69	0,44	0,54
Bio-LarK CR	0,65	0,49	0,56

Tabla 3.7: Rendimiento de los anotadores sobre el corpus de HPO utilizando coincidencia exacta e identificación de conceptos.

En cambio, en la tabla 3.8, se muestran los resultados para el corpus de prueba diseñado por los autores del estudio, en el que el OBO Annotator y el Bio-LarK CR tienen una eficiencia muy dispar, siendo incluso mucho menor la del OBO Annotator frente a los otros dos anotadores considerados. La gran diferencia en el rendimiento de los anotadores ante un *gold standard* (tabla 3.7) y un corpus sintético construido por los propios autores ponen en duda la validez de los resultados de la comparativa de la tabla 3.8.

	Precisión	*Recall*	F1
NCBO Annotator	0,95	0,84	0,89
OBO Annotator	0,54	0,26	0,35
Bio-LarK CR	0,97	0,93	0,95

Tabla 3.8: Rendimiento de los anotadores sobre el corpus de prueba utilizando coincidencia exacta e identificación de conceptos.

3.7 CONCLUSIONES

En el presente trabajo se ha realizado una propuesta de cómo anotar automáticamente fenotipos del conjunto de resúmenes almacenados en PubMed sobre CTX. Aún así, creemos que la metodología propuesta para diseñar el anotador basado en la ontología OBO y evaluar los resultados es lo suficientemente genérica como para aplicarla en la

literatura relacionada con cualquier anomalía fenotípica humana de enfermedades neurogenéticas, ya que el anotador OBO se ha restringido a la ontología específica HPO. Significativamente, evaluamos ampliamente el método y demostramos que cuando los anotadores se configuran correctamente con las ontologías más adecuadas para el dominio, se alcanzan las anotaciones de alta calidad con pocos hallazgos de falsos positivos.

Nuestro enfoque se basa en esta idea. Es una herramienta liviana, con una precisión un poco menor que la del anotador NCBO. Sin embargo, el *recall* se ha mejorado significativamente debido al enriquecimiento del procesamiento previo léxico de los sinónimos de ontología y las cadenas de texto, y al extraer sub-secuencias de palabras en el texto pre-procesado. Aunque el OBO Annotator solo se ha probado en el dominio CTX, la implementación actual se puede implementar en otros dominios de enfermedades neurológicas sin tener que realizar cambios importantes en el código fuente, ya que se basa en la ontología HPO. Los filtros específicos del dominio son los únicos cambios esperados. Además, el anotador OBO es altamente personalizable para ser utilizado con otra ontología OBO diferente.

CAPÍTULO 4
APRENDIZAJE DE SINÓNIMOS EN LA ONTOLOGÍA HPO

A lo largo de los años, se han propuesto diferentes enfoques para ampliar la cobertura de terminologías biomédicas. Las herramientas de anotación basadas en terminologías y ontologías proporcionan una mayor capacidad para reconocer conceptos si la cobertura de las terminologías es amplia y sus conceptos están enriquecidos con una gran cantidad de sinónimos. El enriquecimiento terminológico de una ontología como HPO mejoraría la capacidad de la anotación semántica para extraer la información relevante de fuentes de información biomédica. Es por ello que, en este capítulo de la tesis, nos centraremos en la propuesta de nuevos métodos automáticos para el enriquecimiento terminológico de ontologías, centrados en el caso de uso de HPO.

4.1 INTRODUCCIÓN

En los últimos años, con el objetivo de mejorar el proceso de reconocimiento de conceptos en medicina, se han desarrollado técnicas específicas para generar nuevos sinónimos en los sistemas terminológicos UMLS (Hettne et al., 2010; Bodenreider, 2004) y SNOMED CT (Allones et al., 2014). En estos estudios, la creación de nuevos sinónimos se ha centrado en la sustitución de una o más palabras de los términos de la ontología con sinónimos conocidos. Su principal inconveniente ha sido la generación de una cantidad inmensa de sinónimos candidatos, muchos de los cuales no eran adecuados para el dominio clínico. Además, los métodos no resolvían el problema de los homónimos, ya que reemplazaban los sinónimos sin tener en cuenta el significado original del término.

En el campo de alineamiento de ontologías, existen resultados similares. Dhombres y Bodenreider (2016) han sacado provecho de las

propiedades sintácticas del léxico de HPO y de la estructura lógica de la ontología para descubrir mapeos parciales entre HPO y SNOMED CT. Los autores además han comparado los enfoques léxico-sintáctico y lógico, concluyendo que eran complementarios entre sí. Por su parte, Quesada-Martínez et al. (2015) han propuesto un nuevo método para medir las regularidades léxicas en términos de ontologías biomédicas con el objetivo de descubrir nuevas relaciones entre ellos.

En las últimas dos décadas, diferentes estudios han examinado y aprovechado la estructura compositiva de varias ontologías biomédicas, entre otras, la ontología de genes (GO) y la ontología HPO. No es raro encontrar términos GO que incluyan sus términos principales como sub-cadenas (Ogren y Cohen, 2004; Mungall, 2004; Ogren et al., 2005). Esta propiedad se ha utilizado para aumentar la cobertura de GO, con el reto de mejorar el reconocimiento de las relaciones regulatorias a partir de resúmenes de MEDLINE (Verspoor et al., 2003). Utilizando la naturaleza compositiva de GO, se han identificado patrones sintácticos comunes dentro de GO (Hamon y Grabar, 2008). Este método genera sinónimos como variantes ortográficas o abreviaturas, de la misma manera que las técnicas de sustitución de sinónimos (Hole y Srinivasan, 2000; Huang et al., 2007; Huang et al., 2009; Allones et al., 2014) crean nuevos términos en el paso intermedio.

Un enfoque más reciente, ofrecido por Funk et al. (2016), también basado en la naturaleza compositiva del GO, genera sinónimos aplicando un conjunto de reglas sintácticas y léxicas en los términos constituyentes. Esta técnica de sustitución de sinónimos divide los términos GO en sus componentes y reemplaza estas partes constituyentes con sinónimos GO y variantes derivadas. Mientras que las técnicas de sustitución de sinónimos mencionadas anteriormente (Hole y Srinivasan, 2000; Huang et al., 2007; Huang et al., 2009; Allones et al., 2014) identifican sub-secuencias comunes de palabras compartidas entre pares de sinónimos conocidos, Funk et al. (2006) aplican un conjunto de reglas sintácticas para dividir los términos de la ontología. Además, Funk et al. (2006) producen sinónimos de nivel intermedio mediante la aplicación de reglas de generación de variantes derivadas.

Con el fin de preservar la calidad de GO, independientemente de la técnica utilizada, los términos generados deben seguir convenciones de la expresión de conceptos. Vespoor et al. (2009) han propuesto un método automatizado para garantizar la calidad de la ontología que se basa en la identificación de la aparición de términos que expresan semántica similar con diferentes convenciones lingüísticas. Con respecto a la ontología HPO, algunos términos son frases que utilizan una combinación de entidades y cualidades anatómicas (Gkoutos, et al., 2009). Esta naturaleza compositiva ha brindado la oportunidad de definir lógicamente los términos de HPO, utilizando la estrategia conocida como *Entity-Quality decomposition*. La estrategia se ha aplicado para extraer el mapa conceptual de descripciones de fenotipos de la literatura científica (Groza et al., 2013) e integrar ontologías de fenotipos en múltiples especies (Oellrich et al., 2013). Con el objetivo de mejorar el *recall* en el reconocimiento de fenotipos, Kocbek et al. (2016) proponen construir automáticamente un diccionario de variantes léxicas para las descripciones de fenotipos humanos.

A diferencia de las técnicas descritas anteriormente (Hole y Srinivasan, 2000; Huang et al., 2007; Huang et al., 2009; Allones et al., 2014), que se basan principalmente en las propiedades léxicas de los términos pertenecientes a la ontología, nuestro enfoque también tiene en cuenta la estructura jerárquica de la ontología para generar sinónimos.

4.2 METODOLOGÍA

En este capítulo de tesis nos centraremos en tratar de generar automáticamente nuevos sinónimos de alta calidad para HPO, mediante la identificación de sinónimos comúnmente utilizados en la literatura. El objetivo último de este trabajo es mejorar nuestra capacidad de anotación descrita en el capítulo 3. La figura 4.1 representa esquemáticamente nuestro método de generación de sinónimos. Primero, el método descarta los sinónimos redundantes incluidos en HPO. La redundancia se establece desde el punto de vista del reconocimiento de la entidad nombrada. A continuación, identifica recursivamente todas las superposiciones léxicas en HPO, es decir, todos los pares de términos conectados por una relación jerárquica y

donde el término descendiente incluye el término ascendente como subcadena propia. Este paso aprovecha el cierre transitivo de las relaciones jerárquicas de HPO. Posteriormente, para cada término descendente en cada superposición léxica, el método genera nuevos sinónimos al reemplazar en el término descendente las palabras superpuestas con sinónimos conocidos del término ascendente. A continuación, busca las frases exactas de los sinónimos generados en MEDLINE y descarta aquellas para las que no se haya recuperado ningún resultado. Por último, dado que HPO proporciona diferentes niveles de relación de sinonimia, el método asigna la tipología correspondiente a cada sinónimo. A continuación, describimos en detalle las etapas de nuestro método.

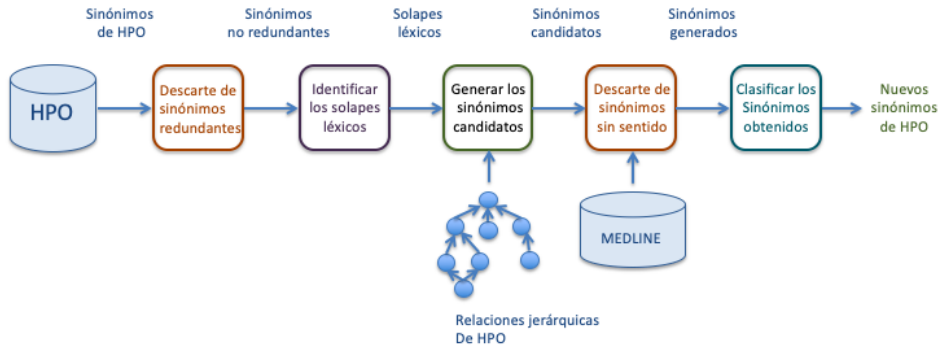

Figura 4.1: Resumen de la metodología aplicada para la generación de sinónimos.

4.2.1 Descartando sinónimos redundantes

En los inicios de nuestro método detectamos que HPO incluía sinónimos redundantes desde el punto de vista de la anotación, lo que llevaba a un rendimiento degradado de nuestro método. Vamos a verlo con un ejemplo. En la figura 4.2 se muestra un extracto de la jerarquía de HPO para el concepto "*hearing impairment*" (HP: 0000365), y en la tabla 4.1 se recogen los sinónimos existentes en HPO correspondientes a cada uno de los nodos de la jerarquía representados en la figura 4.2.

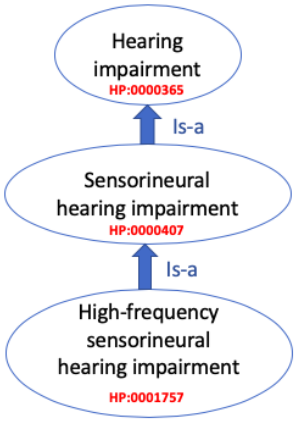

Figura 4.2: Extracto de la jerarquía HPO para el concepto "hearing impairment".

En general, un reconocedor de conceptos que identifica "*congenital hearing loss*" como sinónimo de "hearing impairment" también reconocerá "*hearing loss*", dado que es un sinónimo (Tabla 4.1). Por lo tanto, "*congenital hearing loss*" puede considerarse como un sinónimo redundante desde el punto de vista del reconocimiento de conceptos, y por tanto, puede eliminarse. Por esta razón decidimos eliminar todos los sinónimos redundantes de HPO.

Términos de HPO	Sinónimos de HPO
Hearing impairment	Congenital hearing loss Congenital deafness Deafness Hypoacusis Hearing loss Hearing defect
Sensorineural hearing impairment	Sensorineural deafness Sensorineural hearing impairment
High frequency sensorineural hearing impairment	High frequency sensorineural hearing impairment High-tone sensorineural hearing impairment High-tone sensorineural deafness

Tabla 4.1: Sinónimos de HPO para los términos "*hearing impairment*", "*sensorineural hearing impairment*" y "*high frequency sensorineural hearing impairment*".

4.2.2 Identificando solapes léxicos en HPO

De forma general, un solape léxico entre un par de términos arbitrarios se produce cuando uno engloba a otro como una subcadena. En nuestro trabajo, restringimos los solapes léxicos a un par de términos conectados por una relación jerárquica entre ellos. Por ejemplo, existe un solape léxico entre el término "*hearing loss*" y su término hijo, "*sensorineural hearing loss*". La identificación de este tipo de solapes léxicos es clave en nuestro método para poder restringir la generación de sinónimos.

A continuación describimos el método propuesto para identificar los solapes léxicos restringidos por relaciones jerárquicas. Para cada categoría de nivel superior, se extraen todos los pares de términos de HPO que son solapes léxicos (recogidos en la figura 4.3), desde el nodo raíz de la categoría hasta los nodos inferiores (es decir, el cierre transitivo de las relaciones jerárquicas de HPO). En términos simples, para cada par de términos únicos que se conectan directa o indirectamente entre ellos a través de una relación jerárquica, nuestro método verifica todas las coincidencias entre términos y sinónimos. Por ejemplo, para el par de términos únicos HP: 0000365, "*hearing impairment*", y HP: 0000407, "*high frequency sensorineural hearing impairment*", se identificaron tres solapes léxicos: 1) entre "*hearing impairment*" y "*high frequency sensorineural hearing impairment*" y 2) entre "*hearing impairment*" y "*high-tone sensorineural hearing impairment*", y 3) entre "*deafness*" y "*high-tone sensorineural deafness*".

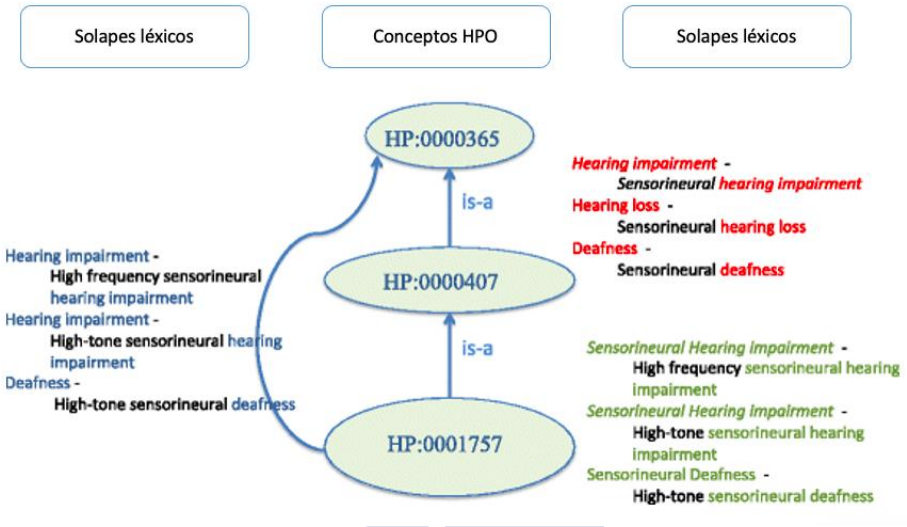

Figura 4.3: Ejemplo de solapes léxicos para términos relacionados jerárquicamente sobre el concepto raíz "hearing impairment".

4.2.3 Generando nuevos sinónimos recursivamente

Para cada solape léxico identificado, el método generó nuevos sinónimos para cada término descendiente. La generación de nuevos sinónimos se llevó a cabo mediante sustitución, es decir, mediante el reemplazo de la subcadena superpuesta en los términos descendientes, utilizando los sinónimos conocidos de los términos antecesores. Veamos esto con un ejemplo. Al reemplazar "*hearing loss*" en "*sensorineural hearing loss*" (ver detalle en la tabla 4.1.) con el sinónimo "*hearing defect*", se obtuvo "*sensorineural hearing defect*" (figura 4.4). Se realiza el mismo proceso en las figuras 4.5, 4.6 y 4.7: se sustituye parte del término por su sinónimo en HPO y se infiere un nuevo sinónimo.

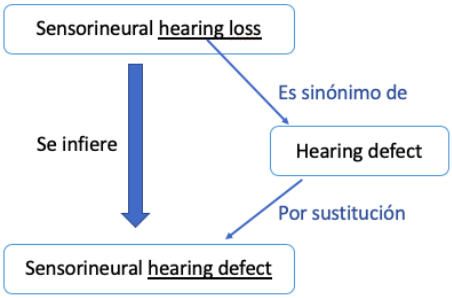

Figura 4.4: Sinónimo obtenido para el término "*sensorineural hearing loss*".

De manera similar, al reemplazar "*hearing impairment*" en "*high-tone sensorineural hearing impairment*" con el sinónimo "*hearing loss*", se generó "*high-tone sensorineural hearing loss*" (figura 4.5).

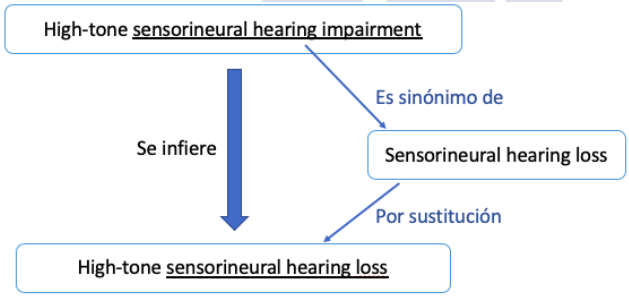

Figura 4.5: Sinónimo obtenido para el término "*high-tone sensorineural hearing impairment*".

De forma análoga en la figura 4.6, como "*deafness*" e "*hypoacusis*" son sinónimos, se sustituye "*deafness*" en "*sensorineural deafness*" y se genera el nuevo sinónimo "*sensorineural hypoacusis*".

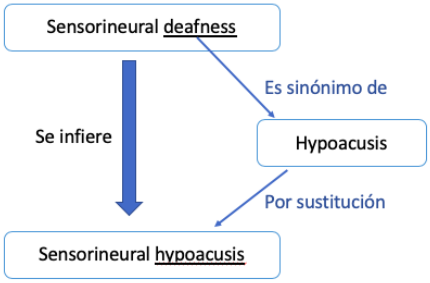

Figura 4.6: Sinónimo obtenido para el término "*sensorineural deafness*".

Finalmente, en la figura 4.7 se identifican como sinónimos de "*sensorineural hearing impairment*" los términos "*sensorineural deafness*", "*sensorineural hearing loss*" y "*sensorineural hypoacusis*" (término ya inferido en la figura 4.6), que por sustitución generan otros tres sinónimos, uno por cada sustitución.

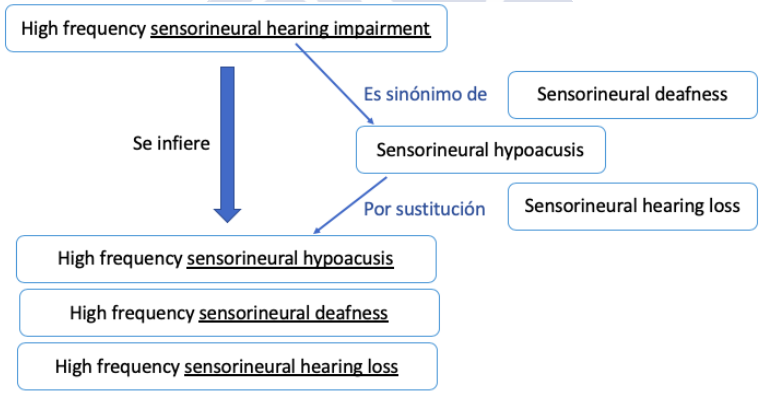

Figura 4.7: Sinónimos obtenidos para el término "high frequency sensorineural hearing impairment".

4.2.4 Descartando sinónimos generados de forma incorrecta

Los pasos anteriores no garantizaron que los sinónimos generados fueran sintácticamente correctos o aceptados en el dominio biomédico. La obtención de términos sin sentido degrada el rendimiento del reconocimiento de la entidad nombrada. Para resolver este problema, decidimos descartar los sinónimos candidatos generados artificialmente. La gran cantidad de publicaciones en MEDLINE, que

se actualizan diariamente y son fácilmente accesibles a través de PubMed, lo hicieron adecuado para verificar los nuevos sinónimos generados de forma rápida, efectiva y precisa. Nuestra hipótesis fue que los sinónimos automáticamente generados que no se encontraban incluidos en ninguna publicación en MEDLINE eran incorrectos. Teniendo esto en cuenta, nuestro método buscó cada uno de los sinónimos generados en MEDLINE, aunque únicamente en los campos de título y resumen. De esta forma, el método no encontró términos como, por ejemplo, "*high frequency sensorineural hypoacusis*", por lo que descartó dicho sinónimo.

4.2.5 Clasificación del tipo de sinónimos

Para cada sinónimo generado, nuestro método obtuvo su tipología teniendo en cuenta el tipo de los términos que originan el sinónimo. En particular, el método obtuvo el tipo más restrictivo de estos términos. En el lado izquierdo de la figura 4.8, se muestra una relación is-a entre "*acute respiratory tract infection*" y "*respiratory tract infection*". El primer término incluye el segundo como una cadena válida. En el lado central, se muestra el conjunto de sinónimos para estos dos términos. El sinónimo "*respiratory infections*" se utilizó para reemplazar la cadena superpuesta, generando el nuevo término "*acute respiratory infections*". Como el tipo del sinónimo sustituido era *related*, el método infirió el tipo *related* para el sinónimo generado, "*acute respiratory infections*", que se muestra en el lado derecho.

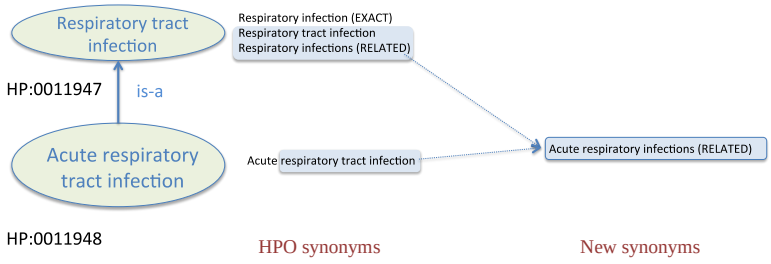

Figura 4.8: Ejemplo del tipo de sinónimos inferidos por nuestro método.

4.3 EVALUACIÓN DEL MÉTODO

Para evaluar nuestro método de generación de sinónimos, medimos el valor de los sinónimos generados de forma extrínseca al evaluar su contribución en la ejecución de un sistema de reconocimiento de conceptos. Específicamente, evaluamos la ejecución de dos aspectos: anotación de conceptos y recuperación de documentos (como ya hicimos en el capítulo 3). Para ello, se utilizaron dos tipos de *corpus* diferentes en la evaluación. El primero es un corpus de 228 resúmenes (Groza et al., 2015) citado por la base de datos en línea de *Online Mendelian Inheritance in Man* (OMIM) (Hamosh et al., 2005) y anotado manualmente por un equipo de tres expertos. Incluye 1933 anotaciones de concepto, que cubren 460 conceptos HPO diferentes (más del 4% de todos los términos únicos). Aunque el conjunto de anotaciones es reducido en relación con el tamaño de la ontología HPO, no hay otro corpus con anotaciones de HPO a nivel de texto. Este corpus se usó como *gold standard* para evaluar la contribución de los nuevos términos en la anotación de conceptos.

En este momento, el desarrollo de HPO no sólo depende de OMIM sino de otros recursos, como la literatura médica. Por lo tanto, nuestro *gold standard* podría no cubrir toda la terminología relevante. Es por ello que decidimos medir la contribución de los nuevos sinónimos en el rendimiento de la recuperación de textos. Para este propósito, preparamos una colección de resúmenes de MEDLINE. Como HPO se usa principalmente en anotaciones de enfermedades hereditarias para permitir cálculos computacionales de gran escala, en los estudios del fenotipo humano, se realizó una búsqueda en PubMed con la palabra clave "*hereditary disease*". En total, se utilizaron 308 resúmenes de los 580 disponibles para nuestra evaluación. Además, calculamos el índice de "*information content"* (IC) de los términos únicos de HPO, en base a las anotaciones seleccionadas proporcionadas por el consorcio HPO (Köhler et al., 2014). El índice IC se cuantifica como la probabilidad negativa de ocurrencia (Resnik et al., 2005):

$$IC = -log_{10}\, p(t)$$

Además, en nuestro trabajo, p (t) fue la probabilidad de que aparezca el término t en las anotaciones seleccionadas.

$$p: T \rightarrow [0,1]$$

siendo T el conjunto de los términos únicos de HPO.

Un término con una puntuación IC baja significa que se está utilizando para anotar muchos conceptos relevantes y debería aparecer con frecuencia en la literatura. En cambio (Funk et al., 2016), los términos con una puntuación IC alta tienen menos probabilidades de aparecer en los textos y, por lo tanto, son más informativos. Debido a esto, los métodos que generan sinónimos con una puntuación IC más alta tendrán un gran impacto en la tarea de reconocimiento de conceptos y, por lo tanto, en la recuperación de documentos. El proceso de evaluación utilizó el OBO Annotator (Taboada et al., 2014), un reconocedor de conceptos orientado a realizar anotaciones automáticas de fenotipos basados en la ontología HPO.

El procedimiento de evaluación consistió en crear dos diccionarios. El primero utiliza el propio HPO como repositorio de sinónimos y el segundo se crea agregando nuevos sinónimos al primer diccionario. Más tarde, el OBO Annotator se utiliza con cada uno de los diccionarios. Sobre estos datos, calculamos la precisión, el *recall* y la *F-measure* para medir la calidad de nuestro método sobre las dos evaluaciones propuestas.

4.4 RESULTADOS

En nuestro trabajo se utilizó la versión disponible en aquel momento de la ontología HPO, la versión publicada el 13/01/2016. Otras fechas importantes y a tener en cuenta en la exposición de resultados son el acceso a MEDLINE a través de PubMed el día 05/05/2016 para filtrar los sinónimos generados y el día 05/05/2017 para generar la recopilación para la evaluación. En los siguientes apartados se analizan los resultados obtenidos para nuestro estudio de sinónimos de HPO.

4.4.1 Superposiciones léxicas de la ontología HPO.

Como ya hemos comentado, cada término en HPO tiene un identificador único, un nombre y una lista de sinónimos. La tabla 4.2 muestra las propiedades principales utilizadas como métricas para las superposiciones léxicas en HPO. En nuestros experimentos, la ontología en formato OBO contenía 11.004 términos únicos. Después

de eliminar 57 términos obsoletos, se tuvieron en cuenta 10.947 términos únicos o términos preferidos (PT). El término preferido es la descripción que se considera más apropiada para expresar un concepto en un registro clínico. Es el sinónimo que se prefiere en un idioma o dialecto. En total, se distribuyeron 18.385 sinónimos en 23 categorías principales representadas por taxonomías. En promedio, hubo 1,68 sinónimos por cada término único. Además, la cantidad de tokens (o palabras) en los que se puede dividir un sinónimo utilizando un carácter de espacio en blanco como delimitador, osciló entre 1 y 12. Sin embargo, el 86% de los sinónimos contenían a lo sumo 4 tokens. En general, 529 sinónimos incluían otros sinónimos del mismo término que las sub-cadenas seleccionadas. Después de la eliminación, nos quedamos con un total de 17.856 sinónimos. Sobre este dato, el número de superposiciones léxicas únicas detectadas en HPO fueron 1.285, que fueron casi el 12% del número total de términos únicos y el 7% del número total de sinónimos.

Propiedad	Número
Total de clases no obsoletas (o términos preferidos, PT)	10.947
Total de etiquetas (PT y sinónimos)	18.385
Número de etiquetas por término/concepto	1,68
Total de solapamientos léxicos diferentes identificados	1.285

Tabla 4.2: Métricas utilizadas en el reconocimiento de los solapes léxicos en HPO.

Para obtener el número de superposiciones léxicas únicas totales, primero las procesamos previamente siguiendo los pasos de a continuación:

- Las palabras con guión fueron divididas en sus palabras constituyentes. Por ejemplo, "*criss-cross atrioventricular valves*" se convirtió en "*criss cross atrioventricular valves*".

- Las palabras o *tokens* entre paréntesis no se contabilizaron, ya que generalmente son aclaraciones o acrónimos y no son adecuadas para soluciones de minería de texto. Por ejemplo, se consideró que "*thyroid stimulating hormone receptor (tshr) defect*" tiene cinco tokens.

Esta etapa de preprocesamiento fue la única parte de nuestro método que involucró la sintaxis especializada de la ontología. En la figura 4.9, podemos ver el número de solapes léxicos únicos desglosados por el número de tokens que incluyeron. Como podría esperarse, cuando el número de tokens se incrementó, el número de superposiciones léxicas disminuyó, excepto en aquellos casos para superposiciones con dos tokens (540 superposiciones con dos tokens frente a 400 solapadas con un solo token). Todas las superposiciones léxicas identificadas se proporcionan como información complementaria[6].

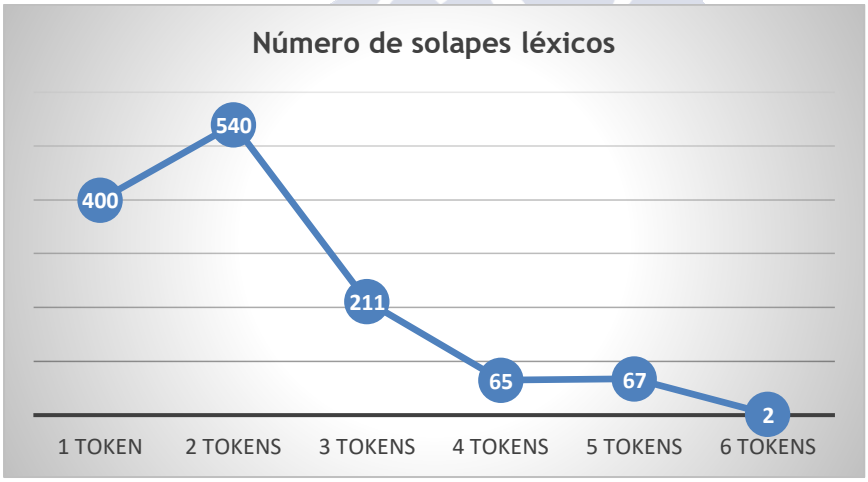

Figura 4.9: Número de solapes léxicos únicos en términos de número de tokens que los componen.

[6] https://nubeusc-my.sharepoint.com/:t:/g/personal/hadriana_rodriguez_rai_usc_es/Ea87-KEaSq5NvvzSTGrKdxQB2czvSUMRSdT-KqQXPjQq2g?e=rvK72b

4.4.2 Generando nuevos sinónimos para la ontología HPO

El número total de sinónimos generados por sustitución fue de 121.594 (ver tabla 4.3), incluyendo 115.630 sinónimos ya existentes en HPO, que fueron eliminados. La diferencia de conjuntos A / B = {x: x ∈ A y x ∉ B} incluye 5.964 sinónimos que representan el 32% del total de sinónimos en HPO.

Método utilizado para la generación de sinónimos	Número de nuevos sinónimos
Grafos de dependencia de etiquetas (A)	121.594
Intersección del conjunto A y los sinónimos originales en la ontología (B)	115.630
Diferencia entre A y B (A/B)	5.964

Tabla 4.3: Número de nuevos sinónimos generados por el método.

4.4.3 Descartando los sinónimos sin sentido

De los 5.964 sinónimos candidatos totales, únicamente 745 de ellos fueron encontrados en MEDLINE por PubMed cuando se buscaron frases exactas[7]. Los sinónimos generados cubren 488 términos HPO únicos. Con respecto al tipo de sinónimo, el 67% de los nuevos sinónimos eran exactos, el 21% eran relacionados y el 12% eran sinónimos sin ningún tipo de relación. El último dato proviene de términos HPO de los cuales no había información sobre la relación de sinonimia.

Después de descartar los sinónimos incorrectos, el número total de sinónimos nuevos fue del 7% del total de términos únicos, el 4% del total de sinónimos y el 58% del total de las superposiciones léxicas. Si comparamos el número total de sinónimos recientemente identificados (745) con el total de superposiciones léxicas (1.285), la proporción fue significativamente mayor (58%).

[7] https://nubeusc-my.sharepoint.com/:w:/g/personal/hadriana_rodriguez_rai_usc_es/EXgl69DldGxAknaA7wEDc1QB29-ISMCeyWykoKvyXim4UA?e=lnBVCQ

4.4.4 Evaluación de la anotación conceptual.

La tabla 4.4 muestra los resultados de los métodos llamados línea base (*baseline*) y sustitución de sinónimos en las superposiciones léxicas (*synonym-substitution in lexical overlaps*). El primer método incorporó el diccionario de datos creado a partir de HPO y el segundo método se desarrolló al ampliar el primer diccionario con los nuevos sinónimos generados. Los resultados muestran un aumento modesto de precisión (0,02) y de *recall* (0,04).

Metódo	Nº de anotaciones	Nº de términos	Precisión	Recall	F-measure
Línea base	1.232	292	0,69	0,44	0,54
Sustitución de sinónimos en solapes léxicos	1.253	308	0,71	0,48	0,57

Tabla 4.4: Resultados para los dos métodos sobre el *corpus*, usando el OBO Annotator, en términos de precisión, *recall* y *F-measure*.

En total, nuestro método generó 745 sinónimos que cubrían 488 términos HPO únicos, aunque únicamente 36 de ellos estaban cubiertos por las anotaciones del *gold standard*. En otras palabras, solo el 8% de los términos únicos que anotaron el *gold standard* eran términos con nuevos sinónimos. Por lo tanto, los resultados sugieren que el reducido incremento en el rendimiento proviene de una baja cobertura de términos con nuevos sinónimos en el *gold standard*.

4.4.5 Análisis del índice IC de términos

En el momento de la evaluación (abril de 2017), el consorcio HPO proporcionó 129.373 anotaciones de términos de HPO a 9.557 síndromes hereditarios humanos enumerados en OMIM, Orphanet y DECIPHER. Estas anotaciones cubrían 8.237 (75%) términos HPO únicos. Las puntuaciones de IC para estos términos se muestran en la tabla 4.5. Estas puntuaciones oscilaron en el intervalo (0–4). Los términos que no se incluyeron en las anotaciones realizadas se clasificaron como “indefinidos”.

Como podemos ver en la tabla 4.5, el 25% de los términos de HPO son indefinidos, mientras que el 65% de los términos tienen una puntuación superior a 2. Con respecto a los sinónimos generados (745),

corresponden a 488 términos de HPO únicos, donde el 80% de ellos tiene una puntuación más alta de 2. Por lo tanto, un alto porcentaje de los sinónimos generados son altamente informativos, por lo que se espera que tengan un impacto positivo en el reconocimiento de conceptos.

IC	Nº de términos únicos	% de términos únicos	Nº de términos únicos para los sinónimos generados	% de términos únicos para los sinónimos generados
(0,1)	269	2%	1	0%
[1,2)	773	7%	37	8%
[2,3)	2.227	20%	154	32%
[3,4)	4.968	45%	232	48%
indefinido	2.710	25%	64	13%
Total	10.947	100%	488	100%

Tabla 4.5: Número de términos únicos de HPO y número de términos únicos para los nuevos sinónimos clasificados por índice IC.

4.4.6 Evaluación de la colección de resúmenes.

A continuación, analizamos el impacto de los sinónimos generados contando el número de resúmenes en los que se reconoce al menos un término único. Las estadísticas de los términos que utilizan HPO (método de línea base) y la ontología HPO extendida con los sinónimos generados se pueden ver en la tabla 4.6. Los resultados están expresados en términos del número de *abstracts* anotados por cada método. La tasa de incremento es el porcentaje de cambio en el total de las anotaciones. Adicionalmente, los resultados están desagregados por IC y el número de *abstracts* anotados por término. La diferencia entre las anotaciones de ambos procedimientos estaba en los 488 términos únicos correspondientes a los 745 sinónimos generados.

En total, se anotaron 142.043 (24%) resúmenes con algunos de los 488 términos únicos. De ese total, 134.367 resúmenes fueron anotados con el método de línea base; y por lo tanto, el 6% de los 142.043 resúmenes anotados se debió a los sinónimos generados (ver la última fila de la tabla 4.6).

De los 488 términos únicos, 13 (3%) términos anotaron más de 1.000 resúmenes (fila "Total" y "> 1000", resaltados en naranja claro

en la tabla 4.7). Estos términos corresponden a valores de IC inferiores a 3. Los sinónimos generados para estos términos se anotaron solo en los rangos de 0-1% de los resúmenes. Un ejemplo es el término "*atopic dermatitis*" (HP: 0001047), que anotó más de 1.000 resúmenes, y el sinónimo generado "*atopic skin inflammation*", que únicamente anotó 18 resúmenes.

En total, 110 (23%) términos anotaron una cantidad de resúmenes en el rango entre 100 y 1.000 (filas resaltadas en verde en la tabla 4.7). Más del 50% de estos términos tenían valores de IC entre 2 y 3, y anotaron el 14% de los resúmenes. Un ejemplo es el término "*progressive hearing impairment*" (HP: 0001730), que anotó más de 110 resúmenes, y el sinónimo generado "*progressive deafness*", que anotó 23 resúmenes más.

Finalmente, 365 (75%) términos anotaron una serie de resúmenes en el rango entre 1 y 100 (filas resaltadas en azul en la tabla 4.7). Más del 70% de estos términos tenían valores de IC superiores a 3 o no estaban definidos, y anotaron el 56% de los resúmenes. Si observamos el total de valores de IC superiores a 3, se anotaron el 33% de los resúmenes. Un ejemplo es el término "*high-output congestive heart failure*" (HP: 0001722), que se anotó en cinco resúmenes, y el sinónimo generado de "*high-output cardiac failure*", que anotó 35 más.

IC	Nº de *abstracts* por término	Nº de términos	% de términos	Nº de *abstracts* anotados (línea base)	Nº de *abstracts* anotados	Incremento de la tasa de abstracts anotados
[0,1)	>1000	0		0	0	0
	[100,1000)	0		0	0	
	(0-100)	1	1	2	2	100
	Total	1	1	2	2	100
[1,2)	>1000	3	8	7.061	7.126	1
	[100,1000)	18	49	6.799	7.648	12
	(0-100)	16	43	308	449	46
	Total	37	100	14.168	15.223	7
[2,3)	>1000	10	6	90.065	90.175	0
	[100,1000)	60	39	17.265	19.655	14
	(0-100)	84	55	3.192	4.007	26
	Total	154	100	110.522	113.837	3
[3,4)	>1000	0	0	0	0	0
	[100,1000)	32	14	7.366	8.377	14
	(0-100)	200	86	2.004	4.041	102
	Total	232	100	9.370	12.418	33
Indef	>1000	0		0	0	
	[100,1000)	0		0	0	
	(0-100)	64	100	305	561	84
	Total	64	100	305	561	84
Total	>1000	13	3	97.126	97.301	0
	[100,1000)	110	23	31.430	35.680	14
	(0-100)	365	75	5.811	9.060	56
	Total	488	100	134.367	142.041	6

Tabla 4.6: Resultados para ambos métodos sobre la colección de abstracts de enfermedades hereditarias usando el OBO Annotator.

4.5 Discusión

4.5.1 Solapes léxicos en la ontología HPO

El análisis propuesto de los solapes léxicos entre pares de términos vinculados por las relaciones taxonómicas de HPO se puede ver como un nuevo método para medir cuantitativamente cómo la ontología está siguiendo la nomenclatura sistemática; especialmente cuando se utilizan diferencias de género en los nombres, es decir, los términos reflejan diferencias entre el término y su término principal (Schober et al., 2009). En los resultados de la tabla 4.3 se cumple esta convención, ya que de todos los sinónimos potenciales que podrían generarse a partir

de las relaciones jerárquicas en la ontología (121.594), el 95% de estos (115.630), ya están incluidos en la ontología. Estas cifras incluyen repeticiones.

4.5.2 Evaluación de la anotación conceptual

Para nuestro estudio, realizar una evaluación adecuada de los resultados es particularmente difícil. El uso de un *gold standard* es la técnica más adecuada para nuestro caso. Sin embargo, los resultados de la evaluación muestran únicamente un aumento modesto en el rendimiento de la anotación de conceptos, lo cual se debe a dos aspectos. En primer lugar, el uso de un número limitado de resúmenes anotados no proporciona la capacidad de evaluar toda la terminología generada, solo una parte reducida. Cabe destacar que, en este contexto, nuestro método de sustitución de sinónimos ayudó en el reconocimiento de 15 resúmenes más (7% del total de resúmenes) para un total de 16 términos únicos nuevos. Esto representa un aumento del 44% de los términos únicos de HPO cubiertos tanto por el *gold standard* como por los sinónimos generados.

En segundo lugar, el *gold standard* no cubre toda la terminología relevante de HPO. De hecho, las anotaciones manuales incluidas en el *gold standard* solo cubrían el 8% de los términos únicos relacionados con los nuevos sinónimos. Algunos ejemplos de los sinónimos obtenidos que mejoraban el rendimiento en el corpus, se muestran en la tabla 4.7. Estos sinónimos son variantes de los términos de HPO existentes. Los resultados sugieren que su uso mejora el rendimiento de la anotación en comparación con el uso exclusivo de la ontología como fuente de sinónimos.

ID HPO	Término	Término ascendente	Sinónimo ascendente	Nivel de la jerarquía	Sinónimo generado
HP:0100019	*cortical cataract*	*cataract*	*lens opacities*	Segundo	*cortical lens opacities*
HP:0008069	*neoplasm of the skin*	*neoplasm*	*cancer*	Segundo	*cancer of the skin*
HP:0012715	*profound hearing impairment*	*hearing impairment*	*hearing loss*	Primero	*profound hearing loss*
HP:0007270	*atypical absence seizures*	*seizures*	*epilepsy*	Cuarto	*atypical absence epilepsy*
HP:0000122	*unilateral renal aplasia*	*renal agnesis*	*renal aplasia*	Primero	*unilateral renal agnesis*

Tabla 4.7: Ejemplo de generación de cinco sinónimos.

4.5.3 Evaluación de la colección de resúmenes

Como se mostró en la tabla 4.6, tanto los términos con el IC más alto (más de 3) como los términos clasificados como no definidos muestran el mayor aumento en el número de resúmenes anotados. Esto confirma que el procedimiento de sustitución de sinónimos conduce a variaciones léxicas que pueden ayudar a reconocer un mayor número de resúmenes que contienen términos más específicos. La diferencia en el número de resúmenes anotados es menos importante para los términos con menor IC; especialmente para aquellos términos que anotan un número de resúmenes superiores a 100.

Con el objetivo de extraer conclusiones adicionales, revisamos una muestra aleatoria del 2% de los resúmenes anotados con los sinónimos generados. Primero, algunos sinónimos generados fueron variaciones morfológicas de los sinónimos HPO, como "*respiratory recurrent infections*". Como el OBO Annotator genera variantes de los términos de la ontología, la inclusión de estas variaciones morfológicas no produjo ningún cambio en el número de resúmenes anotados. En total, detectamos que el 14% de los sinónimos generados fueron variaciones morfológicas. Sin embargo, la adición de estas variaciones morfológicas podría ser útil cuando se usan reconocedores de conceptos distintos del OBO Annotator. En segundo lugar, algunos

sinónimos generados se incluyeron en otros sinónimos HPO como subcadenas adecuadas. Por ejemplo, el método generó el nuevo sinónimo de "*elbow joint dislocation*" para el término HPO "*elbow dislocation*". En casos como este, la inclusión de estos sinónimos no implicó un cambio en el número de resúmenes anotados. En tercer lugar, detectamos algunos errores inusuales en nuestro método. Un ejemplo es el sinónimo de "*anterior spinal fusion*". Este término no se descartó a través de la búsqueda en PubMed, ya que aparece como parte de "*anterior spinal fusion surgery*" en MEDLINE. Sin embargo, este tipo de errores fue extremadamente raro.

Finalmente, un inconveniente potencial de nuestra evaluación es que, llevamos a cabo esta investigación 16 meses después de haber accedido a HPO por primera vez. Para abordar esta limitación, comparamos la versión utilizada en nuestro trabajo (13 de enero de 2016) y la versión posterior del 13 de abril de 2017. La versión más reciente proporcionó 1.222 sinónimos más (incluidos los nombres de los términos y los términos obsoletos). que la versión utilizada para este estudio. Mientras que, nuestro método, generó únicamente un 3% de sinónimos nuevos (20 términos) a partir de los conceptos añadidos a HPO. La lista de estos sinónimos se proporciona en el Apéndice B.

4.6 CONCLUSIONES

La eficacia del enfoque basado en ontología para el reconocimiento de conceptos se basa en la cobertura de sinónimos para el dominio específico y en qué medida estos sinónimos son apropiados para el procesamiento del lenguaje natural. Sin embargo, las ontologías no están diseñadas específicamente para ser la base léxica para la minería de texto o los sistemas de reconocimiento de nombres, por lo que el rendimiento de los enfoques basados en ontología es inferior al requerido. Esta investigación ha demostrado que es posible reconocer automáticamente nuevas variaciones léxicas para los sinónimos HPO, utilizando las propiedades léxicas y lógicas de la ontología.

Además, el motor de búsqueda PubMed proporcionó un método eficaz para filtrar sinónimos sin sentido. Demostramos que los sinónimos generados tienen un impacto positivo en el reconocimiento

de conceptos, principalmente los que corresponden a conceptos HPO altamente informativos.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CAPÍTULO 5
ANOTACIÓN SEMÁNTICA EN EL ANÁLISIS DE TENDENCIAS EN LA INVESTIGACIÓN EN DIRECCIÓN DE PROYECTOS

Dentro del contexto de la anotación semántica, en este capítulo analizaremos el potencial de la herramienta desarrollada sobre otro ámbito de aplicación conocido como la dirección de proyectos (*project management*, PM). En este dominio surge la necesidad de extraer la información relevante de un gran conjunto de publicaciones para poder analizar sus tendencias temáticas. Durante los últimos años se han realizado varios estudios que tratan de analizar cómo es y hacia dónde se dirige realmente la investigación en la dirección de proyectos. Es por ello que, probamos la eficacia nuestra herramienta de anotación para extraer automáticamente la información relevante sobre un diccionario terminológico que agrupa diferentes glosarios de términos elaborados por instituciones del área.

5.1 INTRODUCCIÓN

La investigación en PM ha evolucionado fuertemente en los últimos años, y se espera que esta tendencia continúe en el futuro previsible. Percibir cómo se está conformando la investigación de PM para vislumbrar cualquier cambio importante en este campo, es un desafío importante hoy en día. En las últimas dos décadas se han realizado numerosos estudios para intentar dar respuesta a estas necesidades de información sobre el área de PM (Pollack y Adler, 2015). Estos trabajos han aplicado diferentes enfoques que van desde procedimientos manuales, como la revisión de literatura (Carden y Egan, 2008; Kwak y Anbari, 2009; Smyth y Morris, 2007; Söderlund, 2004a; Söderlund, 2004b) hasta herramientas específicas de revisión documental (Artto et

al., 2007; Pollack y Adler, 2015; Urli y Urli, 2000), como el análisis bibliométrico o el análisis estadístico. En general, para la extracción de conocimiento se realiza la consecución de las siguientes tres etapas: diseño de la consulta de búsqueda (es decir, selección de la fuente de datos y los términos de consulta), extracción de la información (según campos o secciones determinadas) y análisis y visualización de la información (basado en la frecuencia de palabras y análisis de redes, tales como redes de ocurrencia de citas, redes de ocurrencia de palabras clave, etc.). La tabla 5.1 incluye información sobre cómo diferentes estudios han propuesto diferentes alternativas para cada una de las etapas anteriormente descritas en la investigación de PM.

Artículo	Diseño de la consulta de búsqueda	Extracción de la información	Análisis de la información
Betts y Lansley (1995)	IJPM, entre 1983 y 1992	Título	Análisis de frecuencia
Themistocleus y Wearne (1995)	IJPM y PMJ utilizando 44 tópicos sobre la gestión de proyectos en APM, IPMA y PMI, entre 1984 y 1998	Palabras clave	Análisis de frecuencia
Urli y Urli (2000)	La base de datos ABI-INFORM con la palabra clave: "project management"	Palabras clave	Análisis de co-ocurrencia, utilizando Leximappe
Söderlund (2004)	IJPM y otras 14 revistas buscando por: "projects", "project management", "project organization", "project collaboration" y "temporary organization", entre 1993 y 2002	Artículo completo	Revisión de la literatura y categorización de los resultados
Crawford, Pollack, y England (2006)	IJPM y PMJ buscando por: "project management" (excluyendo revisiones), entre 1994 y 2003	Palabras clave	Análisis de keywords y del corpus linguistico
Smyth y Morris (2007)	IJPM en 2005	Palabras clave	Revisión de la literatura

Carden y Egan (2008)	Emerald, IJPM y PMJ entre 1994-2003	Palabras clave	Revisión de la literatura
Kwak y Anhari (2008)	18 revistas de gestión general entre 1950 y 2007	Palabras clave	Revisión de la literatura
Artto y Gemünden (2009)	23 revistas de negocio buscando por: "program" y "project"	Referencias y palabras clave	Análisis de co-ocurrencia y de frecuencia, utilizando UCINET
Pollack y Adler (2015)	La base de datos ISI WoS con el término de búsqueda: "project Management"	Palabras clave de autor y resumen	Análisis de co-ocurrencia y de frecuencia, utilizando Sci Tool
Padalkar y Gopinath (2016)	Scopus, Emerald Insight, ProQuest, ABI/Informs y Google Scholar buscando por "project" en títulos y "literature review" en abstract o keywords; búsquedas hacia atrás desde Pollack y Adler (2015) en las listas de citaciones.	Artículo completo	Revisión de la literatura y categorización de los resultados
Xia, Zou, Griffin, Wang y Zhong (2018)	Scopus, ASCE y Science Direct, buscando por: "("stakeholder" OR "project participant") AND ("construction project" OR "infrastructure project" OR "civil engineering project") AND ("risk")"	Artículo completo	Análisis de contenido (análisis temático y descriptivo)

Tabla 5.1: Revisión de estudios previos en la investigación de tendencias en la gestión de proyectos.

En primer lugar, como PM es un área de investigación interdisciplinaria, la literatura disponible puede ofrecer una indicación palpable de los avances en este campo. Específicamente, el análisis de publicaciones especializadas (revistas o conferencias) y bases de datos bibliográficas puede considerarse una fuente importante de información para detectar tendencias temáticas actuales, ya que este tipo de literatura ofrece información clave sobre la investigación en

cualquier campo. Estudios recientes confirman el éxito y la utilidad del análisis cuantitativo de las publicaciones científicas (Artto et al., 2007; Pollack y Adler, 2015; Urli y Urli, 2000). Por lo tanto, diseñar adecuadamente la consulta de búsqueda en las fuentes de datos más adecuadas es un hito que puede afectar a las conclusiones finales sobre las tendencias actuales de la investigación. En algunos enfoques, el conjunto de datos se construye a partir de grandes bases de datos, como *Web of Sciences* (WoS) (Pollack y Adler, 2015; Urli y Urli, 2000), *Scopus* (Padalkar y Gopinath, 2016; Xia et al., 2018), o combinando múltiples revistas (Artto et al., 2007; Betts y Lansley, 1995; Carden y Egan, 2008; Kwak y Anbari, 2009; Smyth y Morris, 2007; Söderlund, 2004a; Söderlund, 2004b; Themistocleus y Wearne, 2000), lo que genera una gran cantidad de información que debe filtrarse mediante diferentes criterios para su uso posterior. La información de estas fuentes de datos generalmente se filtra utilizando términos de búsqueda explícitos como "*program*" (Artto et al., 2007) o "*project management*" (Pollack y Adler, 2015; Urli y Urli, 2000). Otra estrategia, como la que se sigue en este documento es, en primer lugar, hacer una selección adecuada sobre las fuentes de datos y, a continuación, recuperar toda la información de estas fuentes (sin filtrado de contenido) a través de los términos de búsqueda. Por ejemplo, Padalkar y Gopinath (2016) y Xia et al. (2018) se centran en su estrategia de filtrado y selección de documentos para minimizar los artículos a estudiar, en este caso, Xia et al., (2018) se enfoca en dos áreas concretas: la administración de riesgos y la administración de partes interesadas (*stakeholders*), para que puedan llevar a cabo una revisión manual del contenido completo del documento. Por otra parte, Gemünden (2014) propone extraer la información a través de una encuesta a expertos del dominio objeto del análisis, dejando de lado la extracción de conocimiento automatizado que supone un alto coste pero limitando la extracción de información a la opinión subjetiva del grupo de expertos.

En segundo lugar, el tipo de análisis requerido generalmente determina el enfoque adoptado en la etapa de extracción de información. Los autores suelen utilizar tres campos principales, es decir, el título, el resumen y las palabras clave para describir su trabajo en una publicación (Pollack y Adler, 2015). La identificación de los

temas clave a través de estos campos se considera una buena opción para mostrar los cambios en el campo de investigación. Por un lado, estos campos proporcionan una descripción clara y concisa del contenido de la investigación. Y, por otro lado, este proceso es sencillo y no consume muchos recursos. Sin embargo, se reconoce que los artículos con resúmenes muy similares pueden ser sustancialmente diferentes con respecto al contenido completo y menos del 8% de las afirmaciones científicas se presentan en los resúmenes (Blake, 2010).

Tercero, con el propósito de abordar el problema del estudio de tendencias en PM, se han utilizado tres métodos de análisis principales en los enfoques que se muestran en la tabla 1.1: análisis de frecuencia, revisión de literatura y análisis de redes sociales. El análisis de frecuencia ha sido utilizado por diferentes autores (Betts y Lansley, 1995; Crawford et al., 2006; Themistocleus y Wearne, 2000). Por ejemplo, Themistocleus y Wearne (2000) analizaron la frecuencia de las palabras clave incluidas en las publicaciones revisadas para proporcionar una visualización general sobre el estado del arte. Este trabajo se opone a otros estudios que consideraron la revisión de la literatura como la mejor manera de categorizar los artículos revisados e identificar los diferentes métodos que han dado lugar a cambios en la investigación de PM (Carden y Egan, 2008; Kwak y Anbari, 2009; Smyth y Morris, 2007; Söderlund, 2004a; Söderlund, 2004b;). Entre todos estos enfoques, también hay discrepancias sobre qué parte del artículo es mejor utilizar. Finalmente, estudios previos (Artto et al., 2007; Pollack y Adler, 2015; Urli y Urli, 2000) acaban de demostrar los beneficios de aplicar el análisis de redes sociales para estudiar la literatura de PM.

Finalmente, existen estudios previos que demuestran los beneficios de aplicar el análisis de redes para estudiar la literatura de PM (Urli, 2000; Artto et al. 2009; Pollack y Adler, 2015). El análisis de redes sociales se inició en los años setenta del siglo pasado, con el objetivo de examinar la estructura de relaciones entre entidades sociales, utilizando diferentes herramientas gráficas, matemáticas y estadísticas, que habían sido originalmente empleadas en sociología (Wasserman y Faust, 1994; Newman et al. al, 2006). Si bien el análisis de redes sociales inicialmente se enfoca en individuos, grupos, empresas u

organizaciones, rápidamente se extiende a la capacidad de diseminación de ideas, enfermedades o influencias, procesos de circulación en mercados o redes de tráfico. La alta aplicabilidad del análisis de redes de co-ocurrencia se debe a la sencillez de sus elementos básicos, facilitando casi cualquier modelo relacional. Este modelo, aunque aparentemente simple, permite la realización de análisis complejos que nos ayudan a comprender el funcionamiento interno de la red en su conjunto, así como el comportamiento individual de las entidades de la red.

En trabajos anteriores (Urli, 2000; Artto et al. 2009; Pollack y Adler, 2015), el uso del análisis de redes de co-ocurrencia para detectar tendencias en la investigación de PM se basa en una revisión de las palabras clave con fines de indexación. Por tanto, cada artículo se caracteriza por unas pocas palabras clave vinculadas. Muchas palabras clave se repetirán en diferentes artículos y es esta reiteración la que permite que el dominio de PM esté representado por redes de palabras clave. El conjunto de palabras clave que están más fuertemente relacionadas entre sí que con las otras palabras clave en la red describe un tema. Cada tema se caracteriza por una palabra clave central y relaciones estadísticas con las otras palabras clave en la red.

El objetivo de nuestro estudio es identificar tendencias temáticas en la investigación reciente de PM. Para ello, proponemos analizar la importancia relativa de la terminología propia de la disciplina, así como las relaciones entre términos extraídos de resúmenes de publicaciones en revistas indexadas en el *Journal of Citation Reports* (JCR). El factor clave que distingue la investigación presentada en nuestro trabajo es el análisis de las tendencias temáticas a través de la anotación semántica. El enfoque extrae automáticamente el conjunto de términos estándar sobre PM de los artículos, y los términos extraídos se consideran entidades de la red. Por lo tanto, los temas de PM se representan utilizando redes de co-ocurrencia de términos estándar y se caracterizan por índices habituales, como la cohesión interna (densidad del tema) y la cohesión externa (centralidad).

El contenido del trabajo se ha estructurado de la siguiente forma. Primero, presentamos la metodología utilizada para cada una de las fases en la extracción de conocimiento. A continuación, se presentan

los resultados obtenidos, sobre los que se elabora una pequeña discusión y, finalmente, se resumen las principales conclusiones del trabajo realizado.

5.2 METODOLOGÍA

Esta sección cubre los aspectos clave del método cuantitativo implementado para analizar publicaciones científicas y académicas relevantes con el objetivo de identificar tendencias temáticas en la investigación de PM. La figura 5.1 muestra el flujo de trabajo de nuestra metodología. Primero, se obtiene la información de interés para el análisis. En nuestro caso, se extraen los resúmenes de cada una de las publicaciones de una revista, previamente seleccionada, de investigación en el área. A continuación, con nuestra herramienta de anotación semántica se anotan automáticamente términos en base a un diccionario hecho a medida. Con la información extraída de los documentos, se identifican las diferentes temáticas a través redes de co-ocurrencia de términos y análisis de *clustering*, cuyo análisis permite que los términos que aparecen con mayor frecuencia sean más prominentes en la literatura revisada. Con el análisis de *burst* podremos mostrar los cambios rápidos en las temáticas extraídas y analizar su tendencia.

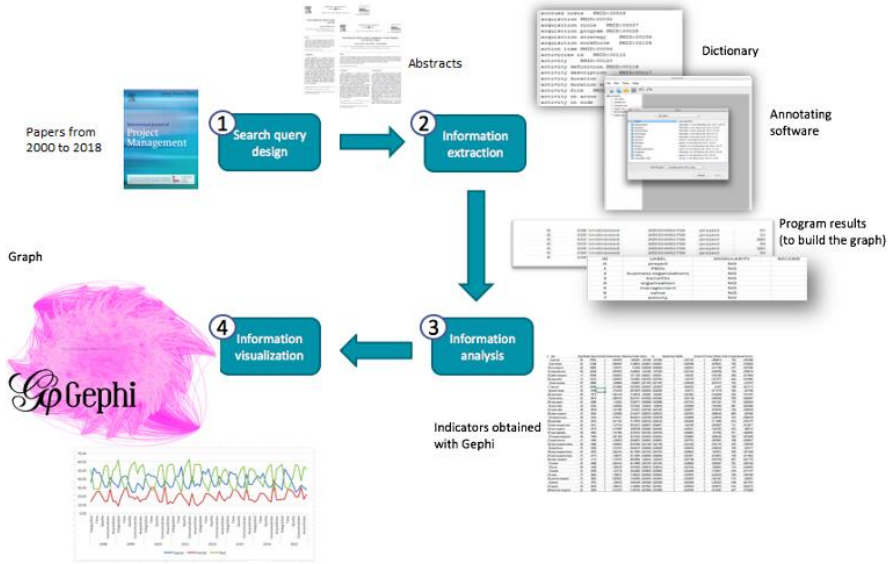

Figura 5.1: Etapas de nuestra metodología para la detección de tendencias temáticas en la investigación sobre dirección de proyectos.

5.2.1 Selección de la fuente de datos

Con el objetivo de seleccionar convenientemente las fuentes de información más adecuadas para nuestro propósito, diseñamos un conjunto de criterios para aplicar sobre nuestro estudio. Primero, con el fin de proporcionar un punto de vista más preciso de la investigación en PM, hemos centrado nuestro estudio en revistas específicas a esta área de trabajo. En segundo lugar, con el objetivo de generar conclusiones con alto impacto y relevancia sobre el desarrollo del campo, seleccionamos revistas incluidas en el *Journal of Citation Reports* (JCR) proporcionadas por Thomson Reuters y/o con el mayor número de publicaciones cada año. El factor de impacto de JCR es una medida reconocida de la calidad científica para evaluar revistas académicas. Actualmente, las revistas incluidas en JCR son artículos revisados por pares y se consideran como los más relevantes en el campo. Finalmente, para obtener un análisis actualizado de las tendencias actuales, hemos analizado los artículos publicados en los últimos diecinueve años, es decir, desde 2000 hasta 2018.

5.2.2 Extracción de conocimiento sobre la fuente de datos seleccionada

La identificación de asociaciones entre palabras clave es sencilla y no consume muchos recursos, ya que los autores proporcionan las principales palabras clave que definen cada uno de sus estudios. Sin embargo, un enfoque general para reconocer automáticamente los términos clave estándar en texto sigue siendo un desafío importante para la investigación. En el área particular de la biomedicina, se han desarrollado diferentes herramientas para reconocer entidades específicas con gran precisión. Sin embargo, estas herramientas específicas generalmente no se pueden usar para identificar términos de otros vocabularios (Doms y Schroeder, 2005), en parte, porque deben capacitarse en colecciones de texto que generalmente no están disponibles y, por otro lado, como resultado de la brecha existente entre el significado de los términos en los diccionarios y en los textos. Como consecuencia, han surgido algunas herramientas basadas en diccionarios para reconocer términos en el dominio biomédico, como NCBO Annotator (Jonquet et al., 2009), MetaMap (Aronson, 2006) y nuestro OBO Annotator (Taboada et al., 2014) (herramienta de nuestro trabajo de investigación), que son herramientas genéricas orientadas a reconocer términos de cualquier vocabulario.

Nuestro enfoque para extraer información de resúmenes en PM se basa en el uso del núcleo OBO Annotator. Este anotador se puede utilizar para anotar automáticamente los textos como un paso intermedio para indexarlos y luego agruparlos por temas. Como ya hemos señalado, fue diseñado específicamente para el ámbito biomédico utilizando ontologías OBO como principal recurso terminológico. Las ontologías son un recurso muy valioso ya que definen el mapa conceptual del dominio para el cual fueron diseñadas, pero, ante la ausencia de una ontología validada por un organismo reconocido en el campo PM, hemos utilizado el núcleo OBO Annotator con un diccionario terminológico hecho a medida, basado en un conjunto de glosarios validados por el *Project Management Institute* (PMI).

En los siguientes apartados especificamos la metodología seguida para la construcción de nuestro diccionario de referencia y exponemos la(s) sección(es) de nuestra fuente de datos utilizadas en el análisis.

5.2.2.1 Un diccionario específico para la dirección de proyectos

Para elaborar el diccionario para PM, hemos utilizado dos fuentes de datos como referencia. El primero consiste en un glosario completo de términos de *Project Management Office* (PMO) y PM (Filicetti, 2016), el cual nos ha proporcionado el núcleo de nuestro diccionario con un total de 733 términos. Hemos ampliado este núcleo con 123 nuevas entradas del glosario de gestión de proyectos de Wikipedia (Wikipedia, 2018). Cabe señalar que toda la terminología seleccionada está validada por PMI (PMI, 2013), por lo que la perspectiva metodológica PMI se garantizó en el análisis, pero no fue dominante en relación con otros términos comúnmente utilizados en la disciplina. Finalmente, el diccionario se ha revisado manualmente por un experto en el dominio que ha añadido la correspondencia entre siglas y términos extendidos, y ha establecido las relaciones de sinonimia entre los conceptos, lo que garantiza la conciliación de las diferencias en el uso de términos sinónimos en textos y entre diferentes documentos.

Por otro lado, para evitar que términos con significado muy amplio (conceptos muy generales) y poco frecuentes emergieran en la red de co-ocurrencia, se eliminaron de la red siguiendo la ley de *Zipf* (Kim et al., 2020).

5.2.2.2 Selección de nuestra fuente de datos a procesar

Un aspecto planteado por muchos expertos en el área, ya no sólo en la investigación sobre gestión de proyectos, es qué parte del artículo debemos analizar para alcanzar nuestro propósito. En la tabla 1.1, hemos visto cómo la gran mayoría se centra en analizar las palabras clave, ya que las consideran como la mejor opción para extraer las temáticas que se tratan en el artículo. Otros, como Padalkar y Gopinath (2016), piensan que es mejor centrarse en la estrategia de búsqueda para reducir al mínimo los artículos utilizados para poder realizar una

revisión manual del contenido completo. Por otro lado, y en menor medida, otros como Pollack y Adler (2015) también consideran el *abstract* o resumen, ya que al ser una breve representación del artículo se deben proporcionar todos los temas a tratar. Es por ello que, hacer una revisión de la literatura, estudiar las palabras clave de autor utilizadas o analizar los resúmenes de las publicaciones son las tres fuentes clásicas para extraer información. Partiendo de la hipótesis de que los resúmenes de los artículos proporcionan mayor información dado que proveen de contexto a los términos clave de autor, hemos probado la eficacia de nuestra herramienta de anotación semántica sobre los resúmenes de la base de datos seleccionada.

5.2.3 Construcción de la red de co-ocurrencia

Una vez seleccionada la fuente de datos a procesar, hemos desarrollado una herramienta, llamada PIMAnnot (*Project Information Management Annotator*) para anotar automáticamente con el diccionario elaborado 1) resúmenes proporcionados por los autores en formato electrónico; y 2) construir redes de co-ocurrencia de términos que representan los temas principales en el dominio de PM. Cada tema está construido alrededor de un término estándar primario que sirve como su definición, y con el cual el resto de los términos del tema están en una relación estadística. En la figura 5.2, se ejemplifica cómo se establece la relación semántica entre términos para obtener la red de co-ocurrencia.

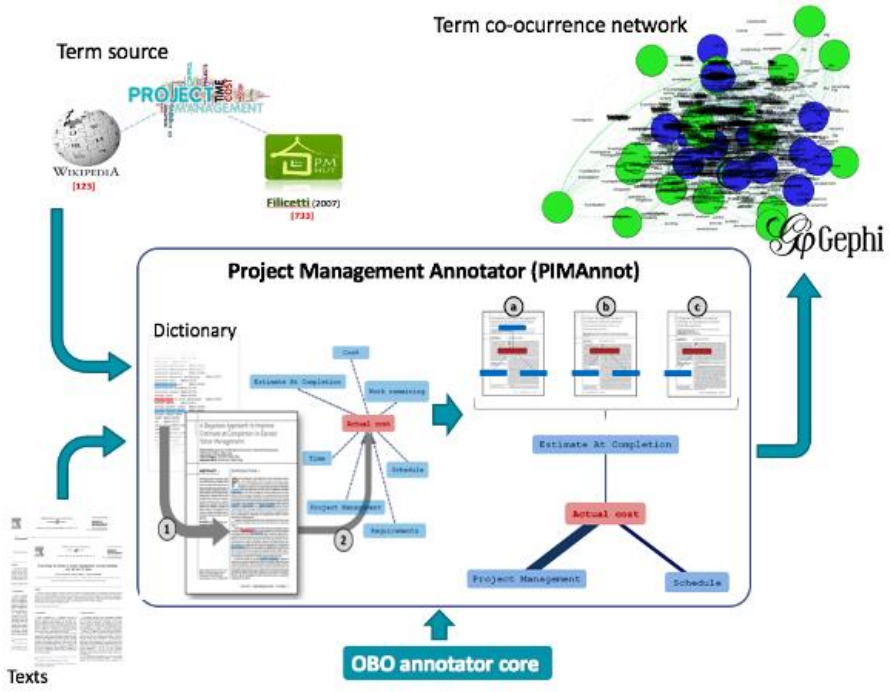

Figura 5.2: Metodología aplicada para la extracción de conocimiento y la posterior construcción de la red de co-ocurrencia.

Primero, la herramienta inicia sesión en el sitio web de la revista para descargar los documentos automáticamente. En segundo lugar, los resúmenes se anotan con el conjunto de términos incluidos en el diccionario. La anotación es el proceso de asignar términos del glosario a los documentos. Para hacer esto, la herramienta primero procesa los términos del glosario y los términos de los documentos, para conciliar las pequeñas diferencias ortográficas, y luego compara los textos del documento con los términos del glosario. El pre-procesamiento incluye transformar palabras en minúsculas, dividir los términos del glosario en palabras, eliminar palabras comunes y signos de puntuación, reemplazar palabras con el mismo lexema (eliminando los sufijos), generar variaciones de términos como permutaciones de los lexemas y filtrar variaciones de términos incorrectos. Esta fase es necesaria en la investigación cienciométrica para garantizar la integridad de los datos, y se ha descrito como costosa y lenta, ya que generalmente requiere el

80 por ciento del esfuerzo total en este tipo de proyectos (Börner, 2010). Sin embargo, esta fase de pre-procesamiento se realiza automáticamente con PIMAnnot. En tercer lugar, la herramienta utiliza el conjunto de anotaciones para construir las redes de co-ocurrencia de términos. Las aristas de la red representan la co-ocurrencia de dos términos utilizados en el mismo artículo de investigación.

Gephi (Gephi v0.9, 2016) es la herramienta de software elegida para analizar y visualizar las redes de co-ocurrencia debido a que nos permite extraer los indicadores necesarios para nuestro análisis de redes: grado, grado ponderado, excentricidad, modularidad, centralidad intermedia y longitud media del camino, entre otras medidas que nos proporciona la herramienta. Los términos con mayor frecuencia relativa, mayor grado y mayor centralidad son términos con mayor relevancia en la investigación actual de PM.

5.2.4 Análisis de tendencias temáticas

Utilizando la propiedad de modularidad sobre nuestra red de co-ocurrencia, se obtuvieron las diferentes agrupaciones de conceptos o subredes. Las subredes extraídas representaban, de esta forma, las diferentes agrupaciones de términos en base a su relación semántica en las publicaciones, constituyendo así las diferentes temáticas.

La evolución temporal de los diferentes temas se visualizó a través de la frecuencia con la que estos aparecieron en cada una de las publicaciones en el periodo analizado. Se consideró que una temática se utilizaba en una publicación si alguna de las palabras clave que componían dicha temática estaba presente. Además, se utilizó la técnica de detección de ráfagas o *burst* de Kleinberg (Kleinberg, 2003) para identificar períodos de tiempo en los que un tema era excepcionalmente popular. Tradicionalmente este análisis se aplica sobre palabras clave pero, en nuestro caso, lo hemos aplicado directamente sobre la temática, dado que ésta se compone de varios términos. Para cada tema y cada período de tiempo, se distinguieron dos eventos: un evento objetivo (que incluía el tema) y un evento no objetivo (no incluía el tema). Los eventos objetivo consistían en todos los documentos que incluían el tema y los eventos no objetivo consistían en todos los demás documentos. Asumimos que un documento incluía un tema si alguno

de los términos que constituían esa temática se producía en el conjunto de términos del documento. También partimos de la hipótesis de que sólo dos estados eran posibles: el *estado base*, correspondiente a la probabilidad más baja de eventos objetivo, y el llamado estado de *ráfaga o burst*, asociado a la mayor probabilidad. Sólo los estados de *burst* se representaron gráficamente.

5.3 RESULTADOS

En esta sección, se muestran los resultados obtenidos para cada una de las etapas previamente definidas para la extracción de información y detección de las principales tendencias temáticas en PM.

5.3.1 Selección de la fuente de datos

Actualmente, hay varias revistas de investigación que publican artículos sobre las últimas tendencias en investigación e innovación de PM, incluidos diferentes temas y áreas de conocimiento. Entre ellos, podemos encontrar la *International Journal of Project Managenment*, la *Project Management Journal*, la *International Journal of Managing Projects in Business* y la *International Journal of Production Research*, que aunque no se centra exclusivamente en PM, con frecuencia publica artículos en esta área. Además, esta lista de revistas podría completarse con otras que no están incluidas en JCR y con una perspectiva que combine la actividad académica y profesional, como el *Journal of Modern Project Management*, *Project Management World Journal* o *International Journal of Project Organization and Management*. En la Tabla 5.2, mostramos un resumen de las principales características de estas seis revistas.

Revista	Publicaciones por año	Indexada en JCR	Centrada en PM
IJPOM	3-4	X	✓
IFRP	24	✓	X
PMWJ	12	X	✓
PMJ	6	✓	✓
IJPM	8	✓	✓
IJPMB	4	✓	✓

Tabla 5.2: Resumen de la metodología aplicada para la extracción de información en la investigación sobre PM.

La *International Journal of Project Organization and Management* (IJPOM), que se publica tres o cuatro veces al año, trata de atraer contribuciones, y especialmente estudios de casos, de un amplio espectro de académicos y profesionales en PM.

El *International Journal of Production Research* (IFPR) se publica 24 veces al año, incluidos documentos sobre gestión de la innovación, diseño de productos, procesos de fabricación, sistemas de producción y logística. Se considera la economía de producción, el comportamiento esencial de los recursos y sistemas de producción, así como los complejos problemas de decisión que surgen en el diseño, la gestión y el control de los sistemas de producción y logística. La buena reputación de IJPR se basa en un fuerte vínculo con las aplicaciones industriales.

Por el contrario, el *PM World Journal* (PMWJ) es un diario electrónico no arbitrado y *PM World* lo publica mensualmente y contiene una amplia gama de artículos, documentos e historias sobre la gestión de proyectos y programas (P / PM). El contenido de la PMWJ está escrito por expertos y profesionales de P / PM de todo el mundo.

El *Project Management Journal* (PMJ) es la revista académica y de investigación del PMI. Publicado seis veces al año, trata de abordar los amplios intereses de la profesión de PM y mantener un equilibrio editorial de contenido sobre investigación, técnica, teoría y práctica.

La *International Journal of Project Management* (IJPM) es una revista con una amplia gama de todas las facetas de la gestión de proyectos que proporciona un enfoque para la experiencia mundial en las técnicas, prácticas y áreas de investigación requeridas. Se publica ocho veces por volumen.

En última instancia, el *International Journal of Managing Projects in Business* (IJMPB), creado recientemente en 2008, ofrece una amplia cobertura de todos los aspectos de la gestión de proyectos, desde la estrategia hasta la planificación y la implementación. Se publica cuatro veces al año y, a pesar de ser una revista joven, su enfoque único y práctico en la gestión de proyectos en los negocios constituye un recurso esencial para todos los involucrados en el campo de la gestión de proyectos.

Aplicando los criterios elegidos para seleccionar las revistas y descritos en la sección de metodología, hemos trabajado únicamente con el *International Journal of Project Management* (IJPM).

El período de análisis seleccionado fue del año 2000 al año 2018, lo que nos ha proporcionado un total de 1.612 artículos de investigación sobre PM.

La figura 5.3 muestra el número de artículos de investigación publicados por año en IJPM. Desde 2000, ha habido un aumento considerable en el número de publicaciones hasta la fecha, pese a que en 2018 se produjo un pronunciado descenso en su proliferación.

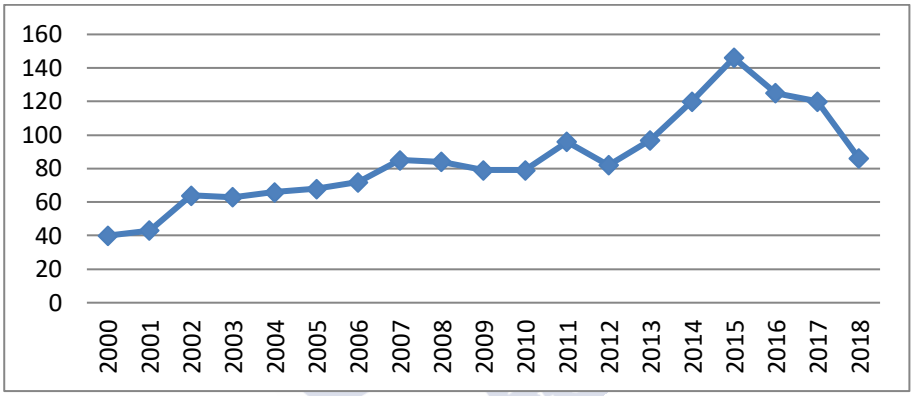

Figura 5.3: Artículos publicados por IJPM entre 2000 y 2018.

5.3.2 Extracción de información

A partir de los glosarios terminológicos Filicetti (2016) y Wikipedia (2018), se creó un diccionario con un total de 926 términos. Para el estudio de las tendencias temáticas, siguiendo a Pollack y Adler (2015), se filtraron los términos que definen el propio ámbito de estudio, es decir, "*project management*", "*project*" y "*PM*". Un experto en el dominio de PM también propuso filtrar los términos de "*ladder"*, "*fordism"* y "*construction".* También propuso añadir términos que estaban en pleno auge en el dominio. Estos términos fueron: "*life cycle assessment"*, "*lean"*, "*extreme programming (XP)"*, "*canvas"*, "*project canvas"*, "*kanban"*, "*scrum ban"*, "*cristal"*, "*crystal method"*, "*agile unified process* (*AUP*)", "*feature driven development* (*FDD*)" y "*kanban method"*. Sobre el listado de términos resultante se

aplicó la ley de *Zipf* y se eliminaron los términos de 1 palabra (con significado más amplio) y de baja frecuencia como, por ejemplo, "*lag*", "*calendar*" o "*chart*". De esta forma nos quedamos con un diccionario compuesto de 874 términos, 799 términos únicos sin contar sinónimos o siglas.

Por otro lado, nuestra herramienta de anotación (PIMAnnot) obtuvo 1.612 artículos de la revista IJPM para el período comprendido de 2000 a 2018 (diecinueve años), de los cuales se extrajo únicamente el resumen. Como todos los artículos constaban de resumen, se utilizaron los 1.612 textos para anotar. Aunque no todos se pudieron anotar, se anotaron 1.597 (99,07%) y se obtuvieron un total de 15.029 anotaciones y 317 conceptos únicos con nuestro diccionario.

En la tabla 5.3 se muestra el número de términos anotados por publicación, teniendo en cuenta si el término estaba compuesto por una palabra, dos palabras o tres o más palabras. Aunque los términos de una palabra no son los más anotados, sí que aparecen en la mayoría de las publicaciones (94,72%). En cambio, los términos de dos palabras son los más abundantes en la literatura puesto que suponen la mayor parte de las anotaciones realizadas (60,88%) y también abundan en la literatura (80,95% de las publicaciones).

	Nº Términos anotados (no repetidos)	% Términos anotados	Nº publicaciones	% publicaciones
1 palabra	87	27,44%	1527	94,72%
2 palabras	193	60,88%	1305	80,95%
3 palabras	37	11,67%	234	14,51%
Total	**317**	**100%**	**1612**	**-**

Tabla 5.3: Número términos anotados por artículo.

En la tabla 5.4 se analiza la frecuencia de aparición de los términos anotados con la herramienta. Como se puede observar, la mayoría de los términos suelen ser poco frecuentes, menos de 5 apariciones en el total de las publicaciones. Los términos más frecuentes suponen el 20% del total.

Frecuencia	Nº términos anotados en el conjunto de datos original	%
> 100	36	11,36%
> 50	62	19,56%
> 25	91	28,70%
> 10	133	41,95%
>5	193	60,88%
> 2	267	84,22%
>1	317	100%

Tabla 5.4: Frecuencia de aparición de los términos anotados.

5.3.3 Análisis de la red de co-ocurrencia

En la figura 5.4 se muestra la red de co-ocurrencia obtenida con la ayuda del software Gephi. En total se obtuvieron 317 nodos (términos) sobre la anotación de 1.612 textos. El tamaño del nodo representa la frecuencia en la cual un término aparece en el conjunto de datos y el peso de la arista muestra la frecuencia con la cual dos términos coinciden en el mismo documento. El grado medio de la red fue de 2, lo que quiere decir que, de media, un concepto está conectado/relacionado con otros dos conceptos. El diámetro obtenido fue de 9 y la longitud media de camino de 4,15, lo que nos da una idea de la proximidad entre los nodos y lo densa que es nuestra red. El valor de modularidad de la red, esto es, la capacidad de la red de descomponerse en subredes o clústeres, fue de 0,622. Este valor implica que se obtendrán de forma más o menos clara las diferentes subredes organizadas por la similitud semántica de los diferentes nodos.

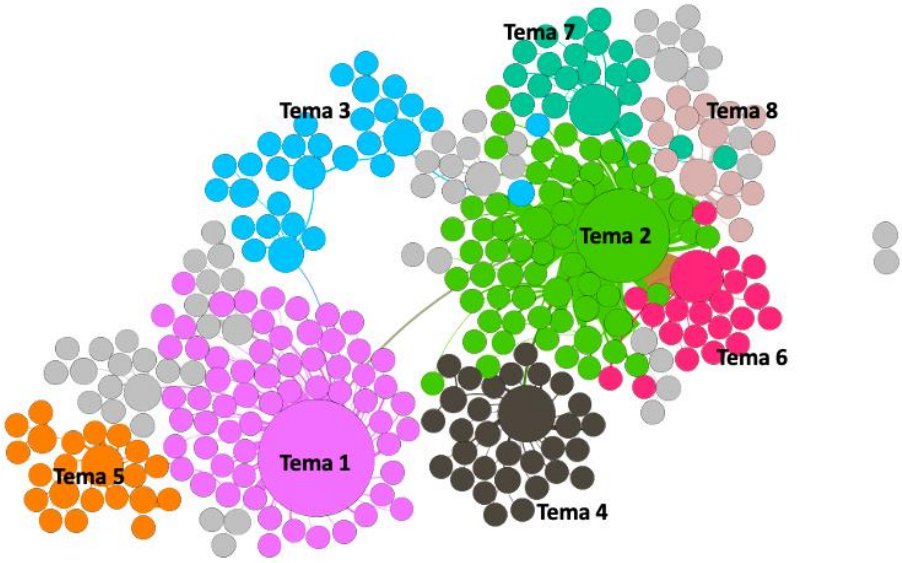

Figura 5.4: Red de co-ocurrencia de términos extraídos de los resúmenes de las publicaciones realizadas entre 2000 y 2018 en IJPM.

5.3.4 Evaluación de la calidad de las temáticas obtenidas

Evaluar si nuestro modelo ha extraído información de calidad y relevante para las necesidades de información en el área de investigación en PM es realmente difícil. Es por ello que para evaluar nuetros resultados realizamos una evaluación en tres fases. Primero, construimos una red de co-ocurrencia de términos a partir del conjunto de conceptos obtenidos en la anotación. A continuación, extraemos los temas, es decir, las subredes emergentes de nuestra red. Y, finalmente, calculamos la métrica de evaluación intrínseca conocida como *coherencia de temática o topic coherence* con el fin de evaluar la calidad de las temáticas identificadas en la red. En nuestro enfoque, se dice que un conjunto de términos es coherente, si el conjunto se puede interpretar en un contexto significativo que cubra todas o la mayoría de los conceptos. La coherencia del tema mide la puntuación individual, cuantificando el grado de similitud semántica entre los términos de puntuación más altos de cada temática (Stevens et al., 2012). Se puede definir como la suma de las puntuaciones dela distribución de similitud en pares sobre el conjunto de términos de cada temática, T.

$$\text{Coherence}(T) = \Sigma\ \text{Score}(k_i, k_j)\ ,\ \forall\ (k_i, k_j) \in T$$

Esta medida de la coherencia entre términos que componen una temática distingue entre los temas que son semánticamente interpretables y los que son artefactos numéricos. En la literatura, podemos encontrar varias medidas de coherencia, entre las que seleccionamos la medida UMass para este trabajo (Mimno et al., 2011). Esta medida se basa en la co-ocurrencia de términos en los documentos.

$$\text{Score}(k_i, k_j) = \log\left((D(k_i, k_j) + 1)\right) / D(k_j))$$

donde D(a, b) es el número de documentos en los que los conjuntos de términos a y b coexisten, y D(a) es el número de documentos etiquetados por el conjunto de términos. La métrica UMass cuantifica estas puntuaciones en el corpus para confirmar que los modelos se crearon a partir de los datos del corpus.

5.3.5 Análisis de tendencias temáticas

En la tabla 5.5 se muestran las ocho tendencias temáticas identificadas a través del análisis de co-ocurrencia de términos. En esta tabla se expone el top 10 de términos ordenados por su valor de grado medio ponderado. El listado de términos que componen cada clúster o tema se utiliza para identificar las temáticas de baja calidad y de alta calidad en función del valor del parámetro UMass obtenido sobre cada clúster o temática. Las temáticas con una puntuación de coherencia temática (o *topic coherence*) entre -2 y 2 fueron calificadas como temáticas de buena calidad semántica, mientras que el resto se consideraron de mala calidad.

En nuestro caso hemos obtenido el 50% de los temas de la red con buena coherencia semántica. Esto implica que, en base a las palabras clave que lo componen, se puede deducir fácilmente una temática de PM. Un experto en el dominio fue el encargado de etiquetar cada temática o clúster bajo un título representativo que definiese correctamente el tema al que hacen referencia los términos que lo componen. Nuestro experto utilizó de base estudios previos en el dominio (Pollack y Adler, 2015; Padalkar y Gopinath, 2016) y la guía

de proyectos del PMI, el PMBoK (PMI, 2017), institución reconocida en la que también se basa el diccionario elaborado.

Entre los temas de alta calidad (valor de UMass cercano a cero) identificados por nuestro experto se encuentran "*Project methods*", "*Risk management*", "*Governance & control*" y "*Project strategy*". A pesar de que el tema 1 se lleva el mayor porcentaje de nodos de la red (19,87%), este tiene un valor absoluto UMass elevado. Lo que significa que nuestro tema es poco coherente, ya que los términos que en él se encuentran tienen poca relación. Por el contrario, si eliminásemos del tema 1 ("*Project life cycle*") los términos de "*supplier*" y "*change management*" que, a priori tienen menos relación con el ciclo de vida del proyecto, y eliminásemos "*phase*" por tener un significado demasiado amplio o general, mejoraríamos sustancialmente el valor de coherencia. En este caso concreto pasaríamos de un valor de UMass de -9,55 a -3,59, es decir, de calidad media. Lo mismo ocurre en el caso del tema 8 ("*Integration management*"), en este caso nuestro experto identifica términos propios del área de conocimiento de gestión de integración del proyecto como "*integration*", "*outsourcing*" o "*concurrent engineering*" y el resto de términos tienen relación con esta área pero no le son propios. Si eliminamos los conceptos menos específicos (con significado más amplio), el valor de UMass mejora de -9,11 a -1,56, y se convertiría también en una temática de calidad.

Nº	Temática	Top 10 de palabras clave de autor	% nodos	UMass
1	Project life cycle	phase, project phase, supplier, execution phase, life cycle, implementation phase, project strategy, project life cycle, commitment, change management	19,87%	-9,55
2	Project methods	dependency, activity, cost, relationship, performing, risk, coordination, contractor, impact, decision	18,93%	-0,84
3	Risk management	project network, critical path method, risk evaluation, management science, risk register, risk assessment, mitigation, network analysis, risk identification, risk analysis	9,78%	-1,68
4	Governance and control	projectized organization, project culture, benefits management, business model, configuration control, functional organization, control account, configuration management, project accounting, performing organization	9,78%	0,00
5	Project environment	multi-project, leadership, capability maturity model, program manager, portfolio manager, process management, resource allocation, six sigma, resource constraint, project environment	6,94%	-6,66
6	Project strategy	project manager, business process modeling, secondary risk, critical chain method, delaying resource, application area, project logic, focused improvement, value planning, impact analysis	6,62%	0,00
7	Scheduling	task, precedence relationship, critical chain project management, critical path, project charter, acceptance criterion, multi-project scheduling, benefits framework, organization design	6,62%	-8,24
8	Integration management	integration, earned schedule, outsourcing, planned start date, quality control, scope definition, cost control, concurrent engineering, risk owner, schedule performance index	4,73%	-9,11

Tabla 5.5: Temáticas identificadas en la red de co-ocurrencia y su cualificación en alta y baja calidad en base a la medida de *Topic coherence*.

La figura 5.5 muestra la frecuencia con la que las temáticas de alta calidad identificadas se utilizaron en cada una de las publicaciones realizadas por año. Para representar la variación anual de cada temática

se consideran los resultados como una fracción del número total de publicaciones por año.

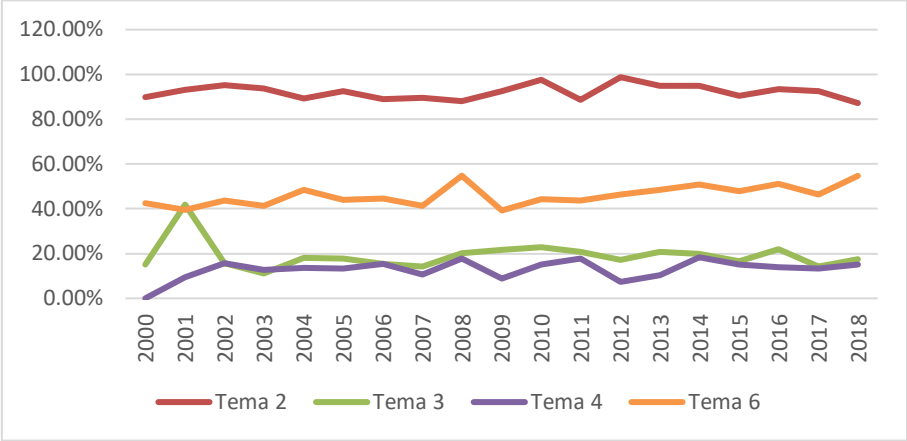

Figura 5.5: Frecuencia temática de los temas de alta calidad sobre las publicaciones realizadas por año.

La figura 5.6 muestra el resultado del análisis de *burst* para las temáticas de alta calidad obtenidas en el período de análisis (2000 a 2018). Para el análisis de *burst* se estableció el parámetro s (distancia multiplicativa entre estados) en 1,3, y el parámetro gamma (dificultad asociada con el ascenso de un estado) en 0,1. Cada punto representado en la figura indica un pico de popularidad (o estado de *burst*) durante el período en el que se muestra. Hay que tener en cuenta que el análisis de *burst* muestra el cambio rápido de frecuencia pero no representa el porcentaje de frecuencia sobre el total (figura 5.5). Por lo tanto, una temática puede presentar un pico de popularidad en un período concreto, pero tener una frecuencia más o menos constante y alta en el resto del período.

Como se puede ver en la figura 5.6 el tema 2 ("*Project methods*") no alcanzó ningún pico de popularidad o estado de *burst*, dado que se mantiene más o menos constante en el tiempo. Por esta razón y dado que está compuesto, principalmente, de términos con significado amplio (generales), se trata de una temática bastante recurrente en el tiempo. El tema 3 ("*Risk management*") muestra un pico de popularidad considerable en el año 2001 y se mantiene más o menos constante en el

resto del periodo, al igual que le sucede al tema 4 ("*Governance & control*") salvo los picos ocasionales de crecimiento. El tema 6 ("*Project strategy*") es el único que muestra una tendencia creciente con picos de popularidad en el 2008 y en el 2018.

Figura 5.6: Análisis *burst* de los temas identificados de buena coherencia temática.

5.4 DISCUSIÓN

En este trabajo hemos puesto a prueba nuestra herramienta de anotación semántica (Taboada et al., 2014) sobre un ámbito totalmente opuesto al ya conocido en capítulos previos, el de investigación en PM. Nuestra herramienta ha demostrado ser eficaz para extraer conocimiento de esta disciplina ya que nos ha permitido diseñar el mapa conceptual del dominio a partir de redes de co-ocurrencia de términos (Callon et al., 1983). Nuestra propuesta de trabajo permitió así obtener ocho temáticas de las cuales, la mitad, resultaron tener una coherencia semántica aceptable para representar temas de PM.

Los resultados obtenidos por nuestra metodología automática son consistentes con otros dos estudios similares realizados en el área de PM. El primero es el estudio de Pollack y Adler (2015), cuyo análisis de tendencias realizado entre los años 1962 y 2012 a partir de resúmenes y palabras clave de publicaciones del área, sirve de referencia para poder comparar nuestros resultados. El segundo es el estudio de Padalkar y Gopinath (2016), que analizaron manualmente las tendencias en PM durante un periodo de 16 años entre 2000 y 2015. Los autores seleccionaron 230 artículos de tres revistas internacionales utilizando una medida basada en el recuento de citas. En total, nuestra fuente de datos (IJPM) proporcionó 145 (63%) artículos en un período cubierto completamente por nuestro análisis (nuestro trabajo incluyó 3 años más al final del período).

Si comparamos nuestros resultados con Pollack y Adler (2015), vemos como en su análisis de resúmenes de publicaciones obtienen únicamente términos muy generales, constituidos por una única palabra. Este hecho respalda la composición de nuestro tema 2 ("*Project methods*"). De hecho, en la identificación del top 20 de términos en frecuencia y *burst* de Pollack y Adler (2015) tenemos dos términos comunes en nuestro clúster, "*cost*" y "*decision*". Aunque nuestro tema 2 ("*Project methods*") esté compuesto por términos con significado amplio, términos como "*dependency*", "*cost*", "*coordination*", "*decision*" e "*impact*", son propios a la gestión del proyecto y con gran uso en el área. De ahí que esta temática se mantenga más o menos constante en el tiempo sin ningún pico de popularidad destacable. Padalkar y Gopinath (2016) sitúan esta misma temática como un tema recurrente en todo su periodo de análisis (2000-2015).

Dado que Padalkar y Gopinath (2016) trabajaron con la misma revista que nosotros y en un período similar, se considera un modelo bastante fiable con el que comparar nuestros resultados. De hecho, la progresión temática obtenida en su estudio, durante el período 2000 a 2015, cubre los 4 temas de alta calidad que surgieron en nuestro análisis. En el caso del tema 3 ("Risk management"), la temática viene claramente definida por todos los conceptos que contienen la palabra "*risk*" y términos vinculados estrechamente a esta área de conocimiento, como "*critical path method*" o "*mitigation*" (PMI, 2017). Padalkar y Gopinath (2016) señalan este tema como relevante de 2000 a 2015, lo que concuerda con nuestro análisis de *burst*, el cual revela picos de popularidad en el año 2001 y 2015. Por otro lado, en el caso del tema 4 ("*Governance & control*"), nosotros identificamos tres estados *burst* en 2008, 2011 y 2014. En el caso del tema 6 ("*Project strategy*") identificamos picos de popularidad en 2008 y en 2018. Estas dos últimas temáticas coinciden con el último periodo de análisis de Padalkar y Gopinath (2016), que sitúan ambas entre 2006 y 2015.

Pero nuestro planteamiento tiene una serie de limitaciones que veremos a continuación. Primero, el anotador utiliza un diccionario con menos de 1.000 palabras. Pese a que nuestro diccionario fue construido a partir de glosarios ya existentes y aprobados por la comunidad

científica, hay que tener en cuenta que muchas veces no concuerda con la terminología utilizada por los autores. La terminología utilizada viene influenciada, muchas veces, por la moda existente en el momento de la publicación que no se suele reflejar en este tipo de glosarios.

Segundo, aunque nuestro diccionario fue revisado por un experto y se eliminaron los términos de menor frecuencia, conceptos con significado amplio o demasiado generales emergieron en la red. Es el caso de términos como "*activity*", "*phase*" o "*decision*", que principalmente se concentraron en el clúster o tema 2 ("*Project methods*"). Esto se debe a que el análisis de co-ocurrencia de palabras se basa en la frecuencia y estos términos son muy recurrentes en el área de estudio aunque, probablemente, en diferentes contextos de uso. Como son términos muy frecuentes no se eliminaron de la red de co-ocurrencia, pero limitaron la identificación de tendencias al añadir términos poco específicos a la composición de los clústeres temáticos.

Y, por último, el etiquetado de un clúster o subred de términos bajo un mismo nombre común es muy subjetivo y vinculado a los términos de análisis. En esta fase de nuestra metodología únicamente se seleccionaron los diez términos con mayor grado medio ponderado en la red, ya que se supone que son los conceptos más representativos de la temática que componen (Chen y Xiao, 2016). A pesar de que la medida de t*opic coherence* o coherencia temática nos ayuda a identificar las temáticas de alta y baja calidad, no tenemos una guía que nos indique cómo etiquetar cada clúster o subred en función de la terminología que lo compone. Es por ello que el etiquetado realizado por nuestro experto es subjetivo y puede ser que fuera necesario una evaluación por parte de un grupo de expertos.

Pese a las diferencias existentes con los estudios previos comentados y de las limitaciones impuestas, principalmente, por las limitaciones del propio dominio, nuestros resultados constatan que 1) nuestra herramienta es eficaz puesto que nos ha permitido extraer las principales tendencias temáticas del área y que se ajustan a planteamientos previos de otros autores (Padalkar y Gopinath, 2016), y 2) nuestra metodología una excelente propuesta automática para el análisis de tendencias, ya que se trata de un procedimiento adaptable a

diferentes ámbitos de estudio que ha arrojado buenos resultados en términos de coherencia temática.

5.5 CONCLUSIONES

En este estudio hemos comprobado la eficacia de nuestra herramienta en la identificación de tendencias temáticas en el área de investigación en dirección de proyectos. Esto se debe a que nuestra metodología nos ha permitido obtener resultados que concuerdan con estudios previos en el área (Pollack y Adler, 2015; Padalkar y Gopinath, 2016). Nuestro principal logro es corroborar que nuestra herramienta es capaz de extraer información relevante y de interés utilizando recursos terminológicos diferentes (utilizamos un diccionario frente al uso de ontologías ya estudiado), en un contexto totalmente diferente y que nuestro análisis concuerda con estudios previos en el área. Se propone, así, un nuevo método para la obtención, de forma semi-automática, de tendencias temáticas que puede ser muy útil en la comunidad científica puesto que se podría aplicar a diferentes ámbitos de estudio. Nuestra propuesta da paso, de esta forma, a nuevos caminos experimentales como el uso de otra fuente de conocimiento, ya sea las palabras clave de autor, o a la búsqueda de recursos para pre-procesar la terminología, con el objetivo de mejorar los resultados a la hora de extraer información.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CAPÍTULO 6
UNIFICACIÓN AUTOMÁTICA DE PALABRAS CLAVE DE AUTOR SOBRE EL ESTUDIO DE TENDENCIAS TEMÁTICAS EN LA INVESTIGACIÓN EN DIRECCIÓN DE PROYECTOS

Las palabras clave son las unidades de texto más valiosas para numerosas tareas de recuperación de información y procesamiento de lenguaje natural, como el resumen de documentos o el análisis de tendencias temáticas. Un problema importante pero poco abordado hasta ahora es la unificación de un gran conjunto de palabras clave altamente heterogéneas extraídas de documentos dentro de un dominio. Para abordar esta brecha, proponemos a continuación un enfoque automatizado para la unificación de palabras clave, que se basa en la combinación de un conjunto de técnicas léxicas, sintácticas y semánticas.

6.1 INTRODUCCIÓN

Las palabras clave son descripciones compuestas de una o varias palabras de los temas principales sobre los que se basa un documento (Liu et al., 2009). Las palabras clave de autor se consideran la unidad mínima de resumen en la que se pretende expresar el contenido fundamental de un documento en pocas palabras. Por este motivo, son muy valiosas para tareas como la extracción de información relevante y el procesamiento de lenguaje natural, tales como el resumen de documentos (Deng et al., 2020), la agrupación de documentos (Kim y Cho, 2020), la indexación automática (Vega-Oliveros et al., 2019), la clasificación (Rinaldi et al., 2020) o el análisis de tendencias temáticas (Mao et al., 2010; Pollack&Adler, 2015; Kim et al., 2020), entre otros.

Por otro lado, el análisis de co-ocurrencia es una poderosa herramienta para estudiar el mapa conceptual y la dinámica de un campo de investigación (Culnan et al., 1986; Mao et al., 2010; Vaughan et al., 2012; Cho, 2014; Yan et al., 2015; Ravikumar et al., 2015). El uso de esta técnica en un grupo de palabras clave tiene el potencial de mostrar los temas subyacentes de los documentos, así como sus relaciones semánticas y evolución temporal (Ronda-Pupo y Guerras-Martin, 2011).

En el análisis de co-ocurrencia de términos clave, cuanto más frecuente sea la co-ocurrencia de dos palabras clave, más fuerte será la correlación entre ellas. Sin embargo, las palabras clave extraídas de diferentes documentos suelen sufrir un alto grado de heterogeneidad, ya que la misma palabra clave se puede expresar utilizando diferentes variantes. Si se supone que las variantes del mismo término corresponden a diferentes palabras clave, la realidad de la unidad de análisis (palabras) puede distorsionarse (Muñoz-Écija et al., 2017). Para garantizar correlaciones correctas, las palabras clave deben estar unificadas antes del análisis de co-ocurrencia. Por unificación de palabras clave, nos referimos al proceso de conciliación de la terminología dispar de palabras clave en diferentes documentos. La unificación de palabras clave permite conciliar, por ejemplo, diferentes formas ortográficas y variaciones de la misma palabra clave (por ejemplo, "*life cycle*", "*life-cycle*" y "*lifecycle*") o superposiciones entre palabras clave (por ejemplo, "*project risk management*" y "*risk management*") (Ding et al., 2001).

En general, el proceso de unificar palabras clave ha recibido poca atención. La unificación de palabras clave se ha realizado principalmente de forma manual, basándose en el juicio de expertos (por ejemplo, Zhang et al., 2016; Khassed et al., 2017; Leung et al., 2017). Este proceso normalmente incluye editar, integrar, eliminar y corregir palabras clave. Sin embargo, la unificación manual de palabras clave tiene desventajas significativas como es el alto grado de subjetividad involucrado, que puede conducir a sesgos significativos en el análisis de palabras clave, o que se trata de un procedimiento bastante laborioso. La unificación preliminar mediante el pre-procesamiento de palabras clave, el cual abarca la derivación, la

lematización y la exclusión de "*stopwords*" también se ha aplicado en diferentes disciplinas (Rokaya et al., 2008; James et al., 2015; Hu et al., 2019; Kim et al., 2020). En las áreas en donde se dispone de terminologías u ontologías estandarizadas, como la biomedicina, la física o la tecnología de la información, estos recursos se pueden utilizar para unificar palabras clave. Por ejemplo, en (Ding et al., 2001), las palabras clave se unifican a través de tres *tesauros* seleccionados. Sin embargo, esta solución no es práctica si no hay terminologías estandarizadas reconocidas por la comunidad de investigación, como en el campo de la investigación en dirección de proyectos (PM). Por lo tanto, en ausencia de terminologías estandarizadas, se requiere de algún método automatizado para unificar palabras clave antes de proceder a su análisis.

Teniendo en cuenta esta problemática, en este documento, proponemos un enfoque automatizado para la unificación de palabras clave llamado KeyUnif, que combina un conjunto de técnicas léxicas, sintácticas y semánticas que son independientes del dominio de aplicación. Hasta donde sabemos, se trata del primer enfoque para unificar automáticamente las palabras clave. Nuestra propuesta ha sido probada en el contexto de la realización de un análisis de tendencias sobre la investigación en PM. Esta disciplina aporta valor sobre nuestra propuesta por dos razones principales. En primer lugar, es una amplia disciplina científica en constante evolución, no sólo por su progreso en la ampliación de conocimientos, sino por el carácter altamente multidisciplinar y la expansión a nuevos campos de aplicación (Pollack y Adler, 2015). Por esta razón, el análisis de tendencias temáticas se ha convertido en una herramienta muy valiosa para examinar las perspectivas y evolución en la investigación de PM en las últimas dos décadas (Urli, 2000; Söderlund, 2004; Smyth y Morris, 2007; Carden y Egan, 2008; Kwak y Anbari, 2009; Artto, 2009; Polack y Adler, 2015; Padalkar y Gopinath, 2016). En segundo lugar, este campo de estudio es particularmente apropiado para probar nuestro enfoque debido a la ausencia de terminologías/ontologías estandarizadas aceptadas por la comunidad investigadora.

Las principales contribuciones de este documento se resumen de la siguiente forma:

1. Proponemos un enfoque léxico, sintáctico y semántico que puede unificar eficazmente las variantes de las palabras clave en términos base, independientemente del dominio de aplicación.

2. La inclusión de una etapa de unificación de palabras clave en el análisis de tendencias de la investigación de gestión de proyectos muestra una mejora sustancial en la coherencia de los temas extraídos.

3. Los resultados experimentales de nuestro enfoque revelan que la unificación de palabras clave reduce el número de términos de significado amplio (generales) que suelen aparecer en una red de co-ocurrencia.

4. Nuestro enfoque proporciona la primera herramienta práctica que automatiza completamente la extracción de tendencias emergentes en el campo de la investigación en Gestión de Proyectos.

5. Nuestros resultados muestran la viabilidad de desarrollar herramientas orientadas al usuario final que automatizan el proceso de unificación de palabras clave y al estudio de las tendencias temáticas.

El resto del capítulo se organiza de la siguiente forma. En la sección 2 se detalla el trabajo relacionado. En la sección 3 se presentan las técnicas y el método general propuesto. La sección 4 muestra los resultados del experimento. Y, finalmente, se debaten los resultados obtenidos y se exponen las principales conclusiones extraídas y el trabajo futuro.

6.2 TRABAJO RELACIONADO

Con el crecimiento exponencial de los datos en la literatura, cada vez más disciplinas científicas están utilizando enfoques científicos para analizar el conocimiento del dominio. Entre estas disciplinas se encuentra PM, un amplio campo que incluye todas las ramas de la ingeniería. PM se puede definir como un enfoque productivo que

aumenta nuestro conocimiento sobre las empresas u organizaciones modernas (Söderlund, 2004a). Actualmente, PM todavía no es una disciplina bien establecida, ya que está en constante cambio y evolución con el fin de proporcionar un área más especializada para llevar a cabo los diversos proyectos que enfrentamos en el día a día. Además, la investigación de PM ha evolucionado mucho en los últimos años, y se espera que esta tendencia continúe en el futuro (Padalkar y Gopinath, 2016). Percibir cómo se está configurando la investigación de PM para imaginar cualquier cambio importante en el campo es un desafío importante. Por lo tanto, en las últimas dos décadas, se han propuesto numerosos estudios que se centran en el análisis de las tendencias temáticas en PM (Pollack y Adler, 2015). Estos estudios han aplicado diferentes enfoques que van desde los procedimientos manuales, tales como la revisión de la literatura (Söderlund, 2004; Smyth y Morris, 2007; Carden y Egan, 2008; Kwak y Anbari, 2009; Padalkar y Gopinath, 2016), al uso de técnicas específicas como el análisis bibliométrico, el análisis estadístico o el análisis de co-ocurrencia de palabras (Urli, 2000; Artto, 2009; Polack y Adler, 2015).

6.2.1 Análisis de co-ocurrencia de palabras clave

El análisis de co-ocurrencia de términos clave de autor examina la relación entre la terminología utilizada en varias partes de un documento, incluido el título, el resumen y las palabras clave del autor. Las palabras clave de autor tienen la ventaja de caracterizar correctamente el contenido de un artículo (Cambrosio et al., 1993). La alta frecuencia de una palabra clave sugiere que el tema al que se refiere es relevante en el dominio (Khassed et al., 2017). Dos o más palabras clave que coexisten en el mismo documento implica que existe un vínculo entre los temas a los que se refieren (Urli, 2000; Themistocleous, 2000; Crawford, 2006; Artto, 2009; Pollack y Adler, 2015), y la presencia de muchas co-ocurrencias alrededor del mismo conjunto de palabras es un indicativo de que el conjunto puede corresponder a un tema de investigación (Cambrosio et al., 1993). Por lo tanto, el análisis de co-ocurrencia de palabras clave revela el mapa conceptual de cualquier disciplina (Khassed et al., 2017).

En el análisis de co-ocurrencia, se pueden identificar tres etapas bien definidas (Ding et al., 2001): la recopilación del conjunto de datos de entrada, la unificación/selección de palabras clave y el análisis/visualización de los resultados. En la primera etapa, el origen de datos de entrada se selecciona según un conjunto de criterios predefinidos. Una vez que se ha extraído un gran número de palabras clave de la fuente de datos de entrada, se requiere un proceso de unificación para armonizar la disparidad en la terminología (Ding et al., 2001) y un proceso de selección para extraer los términos pertinentes en el campo de análisis (Van Eck et al., 2010). La unificación de palabras clave reduce significativamente el número de términos diferentes para designar la misma palabra clave, lo que conduce a una representación más clara y precisa del mapa conceptual de un dominio (Muñoz-Écija et al., 2017). Por último, en la tercera etapa, se crea la red de co-ocurrencia de palabras clave, y a través de su análisis, como el análisis de clúster, se pueden visualizar las tendencias temáticas de un dominio (Cambrosio et al., 1993).

6.2.2 Unificación/selección de palabras clave

Hoy en día, la unificación/selección de palabras clave se realiza principalmente por expertos en el dominio en cuestión, que dedican tiempo y esfuerzo a unificar y seleccionar palabras clave manualmente, antes de continuar con el análisis. La unificación/selección de palabras clave basada únicamente en su frecuencia de aparición a menudo produce una gran cantidad de palabras clave con pocos o ningún significado específico del dominio (Van Eck et al., 2010). Sin embargo, la unificación/selección manual de palabras clave consume mucho tiempo e implica un alto grado de subjetividad, lo que puede sesgar significativamente la red de palabras clave.

El proceso de selección de palabras clave ha recibido cierta atención de la comunidad científica en el pasado. Aunque hay autores que utilizan el conjunto completo de palabras clave para crear la red de co-ocurrencia (Smith y Morris, 2007), la mayoría de las obras preseleccionan un subconjunto utilizando diferentes métodos. Por ejemplo, muchos autores confían en la Ley de Zipf para eliminar palabras clave que son demasiado comunes o poco utilizadas (Kim et

al., 2020). En (Van Eck et al., 2010), el análisis semántico probabilístico (Hofmann, 2001) se utiliza para identificar los temas principales de un corpus de documentos, y luego sólo se seleccionan los términos que están fuertemente asociados con estos temas principales. En (Chen y Xiao, 2016), se han comparado tres métodos alternativos: la frecuencia de los términos (TF), la frecuencia inversa del documento (TF-IDF) y el índice de actividad de palabras clave (TF-KAI). El último método obtiene el mejor rendimiento en los experimentos realizados. En cambio, en (Noh et al., 2015) se sugiere un conjunto de directrices, que se basan en el uso de redes neuronales y análisis semántico, para seleccionar y procesar palabras clave en el ámbito específico del análisis de patentes.

Por otro lado, la unificación de palabras clave se puede comparar con el proceso de normalización del vocabulario en el reconocimiento de entidades con nombre biomédico, que generalmente se lleva a cabo a través de técnicas de procesamiento del lenguaje natural en tres niveles lingüísticos: léxico, sintáctico y semántico (Atkinson y Bull, 2012). En el contexto de la unificación de palabras clave, se han aplicado técnicas léxicas y sintácticas básicas para pre-procesar palabras clave. Ejemplos son el *stemming*/lematización (James et al., 2015; Van Eck et al., 2010; Rokaya et al., 2008, Hu et al., 2019; Kim et al., 2020), identificación/exclusión de palabras y números, o identificación de categorías léxicas (por ejemplo, sustantivos, verbos, etc.) mediante el etiquetado de los textos (Van Eck et al., 2010, Kim et al., 2020).

Las técnicas semánticas basadas en terminologías u ontologías estandarizadas también son recursos valiosos para unificar palabras clave en los campos donde están disponibles. En (Ding et al., 2001), las palabras clave se unifican con el apoyo de tres *tesauros* en el campo de las tecnologías de la información. En este caso, los encabezados de los *tesauros* se explotan para detectar la misma palabra clave que se ha expresado de manera diferente en singular y en plural, utilizando sinónimos o antónimos alternativos, y diferentes homónimos. En (Muñoz-Écija et al., 2017), se construye un tesauro específico con 2.000 correspondencias en el área de la nano ciencia para la identificación de las siglas. Sin embargo, estas soluciones no son

factibles en ausencia de recursos o teniendo que realizar nuevos desarrollos, como en nuestro caso.

6.2.3 Análisis de las tendencias temáticas en PM

PM se ha definido como un enfoque productivo para aumentar nuestro conocimiento sobre las empresas u organizaciones modernas (Söderlund, 2004a). En las últimas décadas, varios investigadores han aplicado diferentes técnicas sobre las palabras clave de autor extraídas de publicaciones para analizar las tendencias temáticas de investigación en el campo de la gestión de proyectos (tabla 6.1). Las metodologías seguidas en este campo van desde métodos completamente manuales como la revisión de la literatura (Smyth y Morris, 2007; Carden y Egan, 2008; Kwak y Anbari, 2009) a métodos automáticos ampliamente aceptados en la comunidad científica como las redes de co-ocurrencia de términos (Urli y Urli, 2000; Crawford et al., 2006; Artto et al., 2009; Pollack y Adler, 2015). Aunque la revisión de la literatura es un método que todavía se utiliza ampliamente hoy en día (Padalkar y Gopinath, 2016), estudios recientes están más orientados hacia procedimientos que automatizan gran parte del proceso, debido al bajo coste, precisión aceptable y eliminación de la subjetividad del autor en la realización del análisis.

Año	Estudio	Fuente de datos	Información utilizada	Técnicas
2000	Urli y Urli (2000)	Base de datos ABI-INFORM con término de consulta: “project management”	Palabras clave extraídas de referencias	Análisis de la red de co-ocurrencia con Leximappe
2006	Crawford et al. (2006)	IJPM y PMJ utilizando el término de consulta: gestión de proyectos y exclusión de revisiones de libros, entre 1994 y 2003	Palabras clave extraídas de textos	Análisis de palabras clave y lenguaje de corpus
2009	Artto et al. (2009)	23 Business Journals con términos de consulta: “program”, “project”	Referencias y palabras clave de autor	Análisis de la red de co-ocurrencia y análisis de frecuencia de palabras clave utilizando UCINET

2015	Pollack and Adler (2015)	Base de datos ISI WoS con término de consulta: "project management" en la categoría denominada topic	Palabras clave de autor y resúmenes	Análisis de la red de co-ocurrencia y análisis de frecuencia de términos en resúmenes usando Sci Tool
2016	Chen et al. (2016)	Palabras clave extraídas de proyectos de fondos: 7304	Palabras clave del autor	Análisis de co-ocurrencia

Tabla 6.1: Estudios previos sobre tendencias de PM basados en el análisis de palabras clave de autor.

6.3 MÉTODOS

La unificación de palabras clave requiere la identificación de las diferentes variantes que pueden producirse en el conjunto de palabras clave en un dominio concreto. En primer lugar, una variante de una palabra clave puede expresar exactamente el mismo significado que la palabra clave base. Un proceso de unificación de palabras clave ideal debe ser capaz de contraer todas las variantes que conservan el significado en las palabras clave base. El *stemming* y la lematización reducen las diferentes variantes a su forma base, facilitando la unificación. Sin embargo, existen más tipos de variantes más allá de los prefijos, sufijos e interfijos. Por ejemplo, la palabra clave *"risk of project"* es una variante que preserva el significado en la palabra clave "*project risk*" (figura 6.1), aunque podríamos encontrar variantes de este tipo que no conservan el significado de su forma base. El uso indistinto de las siglas o de los términos equivalentes extendidos es otro caso de variantes que preservan el significado. Por ejemplo, el acrónimo "*PRM"* se puede unificar en la palabra clave equivalente extendida "project risk management" en la figura 6.1. Las variantes ortográficas con el uso indistinto de guiones, espacios en blanco o ningún espacio entre términos también son habituales. Por ejemplo, "*life cycle*", "*life-cycle*" y "*lifecycle*" son variaciones del mismo término. En segundo lugar, una variante de una palabra clave también puede expresar un significado más amplio (general) o más estrecho (específico) que la palabra clave base. Por ejemplo, la palabra clave "*project risk management"* tiene un significado más concreto

("estrecho") que la palabra clave "*project risk*" (figura 6.1). La primera palabra clave puede verse como una variante de la segunda, expresando el significado de la palabra clave base aumentada con otra información semántica específica, pero ambas hacen referencia al mismo contexto de aplicación.

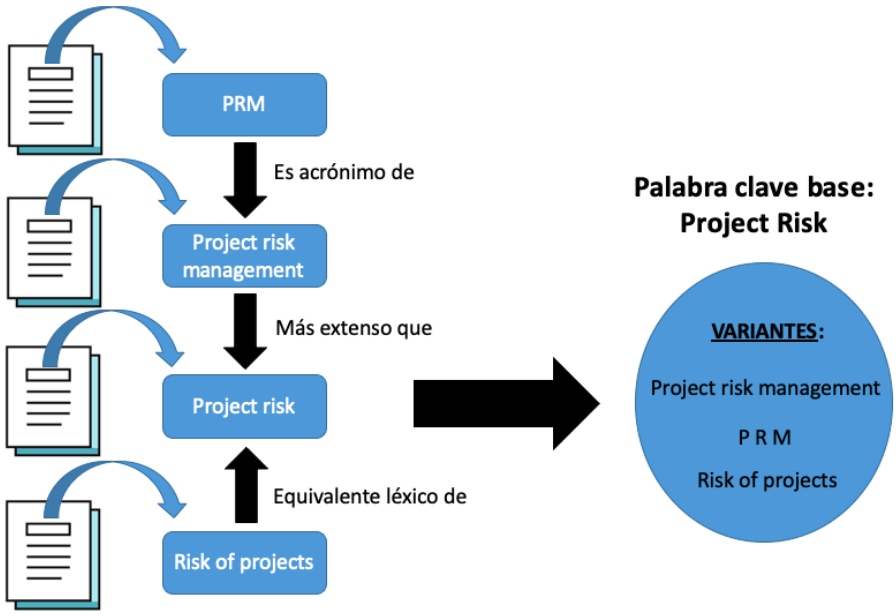

Figura 6.1: Ejemplo de unificación de palabras clave de autor en el campo de PM.

Hay dos problemas que surgen al unificar una palabra clave más estrecha o más amplia. El primer problema es decidir si una palabra clave es un término base o variante. Las palabras clave de alta frecuencia pueden ser de interés, y su unificación en otras palabras clave más amplias/estrechas de baja frecuencia las ocultaría en el mapa de tendencias temáticas. Por lo tanto, en este trabajo, las palabras clave de alta frecuencia se tratan como términos base, mientras que las palabras clave de baja frecuencia se toman como variantes. La segunda pregunta es saber si una palabra clave es realmente una variante de una palabra clave base. En este trabajo, para evitar errores resultantes de la ambigüedad de los términos, sólo las palabras clave de baja frecuencia

que comprenden más de dos palabras se unificaron en palabras clave de alta frecuencia de dos o más palabras. Por ejemplo, la palabra clave de baja frecuencia "*project risk analysis*" (que sólo apareció una vez como palabra clave de autor en el conjunto de documentos analizados, véase doi:10.1016/j.ijproman.2007.02.004) se unificó en la palabra clave más amplia "*project risk*", que se trató como un término base ya que su frecuencia era superior (aparecía en un total de 7 documentos diferentes). Hay que tener en cuenta que el término base "*project risk*" claramente identifica el contenido del documento etiquetado por sus autores con la variante "*project risk analysis*".

El resto de la sección cubre los aspectos técnicos clave del método de unificación de palabras clave. La sección 6.3.1 describe las técnicas léxicas, sintácticas y semánticas utilizadas y la sección 6.3.2 detalla el método completo. La sección 6.3.3 presenta el procedimiento seguido para el análisis de las tendencias temáticas con los resultados obtenidos del método de unificación.

6.3.1 Técnicas léxicas, sintácticas y semánticas

Se desarrolló un conjunto de técnicas independientes del dominio para la unificación de palabras clave (figura 6.2). Llamamos a los métodos léxicos aquellas técnicas que permiten la unificación de palabras clave mediante el uso de propiedades léxicas de los términos como unigramas, bigramas y variantes léxicas. Nos referimos a los métodos sintácticos como aquellos que utilizan las unidades básicas de composición de frases y analizan la relación entre ambos. Además, utilizan características sintácticas que forman parte de las etiquetas de voz (POS) y componentes de un árbol de análisis sintáctico (nombre, verbo, adjetivo, preposición...). Los métodos semánticos hacen uso del significado implícito o latente y las connotaciones de los términos de palabra clave, teniendo en cuenta el contexto en el que aparecen.

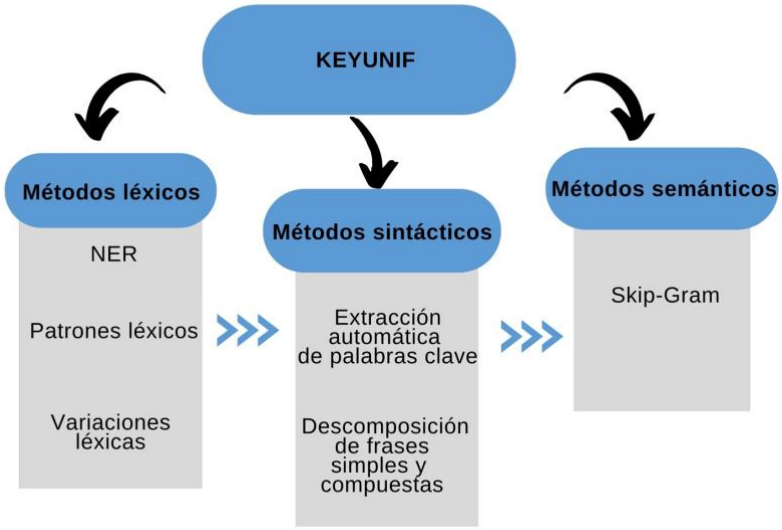

Figura 6.2: Técnicas propuestas para unificar palabras clave.

En nuestro enfoque se implementaron tres tipos diferentes de técnicas léxicas: basadas en patrones léxicos, reconocimiento de variantes ortográficas y reconocimiento de entidades nombradas. En primer lugar, las técnicas léxicas basadas en patrones detectan y unifican automáticamente acrónimos, variantes más amplias/estrechas y variantes ortográficas. En cuanto a las siglas, proponemos un algoritmo sencillo basado en (Pustekhovsky et al., 2001), que detecta pares de acrónimo y término extendido donde un acrónimo es el término corto para la palabra clave. La figura 6.3 muestra un ejemplo de aplicación de nuestro algoritmo. Un acrónimo se identifica en el texto cuando el término está en mayúsculas y no tiene una longitud superior a cinco y, además, suele estar entre paréntesis (Paso 1, figura 6.3). Para cada candidato identificado se crea automáticamente una expresión regular para buscar una secuencia de texto donde cada uno de los caracteres dentro del acrónimo coincide con el primer carácter de cada palabra que compone el equivalente en término extendido (Paso 2, figura 6.3). La búsqueda se limita al resto de palabras clave y al resumen donde se ha detectado el acrónimo (Paso 3, figura 6.3). Entonces, si hay más de una

palabra clave que coincide con el mismo acrónimo (Paso 4, figura 6.3), se escoge la palabra clave que aparece junto al acrónimo más veces (Paso 5, figura 6.3). Los patrones léxicos se elaboraron automáticamente a partir de los términos base y luego se utilizaron para identificar y unificar variantes más amplias/estrechas. En este caso se incluyen también las palabras con guiones y sus variantes, ya que pueden aparecer con él, sin él o con un espacio en blanco en su lugar.

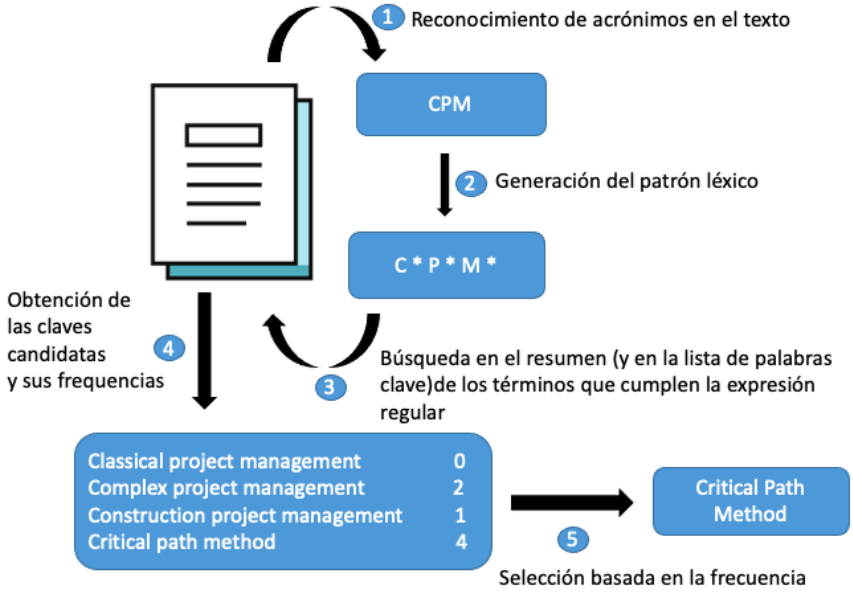

Figura 6.3: Identificación léxica del par acrónimo y término extendido basada en patrones.

Por otro lado, las variantes ortográficas se producen cuando una palabra tiene más de una forma ortográfica aceptada, principalmente se produce entre las convenciones británicas y americanas, ya que las publicaciones científicas están mayoritariamente en inglés. Se desarrolló un algoritmo sencillo para la *unificación de las variantes ortográficas,* con el fin de unificar las palabras clave del autor cuyo lexema difiere por una sola letra consonante (por ejemplo, "*organization*"– "*organisation*"), o aquellas que sólo son diferentes en que se intercambian dos letras consecutivas (por ejemplo, "*center of*

excellence" – "*centre of excellence*") o aquellas que contienen una letra más que la otra (por ejemplo, "*behaviour*" – "*behavior*"). En tercer lugar, *Named Entity Recognition* (NER) implica el procesamiento de texto y la identificación de expresiones que se refieren a pueblos, lugares, organizaciones y empresas (J.Atkinson y V. Bull, 2012). Aunque en el área de estudio la terminología no incluye estas entidades nombradas, su identificación y unificación son importantes en dominios altamente interdisciplinarios, como PM. La unificación de los topónimos en una sola palabra clave llamada *toponyms* se realizó utilizando la biblioteca llamada GeoText[8].

Además de los patrones léxicos, también se utilizaron técnicas sintácticas basadas en la extracción automática de frases clave (Merrouni et al., 2019) para complementar la unificación de variantes más amplias/estrechas. Las frases clave se extrajeron automáticamente del título y se anotaron los resúmenes de los artículos, utilizando SingleRank (Wan y Xiao, 2008), un modelo de clasificación basado en grafos que utiliza el etiquetado POS. Si el conjunto de claves extraídas incluía una palabra clave de autor de alta frecuencia más estrecha que la variante, se identifica como un término base y, a continuación, la variante se unifica en ella. Por ejemplo, en el artículo identificado por doi:10.1016/j.ijproman.2015.09.004, la palabra clave "*cash flow analysis*" *(*frecuencia 1 en el corpus) se unificó en "*cash flow*" *(*frecuencia 5 en el corpus), la cual era una de las palabras clave extraídas del título/resumen del documento por SingleRank. Esta técnica es contextual, ya que tiene en cuenta el contexto específico en el que la palabra clave de autor aparece dentro del texto, a diferencia de las otras técnicas léxico-sintácticas que están libres de contexto. Del mismo modo, las palabras clave estrechas y de baja frecuencia se unificaron en palabras clave de frecuencia más amplias y más frecuentes que se extrajeron automáticamente del título y del resumen. Por ejemplo, en el documento identificado por doi: 10.1016/j.ijproman.2012.09.001, el término clave "*is project team*" se unificó en "*project team*", una palabra clave extraída por *SingleRank*. También se aplicaron técnicas sintácticas para unificar frases compuestas a frases simples de sustantivos. En este caso, restringimos

[8] https://pypi.org/project/geotext/

las frases compuestas a los casos más frecuentes: 1) una secuencia de una frase de sustantivo simple y una frase de sustantivo proposicional (*de*), y 2) una secuencia de dos frases de sustantivo simples vinculadas por la conjunción *y*. Ejemplos de estos son "*success factor of project*", que se unificó en *"project success factor"*, y *"hard and soft project"*, que se unificó en las dos palabras clave "*hard project"* y "*soft project"*.

Con respecto a las técnicas semánticas, éstas incorporaron el modelo de Skip-Gram (Hu et al, 2019), el cual proporcionaba un método indirecto para unificar palabras clave significativas mediante el análisis del texto completo en las publicaciones científicas de PM. En primer lugar, Skip-Gram predijo las palabras del contexto para cada palabra clave de destino. A continuación, se utilizó la similitud del coseno de los vectores de palabras entrenados por dicho modelo para medir la similitud semántica entre las palabras clave de los documentos PM. Por último, se seleccionaron las claves con mejor puntuación. En nuestro caso, el valor mínimo de similitud asignado tenía que ser superior a 0,8 ya que debido a la experiencia adquirida en numerosos ensayos era el valor a partir del cual se generaba menor porcentaje de error. Esto permitió lograr una similitud semántica razonablemente precisa, ya que nos dejaba el menor margen de error. Dicho modelo de Skip-Gram se implementó en paquetes de Python para NLP y aprendizaje automático, incluidos NLTK y Gensim.

6.3.2 El algoritmo propuesto para la unificación de palabras clave

En esta sección, describimos el algoritmo completo que desarrollamos para la unificación de palabras clave. Los principales pasos involucrados en nuestro enfoque fueron: 1) pre-procesamiento y filtrado inicial, 2) procesamiento léxico y procesamiento sintáctico y 3) procesamiento semántico. En primer lugar, las palabras clave se procesaron previamente eliminando las "*stopwords"* (excepto los términos que incluían la preposición "de" o la conjunción "y", que ya se consideraban en la unificación) y los números de las palabras clave, así como la aplicación de *stemming*. Las palabras específicas al dominio, que no eran relevantes en el análisis de tendencias (como

"*project*" y "*project management*" en PM), se excluyeron del análisis (Pollack y Adler, 2015; Kim et al., 2020). Basándonos en la frecuencia de aparición de las palabras clave en un conjunto de documentos, utilizamos la ley de *Zipf* para eliminar palabras clave de baja frecuencia y de una sola palabra. También se eliminaron aquellas palabras clave que consistían en una secuencia de términos separados por comas, debido a que ya existían como palabras clave independientes. En segundo lugar, las palabras clave fueron unificadas por las técnicas léxicas y sintácticas desarrolladas. Y en tercer y último lugar, las palabras clave se unificaron mediante técnicas semánticas basadas en el modelo de Skip-Gram. Cabe destacar que el proceso de unificación fue completamente automático.

6.3.3 Análisis de tendencias basado en palabras clave

En esta sección, describimos el ámbito en el que hemos probado la utilidad de nuestro algoritmo, es decir, el procedimiento para el análisis de tendencias en el dominio PM. El proceso realizado consiste en los siguientes pasos:

Adquisición de datos. Se tomaron los resúmenes y los documentos completos relacionados con el dominio siguiendo los criterios más adecuados.

Unificación de palabras clave. Las palabras clave de autor se extrajeron y unificaron con el método propuesto. El nombre de la palabra clave base se asignó como nombre del conjunto unificado. Si se obtuvieron dos variantes diferentes para un mismo conjunto en un documento, solo se contaron una única vez. Por ejemplo, si un documento (por ejemplo, doi: 10.1016/j.ijproman.2017.01.0109) contiene la variante "*project risk management*" y el término base "*project risk*", la palabra clave base se cuenta una única vez. La co-ocurrencia entre conjuntos unificados existe cuando al menos una palabra clave (término variante o base) de cada conjunto unificado coexiste en el mismo documento.

Análisis de red de co-ocurrencia. Se elaboró una red de co-ocurrencia de términos clave no dirigida utilizando el software de código abierto Gephi (versión 0.9.2), donde cada nodo era un conjunto de palabras clave unificadas y cada relación era una co-ocurrencia

significativa entre dos conjuntos de palabras clave unificadas. Se asignó un peso de N a cada arista entre dos conjuntos de palabras clave unificadas siempre y cuando coexistieran en los documentos N veces (Pollack y Adler, 2015). Utilizando el algoritmo de diseño ForceAtlas2 (Jacomy et al., 2014) proporcionado por Gephi, la red semántica de palabras clave unificadas se transformó y visualizó en un mapa conceptual.

Extracción de las subredes emergentes. Utilizando la propiedad de modularidad de la red de Gephi, la red semántica se dividió en subredes. Las subredes extraídas representaban los clústeres temáticos, es decir, grupos de palabras clave fuertemente interconectados.

Detección de tendencias temáticas. La evolución temporal de los diferentes temas extraídos a partir de la red de co-ocurrencia se visualizó por la frecuencia con la que estos aparecieron en cada una de las publicaciones realizadas en el tiempo. Se consideró que una temática se utiliza en una publicación si alguna de las palabras clave que componían dicha temática estaba presente. Además, se utilizó la técnica de detección de ráfagas o *burst* (Kleinberg, 2003), ya vista en el capítulo 3, para identificar períodos de tiempo en los que un tema era excepcionalmente popular.

6.4 RESULTADOS

En esta sección, evaluamos experimentalmente nuestro método automatizado de unificación de palabras clave. Nos centramos especialmente en el rendimiento de nuestro enfoque en el contexto del análisis de tendencias en la investigación de PM.

6.4.1 Conjunto de datos en el campo de PM

Padalkar y Gopinath (2016) observaron que las revistas revisadas por pares centradas en PM, como la *International Journal of Project Management* (IJPM), la *Project Management Journal* (PMJ) o la *International Journal of Managing Projects in Busines*s (IJMPB), representan la mayor parte de las publicaciones en el área. Para evaluar KeyUnif, recopilamos todos los artículos publicados en IJPM durante un período de 19 años entre 2000 y 2018. Seleccionamos esta revista por su relevancia en el área, el alto número de publicaciones que la

revista ofrece cada año, y que era la única revista revisada por pares centrada en PM a la que teníamos acceso completo a sus publicaciones. En total, se extrajeron para su análisis 1.612 artículos que abarcaban 7.508 palabras clave de autor (figura 6.4). De los documentos extraídos, 1.582 tenían al menos una palabra clave de autor. La mayoría de los autores utilizaron entre 4 y 5 palabras clave por artículo (figura 6.5). La mayoría de las palabras clave (54,97%) eran de dos palabras y se encontraron en el 91,84% de las publicaciones. Por el contrario, menos del 20% eran palabras clave de una o tres palabras, y solo el 7,21% implicaba palabras clave de cuatro o más términos. Aun así, las palabras clave de una sola palabra representaron el 63,15% de las publicaciones. Por lo tanto, los autores utilizan a menudo palabras clave con significado muy amplio que puede variar en función del contexto de uso. Como los autores pueden etiquetar documentos con varias palabras clave, el porcentaje total de publicaciones no suma el 100% (última columna en la tabla 6.2).

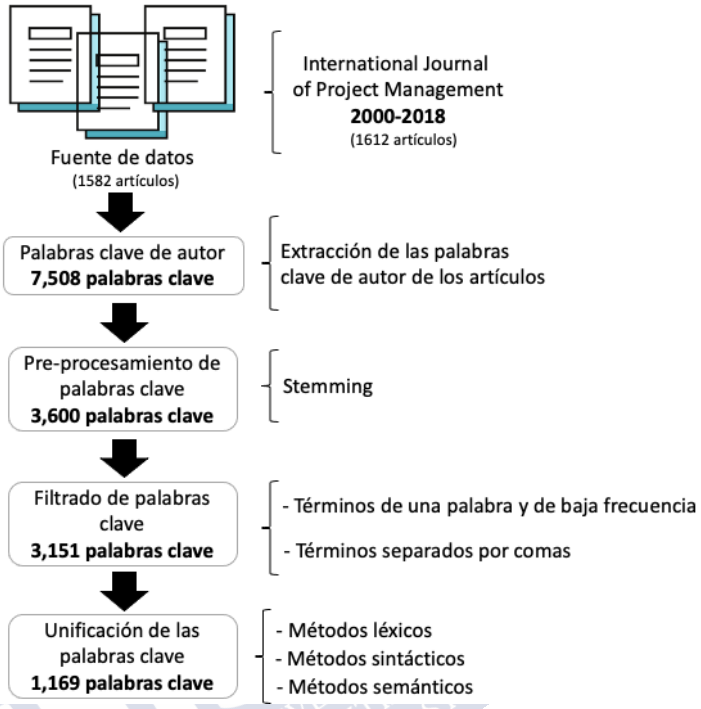

Figura 6.4: Resumen de las etapas del proceso de unificación de palabras clave en el área de PM.

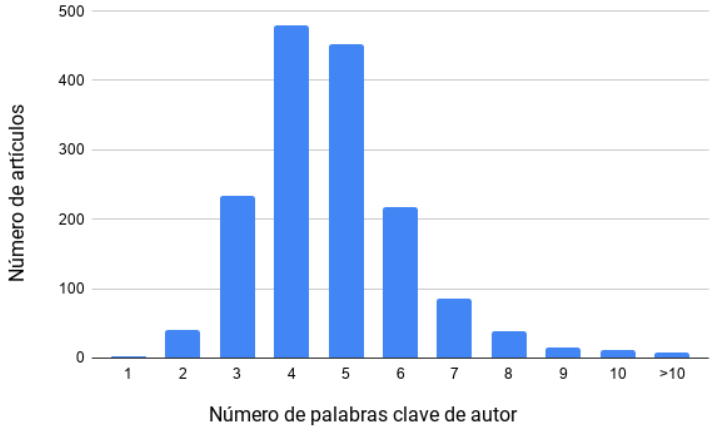

Figura 6.5: Distribución del número de publicaciones en función del número de palabras clave de autor que contienen.

	Nº palabras clave	% Palabras clave	Nº publicaciones	% publicaciones
1 palabra	662	18,39%	999	63,15%
2 palabras	1979	54,97%	1453	91,84%
3 palabras	699	19,43%	786	49,68%
4 o más palabras	260	7,21%	259	16,37%
Total	**3600**	**100%**	**1582**	**-**

Tabla 6.2: Número de palabras clave y publicaciones según el número de palabras que componen un término clave.

6.4.2 Métricas de evaluación

La evaluación de un algoritmo de unificación de palabras clave es un reto, ya que no hay estándares ampliamente aceptados sobre cómo se debe realizar dicha evaluación. Por lo tanto, proponemos evaluar nuestro método de dos formas. En primer lugar, realizamos una evaluación basada en los conceptos conocidos de precisión y *recall*. Dado que no había un estándar de referencia y no había suficientes recursos para crear uno, un experto evaluó el rendimiento del algoritmo de unificación. El procedimiento de evaluación consistió en revisar manualmente un subconjunto de todos los conjuntos unificados de palabras clave obtenidos por el método. En este contexto, la precisión se definió como la fracción entre el número de palabras clave correctas y el número total de palabras clave propuestas por nuestro método de unificación; y el *recall* como la fracción del número de palabras clave correctamente unificadas y el número total de palabras clave relevantes (es decir, las palabras clave creadas y unificadas de forma manual).

A continuación, realizamos una segunda evaluación mediante 1) la construcción de una red de co-ocurrencia de términos clave, 2) extrayendo los temas (es decir, subredes emergentes) y 3) calculando la métrica de evaluación intrínseca, ya vista en el anterior capítulo (sección 5.3.4), conocida como *topic coherence*.

6.4.3 Unificación de palabras clave

En total, se utilizaron 1.612 documentos que cubren 7.508 palabras clave de autor para analizar las tendencias en la investigación de PM desde enero de 2000 hasta diciembre de 2018. Después de la fase inicial

de pre-procesamiento de términos, el conjunto de palabras clave de autor se redujo a 3.600 palabras clave únicas (tabla 6.2). Como se puede observar, las 3.244 (43,2%) palabras clave de frecuencia superior a 5 se redujo drásticamente a 269 (7,5%) después de eliminar duplicados y realizar pre-procesamiento.

Frecuencia	Nº de palabras clave de autor en el conjunto de datos original (incluidas las repeticiones)	%	Nº de palabras clave de autor únicas después de eliminar duplicados y pre-procesamiento	%
> 100	299	3,98%	1	0,03%
> 50	556	7,41%	5	0,14%
> 40	696	9,27%	9	0,25%
> 20	1.334	17,77%	33	0,92%
> 10	2.123	28,28%	94	2,61%
>5	3.244	43,21%	269	7,47%
> 2	4.790	63,80%	883	24,53%
>1	7.508	100%	3.600	100%

Tabla 6.3: Distribución de las palabras clave de autor en el conjunto de datos original y después de pre-procesar y eliminar duplicados.

Después del pre-procesamiento, detectamos que entre las palabras clave más frecuentes (> 40), había palabras clave no significativas específicas del dominio (como "*project*" y "*project management*"), las cuales fueron eliminadas. Entre las palabras clave con una frecuencia inferior a 5, también se eliminaron las palabras clave de una sola palabra (como "*Vignette*" o "*Aristotle*") y las que consistían en una secuencia de términos separados por comas (como "*managing design, planning and appraisal*" o "*design, novate and construct*"). Después de filtrar todas estas palabras clave, el conjunto disminuyó a 3.151 palabras clave únicas (figura 6.4).

La tabla 6.3 muestra el número de palabras clave unificadas después de aplicar cada técnica propuesta. En total, se identificaron y unificaron 853 variantes en 316 palabras clave base, lo que dio como resultado 1.169 (37%) palabras clave unificadas. Las técnicas léxicas identificaron 643 (75,4%) variantes y las técnicas sintácticas 163

(19,1%) variantes. El modelo de Skip-Gram se entrenó sobre los 1.612 documentos completos iniciales, y se seleccionaron únicamente variantes con una puntuación superior a 0,8 en coeficiente de similitud para la unificación. En total, las técnicas semánticas identificaron 47 (5,5%) variantes. Hay que tener en cuenta que hubo superposición entre los métodos para alguna de las claves, y las unificaciones realizadas con las técnicas léxicas no se tuvieron en cuenta en el resto de las técnicas. En resumen, los patrones léxicos identificaron la mayoría de las variantes (67,5%).

Método	Técnica	Nº palabras clave unificadas	% palabras clave unificadas	Total	% palabras clave unificadas
Léxico	Patrones léxicos	576	89,58%	643	75,38%
	Reconocimiento de entidades nombradas	41	6,37%		
	Variaciones léxicas	26	4,04%		
Sintáctica	Palabras clave de frase simple/compuesta	45	27,61%	163	19,11%
	Basado en SingleRank	118	72,39%		
Semántica	Basado en Skip-Gram (Word2vec)	47	100%	47	5,51%
			Total	**853**	**100%**

Tabla 6.4: Número y porcentaje de variantes de las palabras clave identificadas por cada técnica.

La tabla 6.5 muestra la precisión y el *recall* de nuestro enfoque de unificación, que logró una precisión del 91,19% y un 76,88% de *recall*, después de revisar 120 (38%) palabras clave unificadas sobre 316. En total, el número de palabras clave revisadas fue de 467 (43%) de un total de 1.097 palabras clave unificadas. Como se muestra en la tabla 6.5, la mayoría de los conjuntos unificados incluyeron dos palabras clave. Específicamente, nuestro procedimiento obtuvo 201 (63,61%) conjuntos unificados de dos palabras clave de un total de 316 conjuntos unificados. De ellos, revisamos 61 (19%) conjuntos, alcanzando una precisión del 97,47% y un *recall* del 88,36%. La precisión más alta (100%) se logró en los conjuntos unificados de 6 palabras clave, y el

recall más alto (100%) en los conjuntos unificados de 5 palabras clave, aunque en ambos casos el porcentaje de palabras clave revisadas fue el más bajo de todos los conjuntos unificados debido a que tampoco fueron los más mayoritarios. La precisión más baja (78,21%) se alcanzó para agrupaciones de más de 7 palabras clave, y el *recall* más bajo (41,72%) para conjuntos de 7 palabras clave.

Nº de palabras clave por cada conjunto unificado	Nº de conjuntos unificados	Nº de conjuntos unificados revisados	Nº de palabras clave unificadas	Nº de palabras clave revisadas	Precisión	Recall
2	201	61 (19%)	418	122 (11%)	97,47%	88,36%
3	57	28 (9%)	173	84 (8%)	98,77%	85,69%
4	19	9 (3%)	77	36 (3%)	97,22%	78,31%
5	13	5 (2%)	66	25 (2%)	88%	100%
6	4	3 (1%)	24	18 (2%)	100%	90%
7	9	5 (2%)	64	35 (3%)	97,14%	41,72%
> 7	10	9 (3%)	275	147 (13%)	78,21%	89,58%
Total	316	120 (38%)	1097	467 (43%)	91,19%	76,88%

Tabla 6.5: Precisión y *recall* del proceso de unificación de palabras clave.

6.4.4 Análisis de tendencias temáticas

En primer lugar, el Apéndice C detalla las puntuaciones obtenidas para la medida de *topic coherence* alcanzadas en las tres redes de co-ocurrencia resultantes de 1) stemming y filtrado (modelo pre-procesado), 2) unificación léxica y sintáctica (modelo unificado léxicamente y sintácticamente) y 3) unificación semántica (modelo totalmente unificado). Al evaluar la coherencia de cada temática identificada, calculamos los valores de UMass con el top de las 10 palabras clave de cada temática con mayor grado medio ponderado obtenido en la red. En la tabla 6.6 se resume el número de tendencias temáticas identificadas para cada modelo, identificando las temáticas de baja y de alta calidad. Las temáticas con una puntuación de coherencia temática entre -2 y 2 fueron calificados como alta calidad, y el resto de baja calidad.

Para el modelo pre-procesado se obtuvieron 8 tendencias temáticas sin demasiada coherencia. El modelo unificado léxica y sintácticamente produjo 9 tendencias temáticas con 5 (55,6%) muy coherentes. Y, por

último, el modelo unificado completo produjo 9 tendencias temáticas con 6 (66,7%) tendencias muy coherentes.

	Nº de tendencias temáticas	Nº de tendencias temáticas de alta calidad	Nº de tendencias temáticas de baja calidad
Modelo pre-procesado	8	0 (0%)	8 (100%)
Modelo unificado léxica y sintácticamente	9	6 (66.7%)	3 (55.5%)
Modelo totalmente unificado	9	7 (77.8%)	2 (22.2%)

Tabla 6.6: Número de tendencias temáticas de alta calidad y baja calidad basadas en las puntuaciones de coherencia de UMass en los tres modelos analizados.

La figura 6.6 resalta en diferentes colores los principales clústeres (temas) identificados en la red de co-ocurrencia totalmente unificada. La tabla 6.7 muestra estos nueve temas en detalle, con dos de ellos de baja calidad y siete de alta calidad. Para cada uno de ellos un experto eligió etiquetas que resumían brevemente el enfoque del tema, utilizando de base la guía de proyectos del PMI, el PMBoK (PMI, 2017). Entre los temas de alta calidad, "*Project Environment and Integration"* y "*Success Factors"* fueron los temas más *frecuentes*, en términos de porcentaje de nodos en la red, durante el período 2000-2018. Los temas restantes de alta calidad incluyen (en orden descendente de prevalencia): "*Project strategy"*, *"Stakeholder Management"*, *"Risk Management"*, *"Human Resources Management"* y "*Governance and Control"*.

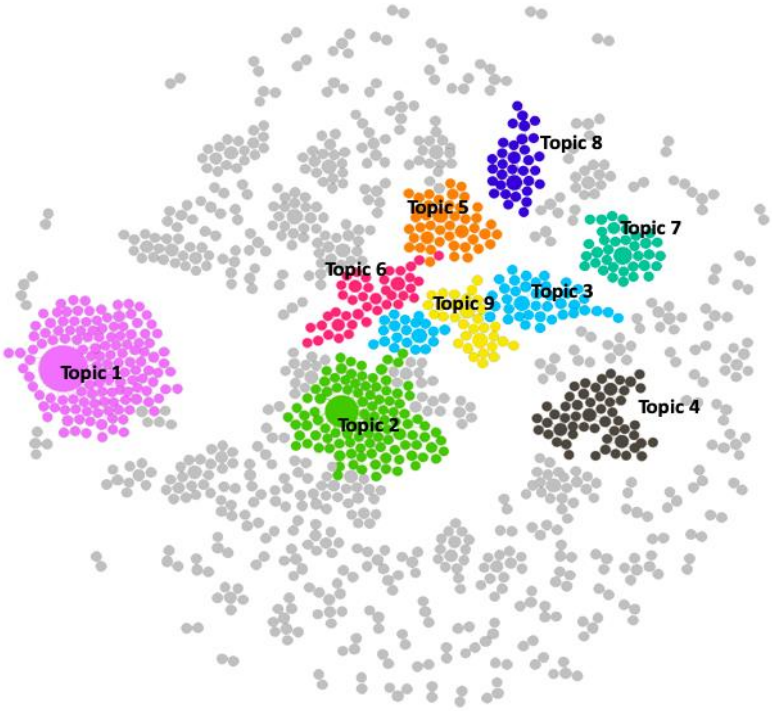

Figura 6.6: Red de co-ocurrencia de palabras clave:2000-2018.

Nº de tema	Temática	Top 10 de palabras clave de autor	% nodos	UMass
1	Project environment y integration	collaborative r y d, business model, system approach, project efficiency, project management information system, cultural gap, cost benefit analysis, knowledge integration, methodology, project-oriented company	10,9%	-0,6
2	Success factors	project success, program management, managing stakeholder, benefit management, management succession, stakeholder and organizational satisfaction, international development, framework, career path, project sponsors	8,5%	-1,2
3	Project strategy	portfolio management, project selection, project evaluation, project procurement, strategic orientation, project contract, project marketing,	4,8%	0,0

		net present value, customer integration, project portfolio		
4	Stakeholder management	project stakeholders, stakeholder analysis, development project, mega projects, stakeholder behavior, performance evaluation, resource planning, change management, organizational citizenship behavior, fuzzy analytic network process	4,4%	0,0
5	Project delivery	construction management, implementation, project delivery, supply chain management, competitive advantage, managing risk, construction risk, engineering construction, innovation, customer satisfaction	3,6%	-6,6
6	Risk management	risk assessment, system analysis and design, information system, linguistic variables, fuzzy number, management technique, building project, international construction, political risk, risk register	3,4%	-2,0
7	Human resource management	financial resources, relationship management, factor analysis, bid price, strategic information system, private finance initiative, new product introduction, withholding support, support profile, roles and responsibility	3%	0,0
8	Scheduling	project network, project scheduling, modelling, resource allocation, critical chain, large project, relational competence, multi project, prevention appraisal failure, resource scheduling	2,7%	-6,5
9	Governance y control	knowledge management, project performance, human resource management, individual control charts, knowledge co-production, knowledge factor, knowledge leadership, data complexity, competence management, complexity management	2,6%	0,0

Tabla 6.7: Temas principales extraídos de la red de co-ocurrencia totalmente unificada.

La figura 6.7a muestra la frecuencia con la que las temáticas de alta calidad identificadas se utilizaron en cada una de las publicaciones realizadas por año. Para representar la variación anual de cada temática se consideran los resultados como una fracción del número total de publicaciones por año. La figura 6.7b es una visualización del análisis de *burst* de los temas de 2000 a 2018, estableciendo el parámetro s (distancia multiplicativa entre estados) en 1,6, y el parámetro gamma

(dificultad asociada con el ascenso de un estado) en 0,1. Cada punto representado en la Figura 6.7b es un indicador de la presencia de un estado de ráfaga durante el período en el que se muestra (nótese que los colores que representan cada temática coinciden entre la Figura 6.7a y b). Vale la pena tener en cuenta que la detección de ráfagas muestra el cambio rápido de frecuencia, no la frecuencia total. Por lo tanto, un tema puede estallar en popularidad, aunque sigue siendo menos significativo que los temas de alta frecuencia que son constantes durante el período de tiempo.

La temática 1 ("*Project Environment and Integration"*) alcanzó un pico de popularidad o estado de *burst* en 2004, seguido de una tendencia media más moderada hasta el final del período. La temática 2 (*"Success Factors"*) tiene una tendencia paralela al tema 1 hasta 2007, luego se mantiene hasta 2012, cuando sufrió un aumento en su popularidad que aumentó ligeramente su tendencia promedio. La temática 4 (*"Satkeholder Management")* también mostró una tendencia ligeramente creciente con un incremento importante al final del período. El resto de las temáticas se han mantenido más o menos en el tiempo, con el pequeño pico ocasional de popularidad dentro del período.

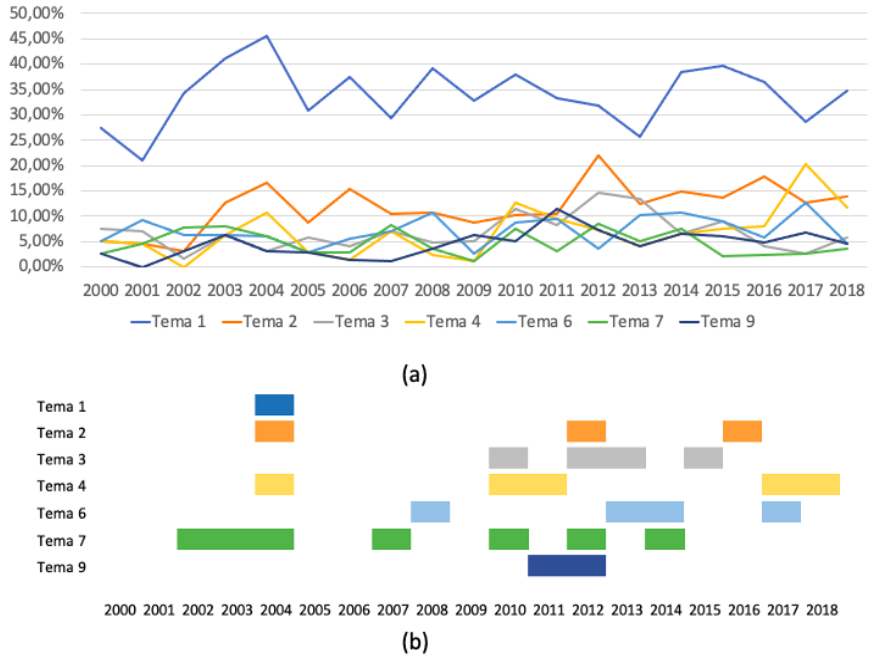

Figura 6.7: Evolución de las tendencias temáticas durante el período 2000-2018.

6.5 DISCUSIÓN

El análisis de co-ocurrencia es eficaz para representar el mapa conceptual de una disciplina (Callon et al., 1983). En este procedimiento, solo se seleccionan aquellas palabras clave cuyas puntuaciones basadas en la red son más altas para revelar los principales temas de investigación y sus relaciones al micro nivel (Chen y Xiao, 2016). Como esta técnica se basa en la frecuencia de palabras clave en lugar de la relevancia, un efecto indeseable es que muchas palabras clave de significado amplio (generales) emergen en la red. Estas palabras clave genéricas pueden ser útiles para representar una visión general aproximada de una disciplina científica, aunque tienen menos éxito en la identificación de temas detallados de un dominio de investigación (Chen y Xiao, 2016). Además, la heterogeneidad de las palabras clave más frecuentes puede sesgar las puntuaciones obtenidas en la red si las palabras clave no están unificadas antes de la

construcción de la red de co-ocurrencia (Muñoz-Écija et al., 2017). Para evitar este problema, en diferentes trabajos se han aplicado diferentes enfoques para pre-procesar la terminología, como los juicios de expertos (Zhang et al., 2016; Khassed et al., 2017; Leung et al., 2017), stemming/lematización (James et al., 2015; Van Eck et al., 2010; Rokaya et al., 2008; Hu et al., 2019; Kim et al., 2020), etiquetado en los textos (Van Eck et al., 2010, Kim et al., 2020) o *tesauros* (Ding et al., 2001; Muñoz-Écija et al., 2017). En este trabajo, se propuso un nuevo método para unificar palabras clave, basado en una combinación de técnicas léxicas, sintácticas y semánticas. En nuestro estudio experimental, se realizó un análisis comparativo entre los modelos logrados después de aplicar técnicas de pre-procesamiento, léxicas-sintácticas y semánticas. Nuestro estudio muestra que el modelo totalmente unificado funcionó mejor que los otros dos modelos. En el modelo basado únicamente en el *stemming*, un mayor número de palabras clave más generales (con significado más amplio) surgieron en la red de co-ocurrencia de palabras clave que en el modelo unificado, lo que llevó a la detección de temáticas de baja coherencia descritas por términos clave más generales. Aunque el análisis de tendencias de PM en (Pollack y Adler, 2015) se ha llevado a cabo en un período de tiempo anterior y ha utilizado diferentes fuentes de información que nuestro trabajo, también se puede señalar que, en el modelo resultante, las 20 palabras clave principales (clasificadas por frecuencia) incluyen términos amplios, como "*cost*", "*construction*" o "*marketing*". Estas palabras clave sólo surgieron en nuestro modelo pre-procesado, mientras que los términos más estrechos que cubren estas palabras clave más generales surgieron en el modelo unificado, como "*cost benefit analysis*", *"construction management*" o "*project marketing*". Por lo tanto, los resultados de nuestro enfoque muestran que la unificación de palabras clave reduce de forma considerable el número de palabras clave generales que surgen en la red de co-ocurrencia.

La primera evaluación de nuestra propuesta se centró en evaluar el rendimiento de nuestro método de unificación. Los resultados experimentales revelaron la eficacia de nuestra propuesta de unificación en términos de precisión y *recall*. Una de las principales consecuencias colaterales positivas de la aplicación de patrones léxicos

fue la identificación y unificación de antónimos, como el término "*monetary factor*" y el término "*non monetary factor*". Algunos autores (Ding et al., 2001) han hecho hincapié en que los antónimos deben ser unificados ya que se refieren al mismo término. Por otro lado, encontramos variantes que diferían del término base por un intercambio de palabras. Por ejemplo, los términos "*public-private partnership*" y "*private/public partnership*" son variantes de la misma palabra clave. Nuestras técnicas no fueron capaces de detectar este tipo de variantes. En el corpus de documentos, también detectamos variantes más estrechas de las siglas. Por ejemplo, "*PMBoK*" es un acrónimo del término base "*Project Management Body of Knowledge*" y "*PMBoK guide*" es una variante del acrónimo. Sin embargo, no es una variante del término base, por lo que nuestros patrones léxicos no pudieron identificarlo y unificarlo. Por último, hay algunos términos que no son variantes de un término base, aunque deben unirse, ya que conservan en gran medida el significado en el campo, como "*interpersonal relation*s" y las "*organizational relationships*". Nuestra técnica semántica funcionó mal en estos casos posiblemente ya que el corpus no era lo suficientemente grande como para entrenarlo adecuadamente.

El porcentaje de unificación mediante el uso de la técnica semántica fue relativamente bajo en comparación con las técnicas léxico-sintácticas. Esto se debe, en parte, al hecho de que los resultados no contaron solapamientos con las otras técnicas, que se aplicaron antes que la técnica semántica. Cuando se analizaron los resultados obtenidos, encontramos una clara diferencia entre las técnicas léxico-sintácticas y las técnicas semánticas. Las técnicas léxicas realizaron la unificación preservando el significado a nivel de palabra clave, mientras que las técnicas semánticas preservaron el significado a un nivel intermedio entre la palabra clave y el tema. Por ejemplo, las técnicas léxicas unificaron las palabras clave "*organizational learning*" y "*organisational learning*", teniendo ambas variantes el mismo significado. Sin embargo, después de aplicar las técnicas semánticas, las siguientes palabras clave se unificaron con las anteriores: "*training programs*", "*career model*" y "*career path*". Todas estas palabras clave conservan el significado a un nivel intermedio entre la palabra clave

base ("*organizational learning*") y el tema en el que se agrupan ("*Success factors*").

La segunda evaluación de nuestra propuesta se centró específicamente en la utilidad de nuestro método para el análisis de tendencias. Los resultados de nuestro enfoque muestran que nuestro proceso de unificación mejoró sustancialmente la coherencia de los temas identificados, en comparación con un modelo basado exclusivamente en la derivación de palabras (*stemming*). Los 7 temas de alta calidad del modelo unificado se podrían interpretar en un contexto significativo que comprendiera la mayoría de las palabras clave principales. Discutimos a continuación los resultados obtenidos comparándolos con otros trabajos relevantes en PM.

En (Padalkar y Gopinath, 2016), los autores analizaron manualmente las tendencias temáticas en PM durante un período de 16 años entre 2000 y 2015. Los autores seleccionaron 230 artículos de tres revistas internacionales utilizando una medida basada en el recuento de citas. En total, el IJPM (nuestra fuente de datos) proporcionó 145 (63%) artículos en un período cubierto completamente por nuestro análisis (nuestro trabajo incluyó 3 años más al final del período). Examinaron la progresión temática durante el período 2000-2015 y encontraron un conjunto de temas principales que cubren los 7 temas de alta calidad que surgieron en nuestro modelo unificado.

Como en (Padalkar y Gopinath, 2016), "*Success Factors*" es un tema principal de interés en la investigación, que abarca palabras clave que se expanden más allá del "*Iron Triangle*" (coste, tiempo y rendimiento). De acuerdo con (Ika, 2009), estas palabras clave enfatizan otros criterios de éxito, como el éxito del proyecto/producto, la gestión del programa, la satisfacción de los interesados y la organización, o los *frameworks*. Específicamente, la palabra clave "*framework*" resume la evidencia sobre el impacto de *frameworks* útiles sobre los factores de éxito o fracaso del proyecto (Padalkar y Gopinath, 2016).

En nuestros resultados, la temática de "*Risk management*" se aborda principalmente a través de elementos de riesgo del proyecto y "*fuzzy methods*" para medir o gestionar dicho riesgo, en consonancia con (Padalkar y Gopinath, 2016). Por otro lado, la aparición de palabras

clave como "*knowledge*" o "*complexity*" en "*Governance and control*" revela la importancia de aprovechar los conocimientos adquiridos para el liderazgo de proyectos complejos (Pitsis et al., 2014). También es destacable la ocurrencia de la palabra clave "*megaproject*" en "*Stakeholder management*", ya que destaca la importancia de la estrategia adoptada en este tipo de proyectos para gestionar los diferentes intereses de cada uno de los participantes del proyecto (Ninan et al., 2019). En "*Human resource management*", surgieron palabras clave que aún no estaban cubiertas en el PMBoK, como "*factor analysis*" o "*new product introduction*". Estas palabras clave se han vinculado a las prácticas de gestión de recursos humanos (HRMP) (Easa y Orra, 2020). El término "*factor analysis*" se refiere a los métodos estadísticos utilizados para validar los estudios sobre HRMP, mientras que "*new product introduction*" se refiere a la innovación de productos en HRMP.

La progresión temática descrita por (Padalkar y Gopinath, 2016) también es consistente con nuestros resultados, excepto para el tercer subperíodo (2011-2015) del tema 1 ("*Project environment & integration*") y el tema 4 ("*Stakeholder management*"), y el Progresión completa del tema 7 ("*Human resource management*"). En nuestro modelo, el tema 1 ("*Project environment & integration*") cubre aspectos de la integración del proyecto, que pueden no tratarse en el tema del entorno del proyecto en (Padalkar y Gopinath, 2016). En cuanto al tema 4 ("*Stakeholder management*"), este es un tema principal hasta 2010 en (Padalkar y Gopinath, 2016), en contraste con nuestros resultados, que lo ubicaron como un tema principal a lo largo del período 2000-2018 con una tendencia ligeramente creciente. Cabe destacar que la gestión de interesados o *stakeholders* apareció como un capítulo independiente en la 5º edición del PMBoK publicada en 2013, lo que justifica su incremento como tema recurrente en los últimos años de nuestro análisis, y el posterior estallido que identificamos en 2017-2018. Finalmente, nuestros resultados muestran que el tema 7 ("*Human resource management*") se mantiene continuo a lo largo del tiempo, con pequeños picos ocasionales de popularidad dentro del período, mientras que en (Padalkar y Gopinath, 2016) este tema es solo uno de los temas principales durante el período 2006-2010. A pesar de las diferencias,

nuestros resultados han demostrado que nuestro método puede ser un excelente complemento a otras metodologías bibliométricas para el estudio de tendencias temáticas en PM por varias razones: bajo coste, posibilidad de utilizar grandes cantidades de datos textuales, experimentos reproducibles y automatización completa del proceso.

6.5.1 Anotación semántica vs unificación de palabras clave de autor

En el anterior capítulo analizamos otra aproximación sobre la misma fuente de datos seleccionada en este estudio. En ella, planteamos la anotación semántica de resúmenes basada en un diccionario creado ad hoc para la extracción de tendencias temáticas. Las principales diferencias entre estas dos aproximaciones las comentamos a continuación. Primero, el diccionario elaborado no unifica terminología y consta únicamente de 874 términos frente a las 3.151 palabras clave de autor. Segundo, la red de co-ocurrencia resultante de la anotación semántica es menos modular que la red creada a partir de palabras clave, pero es más densa y está más interconectada. Tercero, las temáticas de alta calidad, en términos de coherencia, extraídas mediante el análisis de la red basada en anotaciones fueron cuatro, frente a las siete obtenidas en el presente estudio. Además, si nos centramos en los términos que componen cada temática podemos ver como en esta última aproximación y, con ayuda del proceso de unificación, se obtuvieron términos más específicos y concretos. De ahí que su valor de coherencia temática sea mucho mejor. A pesar de que únicamente se obtuvieron cuatro temáticas de buena calidad para el modelo basado en anotación semántica, todas ellas concuerdan con las temáticas extraídas en el modelo unificado de palabras clave. En el caso del tópico que hace referencia a la gestión de riesgos (tema 6 en el modelo unificado y tema 3 en el modelo anotado), ambas aproximaciones comparten los términos de "*risk assessment*" y "*risk register*", y comparten términos bastante relacionados como "*management technique*" (modelo unificado) y "*management science*" (modelo anotado). En el resto de tópicos sucede algo similar, están compuestos por términos léxicamente diferentes pero que hacen referencia a la misma área temática. Por ejemplo, para la temática identificada en los dos modelos

etiquetada como "*Project strategy*", el modelo unificado se centra en diferentes procesos del proyecto como "*project selection*", "*project evaluation*" o "*strategic orientation*", mientras que el modelo basado en diccionario se basa más en la metodología aplicada, con términos como "*impact analysis*", "*business process modeling*" y "*project logic*".

Otro aspecto a destacar en la anotación de resúmenes es la relevancia de términos interesantes como "six sigma", que apenas tiene relevancia en la red resultante del modelo unificado. Este aspecto puede indicarnos que únicamente con el análisis de palabras clave de autor no estamos extrayendo suficiente información como para diseñar el mapa conceptual del dominio analizado. Es decir, aunque las palabras clave de autor se consideren la unidad básica de resumen de la publicación, estaríamos perdiendo información ya que, muchas veces, es imposible resumir el contenido del artículo en tan pocas palabras.

6.6 CONCLUSIONES

En este trabajo, hemos abordado la cuestión de cómo las palabras clave de autor representadas en una red de co-ocurrencia pueden unificarse sin depender en gran medida del juicio de expertos en los diferentes dominios de aplicación. Nuestro principal logro es un método para la unificación automática de palabras clave en un corpus de documentos, que se ha evaluado aplicándolo al dominio PM. Los resultados experimentales de nuestro enfoque han demostrado que la unificación de palabras clave reduce el número de palabras clave más generales que suelen surgir en una red de co-ocurrencia. Nuestra propuesta abre la puerta a la viabilidad de desarrollar herramientas orientadas al usuario final que automaticen el proceso de unificación de palabras clave y al estudio de las tendencias temáticas de un área concreta de estudio.

CAPÍTULO 7
CONCLUSIONES Y TRABAJO FUTURO

Uno de los retos actuales de disponer de grandes cantidades de información es la búsqueda y extracción de la información relevante que se necesita, sin perder ningún dato importante y en el menor tiempo posible. En esta tesis se ha propuesto una herramienta basada en la anotación semántica como una ayuda automática al proceso de búsqueda y extracción de la información almacenada en las publicaciones científicas. La herramienta ha sido probada y validada con éxito en dos dominios de aplicación diferentes, el biomédico y la dirección de proyectos.

7.1 CONTRIBUCIONES Y HALLAZGOS EMPÍRICOS

En el presente trabajo se ha desarrollado una herramienta de anotación semántica para la extracción automática de la información relevante contenida en publicaciones científicas. Se ha aplicado sobre dos ámbitos de estudio: la biomedicina y la investigación en dirección de proyectos. Sobre este núcleo principal, se han llevado a cabo varios experimentos para aumentar el conocimiento empírico de cada una de las áreas y mejorar los resultados de anotación. Es por ello que se enumeran a continuación las principales aportaciones de nuestro trabajo de investigación, centrado en el diseño e implementación de procedimientos automáticos:

1. **Creación de una herramienta de anotación semántica de documentos**. Las principales contribuciones de esta herramienta han sido 1) la identificación de los fragmentos de código relevantes para la anotación mediante el diseño de cinco patrones lingüísticos a partir de las estructuras base reconocidas en los *abstracts*, y 2) la anotación de los textos basada en un

algoritmo que permite comparar éstos contra los términos de la ontología utilizada. El proceso de anotación compara directamente los términos de la ontología con los textos a anotar. Esto permite mejorar la velocidad de anotación, sin perder eficiencia, en comparación con las herramientas basadas en el procesamiento del lenguaje natural. Para aumentar la eficiencia de la anotación, nuestro anotador 1) construye offline un diccionario, que incluye todos los conceptos y sinónimos de la ontología pre-procesados, 2) utiliza una ventana de tamaño ajustable que se va deslizando por el texto para extraer las palabras del texto a comparar, y 3) permuta las secuencias de palabras extraídas para incrementar la capacidad de anotación. Además, nuestro algoritmo utiliza las relaciones jerárquicas, definidas en la ontología, para filtrar anotaciones redundantes, manteniendo únicamente los conceptos más específicos. En el ámbito biomédico, nuestros resultados confirman que es posible, sobre las publicaciones científicas, identificar automáticamente aquellas que corresponden a informes clínicos, con una alta precisión y anotarlas con una calidad satisfactoria (*F-measure* del 74%). La anotación semántica de todas las anomalías fenotípicas encontradas en una enfermedad puede facilitar el diagnóstico temprano de pacientes con enfermedades raras, al proporcionar los fenotipos relacionados junto con su frecuencia. En el ámbito de la investigación en dirección de proyectos (PM) nuestra herramienta también nos ha permitido extraer las principales tendencias temáticas (corroboradas por otros estudios en el área), utilizando glosarios terminológicos como base. De esta forma, podemos afirmar que la herramienta de anotación semántica diseñada puede ser usada tanto sobre ontologías como sobre diccionarios de términos, al haber demostrado su eficacia sobre dos entornos totalmente diferentes.

2. **Diseño de un método automático para la generación de variantes léxicas (sinónimos), a partir de la terminología propia de una ontología**. Nuestro enfoque se basa tanto en las

propiedades léxicas de los términos como en la estructura jerárquica de la ontología. Al identificar las diferencias léxicas entre un término y sus términos descendentes, el método aprende nuevos términos y modificadores, que permiten generar sinónimos para los términos descendentes. La generación sintética de sinónimos puede dar lugar a sinónimos sin sentido. El método filtra estos candidatos sin sentido, simplemente buscando las frases exactas de los candidatos en MEDLINE y descartando aquellos para los que no hay resultados (lo que indica que no son términos utilizados por la comunidad científica). Nuestros resultados mostraron que los sinónimos obtenidos generaban un impacto positivo en el reconocimiento de conceptos, principalmente aquellos sinónimos que correspondían a términos HPO que tenían menos probabilidades de aparecer en los textos, por lo que eran mucho más informativos.

3. **Diseño e implementación de un método basado en anotación semántica para semi-automatizar el análisis de las tendencias temáticas en PM a partir de artículos de investigación**. Nuestra metodología se basa en la combinación de varios procesos. Por un lado, la extracción de conocimiento mediante la anotación basada en glosarios terminológicos. Y, por otro lado, la construcción de la red de co-ocurrencia de los términos anotados. Sobre dicha red se aplica el análisis de clústeres para obtener las principales agrupaciones de temas o subredes. Posteriormente, sobre cada agrupación se analiza, por una parte, la coherencia temática para determinar su calidad y, por otra, su tendencia en el periodo de estudio a través del análisis de *burst*. Nuestro principal logro es corroborar que nuestro enfoque es capaz de extraer información relevante y de interés utilizando recursos terminológicos diferentes (en este caso, un diccionario en vez del uso de una ontología), en un contexto totalmente diferente, y que nuestros resultados están avalados por otros estudios en el área.

4. **Diseño e implementación de un método de unificación automático de palabras clave de autor, basado en técnicas léxicas, sintácticas y semánticas.** Nuestro enfoque ha sido probado en el contexto de la realización de un análisis de tendencias de la investigación de gestión de proyectos. Nuestros resultados han revelado la efectividad de nuestra propuesta, en términos de precisión y exhaustividad *(recall)*, así como de la coherencia de las tendencias temáticas obtenidas. Además, nuestro método de unificación reduce la cantidad de palabras clave de significado amplio que surgen en una red de co-ocurrencia de términos, y simplifica la terminología dispar de un corpus de documentos, aumentando, así, la fiabilidad y validez de los análisis de tendencias basados en palabras clave de autor.

7.2 CONCLUSIONES

Nuestro trabajo de investigación aporta nuevos métodos, técnicas y herramientas para la extracción y análisis del contenido de publicaciones científicas, que han sido validados sobre dos dominios completamente diferentes: el biomédico y el de gestión de proyectos. A continuación, se enumeran algunas lecciones y conclusiones extraídas durante el transcurso de la tesis:

- Nuestro anotador semántico es capaz de extraer información relevante sobre diferentes ámbitos de estudio. En las dos áreas de trabajo ha mostrado buenos resultados en términos de precisión y exhaustividad (*recall).* También ha obtenido temáticas coherentes y concordantes con estudios previos en el campo de la investigación en dirección de proyectos. Nuestros resultados muestran que nuestra herramienta se podría trasladar a cualquier otro entorno, siempre y cuando, se actualice con el recurso terminológico correspondiente.

- Nuestro estudio experimental sobre anotación de casos clínicos de pacientes publicados en revistas científicas mostró que la herramienta de anotación obtuvo mejores

resultados en términos de precisión y *recall* que los anotadores existentes en el momento, como el anotador del NCBO o el facilitado por GoPubMed. Posteriormente, surgió el anotador Bio-Lark CR, que en su evaluación, analizó el rendimiento de nuestro anotador. En esta comparativa realizada por otros investigadores también se obtuvieron buenos resultados, ya que la precisión obtenida en este caso fue superior al Bio-Lark CR. El *recall* fue ligeramente inferior, pero hay que tener en cuenta que en el momento de la evaluación, estos evaluadores utilizaron nuestro anotador con una versión antigua de la ontología HPO, por lo que los resultados obtenidos no eran realmente comparables.

- Nuestro método automático de generación de variantes léxicas (sinónimos) produjo un impacto muy relevante en el reconocimiento automático de conceptos en artículos de investigación. Utilizando las propiedades léxicas y lógicas de la ontología que sirvió de caso de uso (HPO), nuestro método generó 745 nuevos sinónimos para la ontología, que cubrieron 488 conceptos. La evaluación realizada de estos términos identificó una mejora en el desempeño de *F-measure en* las tareas de extracción de información. Los nuevos términos permitieron la recuperación de un 6% más del total de artículos de investigación sobre enfermedades hereditarias, y un 33% cuando consideramos únicamente los conceptos altamente informativos (es decir, aquellos cuyo valor del contenido de información de Resnik es más elevado), lo que constata la efectividad del método.

- La anotación semántica basada en un diccionario terminológico sobre el área, aunque sea limitado, arroja buenos resultados en la identificación de tendencias. El diccionario fue construido a partir de los glosarios terminológicos validados por instituciones reconocidas en el área, por lo que engloba la terminología relevante

utilizada y permite extraer información relevante de los resúmenes de los artículos utilizados en el análisis. Nuestros resultados confirman que el diccionario funciona adecuadamente para extraer las principales temáticas en el área, ya que pudimos extraer automáticamente cuatro temáticas de alta calidad, que coinciden con tendencias identificadas por estudios previos realizados en el área.

- La combinación de técnicas léxicas, sintácticas y semánticas sobre palabras clave de autor proporciona un método automático y eficiente de unificación de terminología. Nuestro enfoque permitió analizar de forma más eficiente la red de co-ocurrencia en el ámbito de la investigación en PM por dos razones. Primero, se redujo el conjunto de términos representados en la red y, segundo, se eliminaron términos con significado demasiado amplio que lo único que proporcionaban eran más interconexiones en la red que no ayudaban a su clarificación. Además, permitió obtener siete temáticas de alta calidad, en términos de coherencia temática, frente al modelo no unificado, que no proporcionó ninguna. De esta forma, comparando el modelo unificado frente al modelo sin unificar pudimos comprobar la validez de nuestra metodología de unificación.

7.3 LIMITACIONES Y TRABAJO FUTURO

Como hemos podido comprobar, nuestra herramienta de anotación semántica ha proporcionado buenos resultados en los dos entornos de estudio: el biomédico y el de investigación en dirección de proyectos. Pero pese a los resultados obtenidos, nuestra herramienta comparte las mismas limitaciones que el resto de anotadores semánticos. Primero, no disponemos del contexto suficiente para interpretar el concepto o entidad anotada con la herramienta. A pesar de que nuestra herramienta demostró mejorar la precisión y el *recall* de otros anotadores ya existentes en el área de la biomedicina, queda un largo camino por recorrer. Planteamos mejorar este aspecto aumentando la capacidad de

anotación a través de la mejora de los recursos terminológicos disponibles o la introducción de técnicas de aprendizaje capaces de incorporar la semántica del contexto en nuestros análisis añadiendo, por ejemplo, términos referidos a nuestras anotaciones. Y, por último, en el proceso de anotación semántica no se tiene en cuenta las diferencias entre la literatura biomédica descrita en las ontologías y los textos clínicos. Las ontologías no recogen toda la jerga biomédica por lo que sería importante poder reconocer todo el vocabulario utilizado en los textos clínicos sobre pacientes. En este punto, hemos tratado de mejorar este aspecto con el diseño de nuestro método de identificación de sinónimos a partir de la ontología HPO. Aunque esta propuesta está muy ligada al recurso ontológico utilizado, proporciona un buen punto de partida para resolver dicha limitación.

Para mejorar la capacidad de anotación, una alternativa consiste en ampliar la terminología de los recursos de los que disponemos en cada uno de los ámbitos de análisis. Como ya hemos indicado, nuestro enfoque para la identificación de sinónimos es un método útil y eficaz que incrementa la capacidad de anotación de nuestra herramienta para el campo biomédico. Pese a los buenos resultados obtenidos en este caso, sólo pudimos quedarnos con un 8% de los términos nuevos después de su evaluación. Por esta razón, proponemos extender nuestro procedimiento de sustitución de sinónimos mediante la identificación de regularidades y superposiciones léxicas. Podríamos, incluso, extender automáticamente nuestra propuesta de acuerdo con un principio similar al definido en la herramienta *IntelliGO*[9], pero limitado a los conceptos de HPO. De esta forma, buscaríamos todas las palabras cercanas a un concepto (excluyendo las "*stopwords*") y podríamos identificar el nuevo modificador del término. La metodología propuesta podría adaptarse para seleccionar automáticamente los sinónimos más apropiados para las tareas de reconocimiento de conceptos y, de esta forma, enriquecer la ontología HPO que utilizamos en nuestro anotador.

[9] http://bioinfo.loria.fr/Members/benabdsi/intelligo_project/

Por otro lado, hemos visto cómo nuestra herramienta de anotación permitió extraer el conocimiento de un ámbito completamente diferente como es el de la investigación en dirección de proyectos, usando un diccionario terminológico. Pero creemos que este recurso no es suficiente para extraer todo el conocimiento que se alberga en los resúmenes de las publicaciones. Dado que el análisis de palabras clave de autor es otra herramienta muy útil para extraer conocimiento de las publicaciones, como hemos podido constatar, y dado que la metodología propuesta para la unificación de terminología proporciona una mejora sustancial frente a los métodos clásicos de pre-procesamiento (stemming, lematización, ley de Zipf, etc.), se propone el estudio de un modelo híbrido. De esta forma tendríamos, por un lado, el conocimiento recogido por glosarios elaborados por instituciones en el área y, por otro, la terminología aportada por los autores en el campo de investigación. Sobre este conjunto de datos se aplicaría nuestro método de unificación para conciliar las principales diferencias terminológicas, añadir las relaciones semánticas y reducir, así, la heterogeneidad ya mostrada por las palabras clave de autor. Creemos que la combinación de estas dos fuentes de información dotará a nuestra herramienta de anotación la capacidad necesaria para extraer todo el conocimiento relevante de las publicaciones y mejorar la calidad de los resultados obtenidos (es decir, extraer mayor cantidad de temáticas coherentes y específicas al área de análisis).

APÉNDICE A
PATRONES LINGÜÍSTICOS PARA EXTRAER FRAGMENTOS RELEVANTES DE *ABSTRACTS*

La tabla de a continuación muestra los patrones lingüísticos confeccionados para la identificación de los fragmentos relevantes del conjunto de 515 *abstracts* o resúmenes de informes clínicos de pacientes de CTX obtenidos de PubMed.

Patrón léxico	Ejemplos
we <action1> …<patient>	We report a 25-year-old young man presenting … (PMID: 2303834) Here we present such a case which … (PMID: 20329433) We studied cholesterol and phytosterol profiles in two siblings with CTX … (PMID: 18949577)
<paper> … <action2>…<patient>	This report concerns two new mutations in the sterol 27-hydroxylase gene in two patients with … (PMID: 8931710) The present paper describes two cases, with (PMID: 2114502)
<author>… <action3>…<patient>	The authors describe four patients with cerebrotendinous xantomatosis… (PMID: 3344851) Here, the authors present a case of a drug-resistant epilepsy patient with … (PMID: 22197981)

	The authors report the case of a 22-year old man presenting with cerebrotendinous … (PMID: 1649488)
<patient> <action4>	A 26-year-old female developed mental deterioration, … (PMID: 1934787) A 36-year-old female with typical CTX clinical manifestation had Spindle shaped ... (PMID: 22018287) A 63-years-old woman noticed unsteady gait … (PMID: 10885331)
<patient> <be><action5>	Four cases without xanthomas among the presenting symptoms are described … (PMID: 1320501) The present study, therefore, was carried out to examine the metabolism of LDL in a 58-year-old black man with CTX… (PMID: 3821507)

APÉNDICE B
SINÓNIMOS ADICIONALES GENERADOS POR NUESTRO MÉTODO

En la siguiente tabla se muestra la lista completa de los 20 sinónimos adicionales generados por nuestro método para la versión de la ontología de HPO del 13 de abril de 2017.

Nuevo sinónimos	ID HPO	Término preferido en HPO
acute fatty liver	006573	acute hepatic steatosis
acute liver inflammation	200119	acute hepatitis
atypical pulmonary carcinoid	030446	atypical pulmonary carcinoid tumor
bilateral facial muscle weakness	001349	bilateral facial weakness
bilateral nanophthalmos	007633	bilateral microphthalmos
bladder cancer	009725	bladder neoplasm
calcium oxalate kidney stones	008672	calcium oxalate nephrolithiasis
chronic liver failure	100626	chronic hepatic failure
chronic liver inflammation	200123	chronic hepatitis
epidermal nevi	010816	epidermal nevus
exercise-induced lactic acidosis	004901	exercise-induced lactic acidemia
generalized muscle wasting	003700	generalized amyotrophy
infantile hypotonia	008947	infantile muscular hypotonia
neoplasm of the nasopharynx	100630	neoplasia of the nasopharynx
segmental demyelination	007107	segmental peripheral demyelination
severe deafness	012714	severe hearing impairment
severe hearing loss	012714	severe hearing impairment

severe sensorineural deafness	008625	severe sensorineural hearing loss
severe sun sensitivity	007537	severe photosensitivity
unilateral facial weakness	012799	unilateral facial palsy

APÉNDICE C
VALORES DE COHERENCIA TEMÁTICA

La siguiente tabla muestra el valor de UMass para cada temática identificada en el modelo pre-procesado, modelo léxico y sintácticamente unificado y el modelo totalmente unificado, clasificado en temas de alta y baja calidad en función del valor UMass de coherencia obtenido.

	TEMÁTICAS DE ALTA CALIDAD	UMASS	TEMÁTICAS DE BAJA CALIDAD	UMASS
MODELO PRE-PROCESADO			Tema 1	-5,9
			Tema 2	-4,3
			Tema 3	-6,9
			Tema 4	-2,6
			Tema 5	-7,3
			Tema 6	-4,0
			Tema 7	-3,2
			Tema 8	-11,5
MODELO UNIFICADO LÉXICO-SINTÁCTICO	Tema 1	-0,93	Tema 2	-5,3
	Tema 3	0,0	Tema 4	-3,4
	Tema 5	0,0	Tema 6	-8,2
	Tema 7	0,0		
	Tema 8	-1,2		
	Tema 9	-0,4		
MODELO UNIFICADO	Tema 1	-0,6	Tema 5	-6,6
	Tema 2	-1,2	Tema 8	-6,5
	Tema 3	0,0		
	Tema 4	0,0		
	Tema 6	-2,0		
	Tema 7	0,0		
	Tema 9	0,0		

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

APÉNDICE D
TOP 10 DE PALABRAS CLAVE DE AUTOR PARA CADA TEMÁTICA

La siguiente tabla muestra las 10 palabras clave principales, ordenadas por grado ponderado en el análisis de co-ocurrencia para el modelo pre-procesado, modelo léxico y sintácticamente unificado y modelo totalmente unificado.

Nº	Modelo pre-procesado	Modelo léxico y sintácticamente unificado	Modelo totalmente unificado
1	project management doctrinal supremacy the project office collaborative r&d national culture psychological climate stakeholder participation higher education overdesign project work	collaborative r y d project efficiency global virtual team operations management cost benefit analysis managing stakeholder project complexity radical innovation business value maturity model	collaborative r y d bussiness model system approach project efficiency project management information system cultural gap cost benefit analysis knowledge integration methodology project-oriented company
2	project risk management best practice change project scheduling risk theory project control earned value management cost culture	agile project management construction project construction industry management practice PMBoK delay claim contractor selection delay analysis construction planning	project success program management managing stakeholder benefit management management succession stakeholder and organisational satisfaction international development framework organizational learning project sponsors
3	project success socialization management succession services	evaluation and monitoring adaptive programme management management education	portfolio management project selection project evaluation project procurement strategic orientation project contract

	project success criteria project strategy early planning entrepreneurship sustainability contractor satisfaction	project education project simulation building project career development cash in/out cost and schedule cost variance	project marketing net present value customer integration project portfolio
4	systems lifecycle public project infraestructure project resource planning routinization dynamic capabilities social risk project capabilities process social network analysis	project stakeholder mega project project governance integration management IT-enabled business project benefit management benefit realization co creation management turnover project management performance	project stakeholder stakeholder analysis development project mega projects stakeholder behaviour performance evaluation resource planning change management organizational citizenship behavior fuzzy analytic network process
5	stakeholders fuzzy logic public private partnerships risk allocation evaluation critical factors international development csfs concession period independent power producer	development project Stakeholder analysis international project public project information system project performance management resource planning dynamic feasinility information technology mental model	construction management implementation project delivery supply chain management competitive advantage managing risk construction risk engineering construction innovation customer satisfaction
6	project management office construction industry knowledge sharing portfolio performance absorptive capacity design choices governance innovation knowledge management organizational design	progress earned schedule cash flow fuzzy logic earned value earned value management earned duration fuzzy project scheduling construction duration genetic algorithm	risk assessment system analysis and design information system linguistic variables fuzzy number management technique building project international construction political risk risk register

7	portfolio management portfolio value management ideation portfolio management agile project management project complexity project modularity sensemaking front end networking	project leadership risk management virtual project transformational leadership horizontal leadership virtual team subway project engaged leaders sensitive material social isolation	financial resources relationship management factor analysis bid price strategic information system private finance initiative new product introduction withholding support support profile roles and responsibility
8	team human resource management partnership construction pressure alliance collaboration project performance project management maturity model critical theory	change management program management best practice communication management business change change program design change process call centre business objective	project network project scheduling modelling resource allocation critical chain large project relational competence multi project prevention appraisal failure resource scheduling
9		project selection portfolio management customer integration project portfolio goal programming organizational learning strategic management theory emergent strategy protestant work ethic doctrinal supremacy	knowledge management project performance human resource management individual control charts knowledge co-production knowledge factor knowledge leadership data complexity competence management complexity management

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

BIBLIOGRAFÍA

[1] Ahlemann, F., El Arbi, F., Kaiser, M.G., Heck, A., (2013). A process framework for theoretically grounded prescriptive research in the project management field. International Journal of Project Management, 31:43-56.

[2] Allones, J.L., Martínez, D., y Taboada, M. (2014). Automated mapping of clinical terms into SNOMED-CT: An application to codify procedures in pathology. J Med Syst, 38(10):134.

[3] Atkinson, J., y Bull, V., (2012). A multi-strategy approach to biological named entity recognition. Expert Systems with Applications, 39, 12968-12974. https://doi.org/10.1016/j.eswa.2012.05.033

[4] Aronson, A.R., (2006). MetaMap: Mapping Text to the UMLS Metathesaurus.

[5] Artto, K., Martinsuo, M., Gemünden, H. G., y Murtoaro, J., (2007). Foundations of program management: A bibliometric view. International Journal of Project Management, 27, 1-18. https://doi.org/10.1016/j.ijproman.2007.10.007

[6] Balakrishnan, R., Harris, M.A., y Huntley,R. (2013). A guide to best practices for Gene Ontology (GO) manual annotation. Database, 2013, bau054.

[7] Benabderrahmane, S., Smail-Tabbone, M., y Poch, O. (2010). IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics,11, 588.

[8] Bettembourg, C., Diot, C., Burgun,A., y Dameron, O. (2012). GO2PUB: querying PubMed with semantic expansion of gene ontology terms. J. Biomed. Semantics, 3, 7.

[9] Betts, M., y Lansley, P., (1995). International Journal of Project Management: a review of the first ten years. International Journal of Project Management 13: 2017-217.

[10] Bodenreider, O. (2004). The Unified Medical Language System UMLS: integrating biomedical terminology. Nucleid Acids Research, 32:267-270.

[11] Bodenreider, O., Rindflesch, T.C., y Burgun, A., (2002). Unsupervised, corpus-based method for extending a biomedical terminology. Workshop on Natural Language Processing in the Biomedical Domain (ACL) Proc; Philadelphia, PA: Association for Computational Linguistics; p. 53-60.

[12] Börner, K., (2010). Atlas of Science: Visualizing what we know. MIT press, London.

[13] Buza, T.J. McCarthy, F.M., y Wang, N., (2008) Gene Ontology annotation quality analysis in model eukaryotes. Nucleic Acids Res. 36, e12.

[14] Cambrosio, A., Limoges, C., Courtial, J. P., y Laville, F., (1993). Historical scientometrics Mapping over 70 years of biological safety research with co-word analysis. Scientometrics, 27, 119-143. https://doi.org/10.1007/bf02016546

[15] Carden, L., y Egan, L. (2008). The search for quality publications relevant to nontraditional industries. Project Management Journal, 39, 6-27. https://doi.org/10.1002/pmj.20068

[16] Chen, X., Chen, J., Wu, D., Xie, Y., y Li, J., (2016). Mapping the Research Trends by Co-word Analysis Based on Keywords from Funded Project. Procedia Computer Science, 91, 547-555. https://doi.org/10.1016/j.procs.2016.07.140

[17] Chen, G., y Xiao, L., (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. Journal of Informetrics, 10, 212-223. https://doi.org/10.1016/j.joi.2016.01.006

[18] Cho, J., (2014). Intellectual structure of the institutional repository field: A co-word analysis. Journal of Information Science, 40, 386-397. https://doi.org/10.1177/0165551514524686

[19] Cohen, A.M., Hersh, W.R., (2005). A survey of current work in biomedical text mining. Brief Bioinform; 6(1): 57-71.

[20] Collier, N., Groza, T., Smedley, D., Robinson, P.N., Oellrich, A., y Rebholz-Schuhmann, D., (2015). PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. Database (Oxford); bav104.

[21] Crawford, L., Pollack, J., y England, D., (2006). Uncovering the trends in project management: Journal emphases over the last 10 years. International Journal of Project Management, 24, 175-184. https://doi.org/10.1016/j.ijproman.2005.10.005

[22] Culnan, M., (1986). The intellectual development of management information systems, 1972-1982: a co-citation analysis. Management Science, 32, 156-172. https://doi.org/10.1287/mnsc.32.2.156

[23] Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. (2013) Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput. Biol., 9, e1002854. http://gate.ac.uk/.

[24] Deng, Z., Ma, F., Lan, R., Huang, W., y Luo, X., (2020). A Two-stage Chinese text summarization algorithm using keyword information and adversarial learning. Neurocomputing. https://doi.org/10.1016/j.neucom.2020.02.102

[25] Dhombres, F., y Bodenreider, O., (2016). Interoperability between phenotypes in research and healthcare terminologies-Investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics;7:3.

[26] Dill, S., Eiron, N., Gibson, D., y Gruhl, D., (2003). A case for automated large-scale semantic annotation. J. Web Semantics, 1, 115–132.

[27] Ding, Y., Chowdhury, G., y Foo, S., (2001). Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing and Management, 37, 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0

[28] Dolan, M.E., Ni, L., Camon, E. y Blake, J.A., (2005). A procedure for assessing GO annotation consistency. Bioinformatics 21 (Suppl 1), i136–143.

[29] Doms, A., y Schroeder, M., (2005). GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res., 33, W783–W786.

[30] Easa, N.F., y Orra, H.E., (2020). HRM practices and innovation: an empirical systematic review. International Journal of Disruptive Innovation in Government, Vol. 1 No. 1, pp. 15-35. https://doi.org/10.1108/IJDIG-11-2019-0005

[31] Fattori, M., Pedrazzi, G. y Turra, R., (2003). Text mining applied to patent mapping: a practical business case. World Patent Information, 25, 335-342, https://doi.org/10.1016/S0172-2190(03)00113-3.

[32] Federico, A., Dotti, M.T., Gallus, G.N. (1993). Cerebrotendinous xanthomatosis. In: Pagon R.A., Adam M.P., Bird T.D., et al., (eds). GeneReviewsTM [Internet]. University of Washington, Seattle, WA, pp. 1993–2014. http://www.ncbi.nlm.nih.gov/books/NBK1409/

[33] Filicetti, J. 2007. PMO and Project Management Dictionary. The Project Management Hut. Disponible en internet, URL: http://www.pmhut.com/pmo-and-project-management-dictionary.

[34] Floricel, S., Bonneau, C., Aubry, M., Sergi, V. 2014. Extending project management research: Insights from social theories. International Journal of Project Management,In Press.

[35] Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K., Hunter, L., y Verspoor, K., (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinform;15(1):59.

[36] Funk, C.S., Cohen, K.B., Hunter, L.E., y Verspoor, K.M., (2016). Gene Ontology synonym generation rules lead to increased performance in biomedical concept recognition. J Biomed Semantics. 2016; 7:52.

[37] Gemünden, H.G., y Schoper, Y., (2015). Future trends in Project management. Research Gate: https://www.researchgate.net/publication/303375998_Future_Trends_in_Project_Management

[38] Gephi, 2016. The Open Graph Viz Platform. https://gephi.org/ (University of Technology of Compiègne, UTC, France)

[39] Gkoutos, G.V., (2009). Entity/Quality-based logical definitions for the human skeletal phenome using PATO. Proc 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Minneapolis, MN. 2009; p. 7069-72.

[40] Groza, T., Hunter, J., Zankl, A., (2013). Mining Skeletal Phenotype Descriptions from Scientific Literature. PLoS ONE;8(2):e55656.

[41] Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F.M., Baynam, G., Zankl, A., y Robinson, P.N., (2015). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. Database(Oxford): bav005.

[42] Gruber, T. R., (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5, 199-220. https://doi.org/10.1006/knac.1993.1008.

[43] Guerreiro, J. y Rita, P., (2020). How to predict explicit recommendations in online reviews using text mining and sentiment analysis. Journal of Hospitality and Tourism Management, 43, 269-272. https://doi.org/10.1016/j.jhtm.2019.07.001.

[44] Hamon, T., Grabar, N., (2008). Acquisition of elementary synonym relations from biological structured terminology. In: Computational Linguistics and Intelligent Text Processing. Springer, p. 40-51.

[45] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A., (2005). Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucleic Acids Res; 33: D514–D517.

[46] Harris, M.A. (2004) Consortium GO: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res., 1, D258–D261.

[47] Henry, S., Cuffy, C., McInnes, B.T., (2018). Vector representations of multi-word terms for semantic relatedness. Journal of Biomedical Informatics, 77, 111-119. https://doi.org/10.1016/j.jbi.2017.12.006.

[48] Hettne, K.M., van Mulligen, E.M., Schuemie, M.J., Schijvenaars, B.J., Kors, J.A., (2010). Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics 1: 5.

[49] Hofmann, T., (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42, 177–196. https://doi.org/10.1023/A:1007617005950

[50] Hole, W.T., y Srinivasan, S., (2000). Discovering missed synonymy in a large concept-oriented Metathesaurus. Proc AMIA Symp; p. 354-358.

[51] Holton, C., (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. Decision Support Systems, 46, 853-864. https://doi.org/10.1016/j.dss.2008.11.013.

[52] Hu, K., Luo, Q., Qi, K. L., Yang, S. L., Mao, J., Fu, y X. K., (2019). Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. Information Processing and Management, 56(4), 1185–1203. https://doi.org/10.1016/j.ipm.2019.02.014

[53] Huang, K.C., Geller, J., Halper, M., y Cimino, J.J., (2007). Piecewise synonyms for enhanced UMLS source terminology integration. AMIA Annu Symp Proc; Chicago, IL: American Medical Informatics Association; p. 339-343.

[54] Huang, K.C., Geller, J., Halper, M., Perl, Y., y Xu, J., (2009). Using WordNet synonym substitution to enhance UMLS source integration. Artif Intell Med; 46(2):97-109.

[55] Ika, L.A., (2009). Project Success as a Topic in Project Management Journals. Project Management Journal, 40, 6-19. https://doi.org/10.1002/pmj.20137

[56] Jacomy, M, Venturini, T, Heymann, S, Bastian, M., (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. Plos One 9: e98679. http://dx.doi.org/10.1371/journal.pone.0098679

[57] James, T., Cook, D., Conlon, S., Keeling, K., Collignon, S., y White, T., (2015). A framework to explore innovation at SAP through bibliometric analysis of patent applications. Expert Systems with Applications, 42, 9389-9401. https://doi.org/10.1016/j.eswa.2015.08.007

[58] Jovanovic, J., y Bagheri, E., (2017). Semantic annotation in biomedicine: The current landscape. Journal of Biomedical Semantics. 8. 10.1186/s13326-017-0153-x.

[59] Jonquet, C., Nigam, H. S., y Mark, A. M., (2009): The open biomedical annotator. Summit Trans Bioinformatics, p 56-60.

[60] Khan, Z. y Qamar, U., (2016). Text Mining Approach to Detect Spam in Emails. Conference: International Conference on Innovations in Intelligent Systems and Computing Technologies (ICIISCT2016).

[61] Khasseh, A.A., Soheili, F., Moghaddam, H.S., y Chelak, A.M., (2017). Intellectual structure of knowledge in iMetrics: A co-word analysis. Information Processing and Management, 53, 705-720. https://doi.org/10.1016/j.ipm.2017.02.001

[62] Kim, S., Park, H., y Lee, J., (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. Expert Systems with Applications, 152, 113401. https://doi.org/10.1016/j.eswa.2020.113401

[63] Kiryakov, A., Popov, B., y Terziev, I., (2005). Semantic annotation, indexing, and retrieval. J. Web Semantics, 2, 49–79.

[64] Kiyavitskaya, N., Zeni, N., y Cordy, J.R., (2009). Cerno: light-weight tool support for semantic annotation of textual documents. Data Knowl. Eng., 68, 1470–1492.

[65] Kleinberg, J., (2003). Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7, 373–397. https://doi.org/10.1023/A:1024940629314

[66] Kocbek, S., y Groza, T., (2016). Building a dictionary of lexical variants for human phenotype descriptors. Proc 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany; pp. 186-190.

[67] Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data, Nucleic Acids Res. 2014; 4:D966-D974.

[68] Kwak, Y. H., y Anbari, F. T., (2009). Analyzing project management research: Perspectives from top management journals. International Journal of Project Management, 27, 435-446. https://doi.org/10.1016/j.ijproman.2008.08.004

[69] Leung, X., Sun, J., y Bai, B. (2017). Bibliometrics of social media research: A co-citation and co-word analysis. International Journal of Hospitality Management, 66, 35-45. https://doi.org/10.1016/j.ijhm.2017.06.012

[70] Li, A., Zang, Q., Sun, D., y Wang, M., (2016). A text feature-based approach for literature mining of lncRNA-protein interactions. Neurocomputing; 206: 73-80.

[71] Li, Y., Wu, H., (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia, 25, 1104-1109. https://doi.org/10.1016/j.phpro.2012.03.206.

[72] Lu, Z., Kim, W., y Wilbur, W.J., (2009). Evaluation of query expansion using MeSH in PubMed. Inf. Retr. Boston, 12, 69–80.

[73] Ma, L., y Zhang, Y., (2015). Using Word2Vec to process big text data. https://ieeexplore.ieee.org/document/7364114

[74] Mao, N., Wang, M. y Ho, Y. (2010) A Bibliometric Study of the Trend in Articles Related to Risk Assessment. Human and Ecological Risk Assessment: An International Journal, 16, 801-824. https://doi.org/10.1080/10807039.2010.501248

[75] McGuinness, D.L. y F. van Harmelen, (2004). OWL Web Ontology Language. Disponible en: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[76] McKenzie, B.C., (1996). Medicine and the Internet: Introducing Online Resources and Terminology, Oxford University Press https://www.ncbi.nlm.nih.gov/PubMed/

[77] Medicine, National Library of Medical Subject Headings (MeSH). Disponible en: http://www.nlm.nih.gov/mesh/meshhome.html, 2003.

[78] Merrouni, Z. A., Frikh, B., y Ouhbi, B. (2019). Automatic keyphrase extraction: A survey and trends. Journal of Intelligent Information Systems, 54, 391-424. https://doi.org/10.1007/s10844-019-00558-9

[79] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., y McCallum, A., (2011). Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 262–272. http://dirichlet.net/pdf/mimno11optimizing.pdf

[80] Moreira, M., (2013). Being Agile: Your roadmap to successful adoption of Agile. Apress.

[81] Morris, P. 2013. Reconstructing Project Management Reprised: A knowledge perspective. Project Management Journal, 44: 6-23.

[82] Müller, H.M., Kenny, E.E., y Sternberg, P.W., (2003). Textpresso: An ontology-based information retrieval and extraction system for biological literatura PLoS Biol. 22003. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020309

[83] Mungall, C.J., (2004). Obol: integrating language and meaning in bio-ontologies. Comp Funct Genomics;5:509-20.

[84] Muñoz-Écija, T., Vargas-Quesada, B., y Chinchilla-Rodríguez, Z., (2017). Identification and visualization of the intellectual structure and the main research lines in nanoscience and nanotechnology at the worldwide level. Springer, 19, 62. https://doi.org/10.1007/s11051-016-3732-3

[85] Musen, M.A. Noy, N.F., y Shah, N.H., (2012). The National Center for Biomedical Ontology. J. Am. Med. Inform. Assoc., 19, 190–195.

[86] Naili, M., Chaibi, A.H., Ghezala, H.H., (2017). Comparative study of word embedding methods in topic segmentation. Procedia Computer Science, 112, 340-349. https://doi.org/10.1016/j.procs.2017.08.009.

[87] Ninan, J., Mahalingam, A., y Clegg, S., (2019). External Stakeholder Management Strategies and Resources in Megaprojects: An Organizational Power Perspective. Project Management Journal, *50*, 625–640. https://doi.org/10.1177/8756972819847045

[88] NLM: UMLS Reference Manual. Bethesda (MD): National Library of Medicine (US), Septiembre 2009. http://www.ncbi.nlm.nih.gov/books/NBK9676.

[89] Noh, H., Jo, Y., y Lee, S., (2015). Keyword selection and processing strategy for applying text mining to patent analysis. Expert Systems with Applications, 42, 4348-4360. https://doi.org/10.1016/j.eswa.2015.01.050

[90] Padalkar, M., y Gopinath, S., (2016). Six decades of project management research: Thematic trends and future opportunities. International Journal of Project Management 34, 1305-1321. https://doi.org/10.1016/j.ijproman.2016.06.006

[91] Pitsis, T.S., Sankaran, S., Gudergan, S., y Clegg, S.R., (2014). Governing projects under complexity: theory and practice in project management. International Journal of Project Management, 32, 1285-1290. https://doi.org/10.1016/j.ijproman.2014.09.001.

[92] Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X. y Jensen, L.J., (2015). DISEASES: Text mining and data integration of disease–gene associations. Methods, 74, 83-89. https://doi.org/10.1016/j.ymeth.2014.11.020.

[93] Pollack, J., y Adler, D., (2015). Emergent trends and passing fads in project management research: a scientometric analysis of changes in the field. International Journal of project management, 33, 236-248. https://doi.org/10.1016/j.ijproman.2014.04.011

[94] PMI, 2017. A guide to the project management body of knowledge (PMBoK guide), 6th. Ed. Project Management Institute, Pennsylvania, EEUU.

[95] Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M., y Morrell, M., (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. Stud Health Technol Inform, 84, 371-375. https://arxiv.org/abs/1903.12180

[96] Oellrich, A., Grabmuller, C., y Rebholz-Schuhmann, D., (2013). Automatically transforming pre-to post-composed phenotypes: EQ-lising HPO and MP. J Biomed Semantics; 4:29.

[97] Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J., y Hunter, L., (2004). The compositional structure of Gene Ontology terms. Pacific Symposium on Biocomputing; p. 214-225.

[98] Ogren, P.V., Cohen, K.B., Hunter, L., (2005). Implications of compositionality in the gene ontology for its curation and usage. Pacific Symposium on Biocomputing; p. 174-185.

[99] Quesada-Martinez, M., Mikroyannidi, E., Fernandez-Breis, J.T., y Stevens, R., (2015). Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. Artif Intell Med; 65(1): 35-48.

[100] Ravikumar, S., Agrahari, A., y Singh, S.N., (2015). Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics (2005–2010). Scientometrics, 102, 929–955. https://doi.org/10.1007/s11192-014-1402-8

[101] Resnik, P., (1995). Using information content to evaluate semantic similarity in a taxonomy. Proc International Joint Conferences on Artificial Intelligence (IJCAI); pp. 448–45.

[102] Rinaldi, A.M., Russo, C., y Tommasino, C., (2020). A semantic approach for document classification using deep neural networks and multimedia knowledge graph. Expert Systems with Applications, 114320. https://doi.org/10.1016/j.eswa.2020.114320

[103] Robinson, P., y Webber, C., (2014). Phenotype ontologies and cross-species analysis for translational research. PLoS Genet;10:e1004268.

[104] Robinson, P.N., Kohler, S., Bauer, S., Seelow, D., Horn, D., y Mundlos, S., (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet; 83:610–5.

[105] Rokaya, M., Atlam, E., Fuketa, M., Dorji, T.C., y Aoe, J., (2008). Ranking of field association terms using co-word analysis. Information Processing y Management, 44, 738-755. https://doi.org/10.1016/j.ipm.2007.06.001

[106] Ronda-Pupo, G.A. y Guerras-Martin, L.A., (2011). Dynamics of the evolution of the strategy concept 1962-2088: a co-word analysis. Strategic management journal, 33, 162-188. https://doi.org/10.1002/smj.948

[107] Rodríguez-García, M.A., Valencia-García, R., García-Sánchez, F., y Samper-Zapater, J.J., (2014). Ontology-based annotation and retrieval of services in the cloud. Knowl. Based Syst., 56, 15–25.

[108] Schober, D., Smith, B., Lewis, S.E., Kusnierczyk, W., Lomax, J., Mungall, C., Taylor, C.F., Rocca-Serra, P., Sansone, S.A., (2009). Survey-based naming conventions for use in OBO Foundry ontology development. BMC Bioinform;10:125.

[109] Schulz, S., Jansen, L., (2013). Formal ontologies in biomedical knowledge representation. YearB Med Inform;8:132-146.

[110] Shah, N., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A., y Musen, M., (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics; Suppl 9,10:14.

[111] Skunca, N., Altenhoff, A., y Dessimoz, C., (2012) Quality of computationally inferred gene ontology annotations. PLoS Comput. Biol., 8, e1002533.

[112] Slater, L., (2014). Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada, 33(2) 106. https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi

[113] Smalheiser, N.R., Torvik, V.I, Bischoff-Grethe, A., Burhans, L.B., Gabriel, M., Homayouni, R., Kashef, A., Martone, M.E., Perkins, G.A., Price, D.L., Talk, A.C., y West, R., (2006). Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. Journal of Biomedical Discovery and Collaboration. https://j-biomed-discovery.biomedcentral.com/articles/10.1186/1747-5333-1-8

[114] Smith, B. Ashburner, M., y Rosse, C., (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol., 25, 1251–1255.

[115] Smyth, H. J., y Morris, P. W. G., (2007). An epistemological evaluation of research into projects and their management: Methodological issues. International Journal of Project Management, 25, 423-436. https://doi.org/10.1016/j.ijproman.2007.01.006

[116] Söderlund, J., (2004) b. On the broadening scope of the research on projects: a review and a model for analysis. International Journal of Project Management, 22, 655-667. https://doi.org/10.1016/j.ijproman.2004.05.011

[117] Söderlund, J., (2004) a. Building theories of project management: past research, questions for the future. International Journal of Project Management, 22, 183-191. https://doi.org/10.1016/S0263-7863(03)00070-X

[118] Stevens, K., Kegelmeyer, P., Andrzejewski, D., y Buttler, D., (2012). Exploring Topic Coherence over many models and many

topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 952–961. https://www.aclweb.org/anthology/D12-1087.pdf

[119] Taboada, M., Rodriguez, H., Martínez, D., Pardo, M., y Sobrido, M.J., (2014). Automated semantic annotation of rare disease cases: a case study. Database (Oxford) : bau045.

[120] Tao C., Song, D., Sharma, D., Chute, C.G., (2013). Semantator: Semantic annotator for converting biomedical text to linked data, Journal of Biomedical Informatics. Volume 46, Issue 5, pages 882-893. https://doi.org/10.1016/j.jbi.2013.07.003.

[121] Tchechmedjiev, A., Abdaoui, A., Emonet, V., (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. BMC Bioinformatics 19, 405. https://doi.org/10.1186/s12859-018-2429-2

[122] Themistocleous, G., y Wearne, S. H., (2000). Project management topic coverage in journals. International Journal of Project Management, 18, 7-11. https://doi.org/10.1016/S0263-7863(99)00030-7

[123] Tripathi, S., Christie, K.R., y Balakrishnana, R., (2013) Geneontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large scale curation effort. Database, 2013, bau062.

[124] Tsatsaronis, G., Schroeder, M., y Paliouras, G., (2012) BioASQ: a challenge on large-scale biomedical semantic indexing and question answering. AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical. Text. AAAI Press, Arlington, VA, pp. 92–98.

[125] Uçkan, T., y Karcı, A., (2020). Extractive multi-document text summarization based on graph independent sets. Egyptian Informatics Journal, 21, 145-157. https://doi.org/10.1016/j.eij.2019.12.002

[126] Urbain, J., (2015). Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. J Biomed Inform; Suppl 58:143-149.

[127] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F., (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics, 4, 14-28. https://doi.org/10.1016/j.websem.2005.10.002.

[128] Urli, B., y Urli D., (2000). Project Management in North America, stability of the concepts. Project Management Institute, 31, 33-43. https://doi.org/10.1177/875697280003100305

[129] Van Eck, N. J., Waltman, L., Noyons, E., y Buter, R., (2010). Automatic term identification for bibliometric mapping. Scientometrics, 82, 581-596. https://doi.org/10.1007/s11192-010-0173-0

[130] Vanteru, B.C., Shaik, J.S., y Yeasin, M., (2008). Semantically linking and browsing PubMed abstracts with gene ontology. BMC Genomics, 9 (Suppl 1), S10.

[131] Vaughan, L., Yang, R., y Tang, J., (2012). Web co-word analysis for business intelligence in the Chinese environment. Aslib Proceedings, 64, 653-667. https://doi.org/10.1108/00012531211281788

[132] Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., y Berton, L., (2019). A multi-centrality index for graph-based keyword extraction. Information Processing y Management, 56, 102063. https://doi.org/10.1016/j.ipm.2019.102063

[133] Verspoor, C.M., Joslyn, C., Papcun, G.J., (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. SIGIR workshop on Text Analysis and Search for Bioinformatics; p. 51-56.

[134] Verspoor, K., Dvorkin, D., Cohen, K.B., Hunter, L., (2009). Ontology quality assurance through analysis of term transformations. Bioinform; 25(12):77-84.

[135] Vidal, J.C., Lama, M., Otero-García, E., y Bugarín, A., (2014) Graph-based semantic annotation for enriching educational content with linked data. Knowl. Based Syst., 55, 29–42.

[136] W3 Consortium. OWL 2 Web Ontology Language Document Overview (Second Edition), 2012. Disponible en: http://www.w3.org/TR/2012/REC-owl2-overview-20121211/ .

[137] Wan, X., y Xiao, J., (2008). Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd national conference on Artificial intelligence, 2, 855–860. https://dl.acm.org/doi/10.5555/1620163.1620205

[138] Wasserman, S., & Faust, K. 1994. Social Network Analysis: Methods and Applications.Cambridge: Cambridge University Press.

[139] Wessman, A., Liddle, S.W., y Embley, D.W., (2005). A generalized framework for an ontology-based data-extraction system. 4th Int. Conference on Information Systems Technology and its Applications. Palmerston North, New Zealand, pp. 239–253.

[140] Westbury, S.K., (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med;7:36.

[141] Whetzel, P.L., y NCBO Team (2013) NCBO Technology: powering semantically aware applications. J. Biomed. Semantics, 4 (Suppl 1), S8.

[142] Wikipedia (2014). Glossary of Project Management. Wikipedia, version en Inglés. Disponible en internet, URL: http://en.wikipedia.org/wiki/Glossary_of_project_management [Fecha último acceso: 22/05/2018].

[143] Wolfe, J.. Annotation technologies: A software and research review. Computers and Composition, 19-4, 2002, 471-497. https://doi.org/10.1016/S8755-4615(02)00144-5.

[144] Xia, N., Zou, P. X.W, Griffin, M.A., Wang, X., y Zhong, R., (2018). Towards integrating construction risk management and stakeholder management: A systematic literature review and future research agendas.

[145] Yan, B.N., Lee, T.S., y Lee, T.P., (2015). Mapping the intellectual structure of the Internet of Things (IoT) field (2000–2014):

a co-word analysis. Scientometrics, 105, 1285–1300. https://doi.org/10.1007/s11192-015-1740-1

[146] Zhang J., Yu Q., Zheng F., Long C., Lu Z., y Duan Z., (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. Journal of the Association for Information Science and Technology, 67, 967-972. https://doi.org/10.1002/asi.23437

[147] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., Shen, B., (2013). Biomedical text mining and its applications in cancer research. J Biomed Inform;46: 200-211.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS