UNIVERSIDAD PONTIFICIA
ICAI ICADE
COMILLAS
MADRID
ESCUELA TÉCNICA SUPERIOR
DE INGENIERÍA
DEPARTAMENTO DE TELEMÁTICA Y COMPUTACIÓN

ANÁLISIS Y OPTIMIZACIÓN DEL RECURSO UMLS EN LA RECUPERACIÓN DE INFORMACIÓN BIOMÉDICA MEDIANTE MÉTRICAS DE SIMILITUD SEMÁNTICA

Tesis para la obtención del grado de Doctor
Director: Prof. Dr. D. David Contreras Bárcena
Autor: Israel Alonso Martínez

Madrid 2016

A Toñi, por todo el apoyo que me ha
dado durante este largo camino
estando siempre ahí.

A mis padres, darles las gracias por
todo lo que hicieron por mí, nunca
os olvidaré y dejaré de querer.

Agradecimientos

Parecía que este momento no iba a llegar nunca, pero al final después de mucho esfuerzo y no sin complicaciones parece que el camino llega a su final, o por lo menos a un punto y aparte. Tengo que reconocer que he pasado por momentos muy complicados en el desarrollo de esta tesis, cuando las cosas parecían no fluir como deberían, pero también otros muy satisfactorios cuando todo empezaba a tomar forma y se empezaban a ver destellos, hasta por fin ver la luz al final del camino.

Para llegar a este punto, no me puedo olvidar de todas aquellas personas que han estado ahí durante este largo tiempo ayudándome de diferentes maneras, cada uno dentro de sus posibilidades, para la consecución de esta tesis.

Quiero dar las gracias a Toñi, por todo su apoyo incondicional y amor que me ha demostrado durante todo este largo camino, no solo en las dificultades que han surgido derivadas del desarrollo de esta tesis, si no por todos los obstáculos que nos ha ido poniendo la vida por delante y que gracias a su forma de ser, optimismo y paciencia hemos podido sortear juntos. Creo que no podría haber llegado a este punto sin ella.

También quiero acordarme y dar las gracias en estos momentos tan especiales a mi familia. Especialmente a mis padres, que sin duda si no hubiera sido por ellos nunca habría llegado hasta aquí. A mi madre, que fue y será el pilar central de mi vida y educación, a pesar de separarnos tan pronto, sé que la llevo conmigo ayudándome siempre. A mi padre, me entristece el que no haya podido llegar a verme terminar este camino y poder decirle "ya he terminado". A los dos, nunca os olvidaré. También quiero dar las gracias a todos los familiares que me han ayudado de alguna u otra manera y me demostraron o me siguen demostrando que están ahí para cuando los necesitas, gracias Jesusa y Modesta.

Y como no, no puedo olvidarme de David. Si alguien es el principal responsable de que llegara hasta este punto es él. Quiero darte las gracias por todo lo que me has trasmitido, no solo desde el punto de vista de investigación, sino por tu sincera amistad, tu paciencia y apoyo constante. Incluso en esos momentos que siempre querías "algo más" y parecía que nunca íbamos a terminar, me demostrabas el interés por mi trabajo y por las cosas bien hechas. Sé que no ha tenido que ser fácil lidiar conmigo, pero nunca podré dejar de agradecerte tu dedicación, confianza en mí y sobre todo por ser mi "amigo". Aunque en algunos momentos hubiera preferido que este camino hubiera sido más corto, también puedo decir que hemos pasado ratos divertidos. También quiero acordarme de Olga, Álvaro y Aitor por el tiempo que les he robado de forma indirecta.

Igualmente quiero dar las gracias a Francisco, como parte responsable de iniciarme en este trabajo y por su colaboración desinteresada. A Mario, por sus buenos consejos y confianza mostrada y a Mateo por ser el primero en involucrarme en esta aventura. También, dar las gracias a todos aquellos compañeros que me han apoyado y han demostrado su interés por mi trabajo.

Y ya por último, a todos los amigos que han sufrido este proceso y que siempre han tenido una palabra de ánimo y apoyo.

Gracias a todos.

Índice de figuras

Índice de tablas

Capítulo 1

1. Introducción

Actualmente el volumen de datos almacenados en soportes digitales crece de forma exponencial, siendo este cada vez mayor. Pero no es solo el volumen de datos lo que conlleva su mayor importancia, sino la información en ella contenida. Hoy en día uno de los campos que están teniendo un mayor auge en el procesamiento digital de la información, es el de la biomedicina. Donde frecuentemente se incorporan nuevas tecnologías e investigaciones en apoyo y mejora de múltiples disciplinas (Wang, Ranjan, Kolodziej, Zomaya & Alem, 2015), como pueden ser: el almacenamiento y procesado de informes médicos o imágenes médicas digitalizadas; la aplicación de la telemedicina; la interacción con aplicaciones móviles de salud; la personalización de los cuidados médicos; el intercambio de información biomédica; estudios epidemiológicos de población; la robótica clínica; etc.

Dentro de este amplísimo entorno tecnológico asociado a la biomedicina, el trabajo desarrollado en esta tesis se centra en la Recuperación de Información (R.I.) basada en registros médicos electrónicos (Electronic Health Records, o EHR) (Hoffman, 2010; Prokosch, & Ganslandt, 2009). Donde uno de los mayores retos a los que se enfrentan los sistemas de R.I. en este entorno, es el de la identificación de aquellos registros médicos (EHR) que respondan eficazmente a las necesidades de información definidas por los usuarios de dichos sistemas. Para la consecución con éxito de esta tarea, será crítico el reconocimiento de ciertos patrones en el conjunto de historias médicas tratadas que permitan, la identificación de pacientes en grupos de cohortes, la detección de brotes o epidemias, así como la anticipación a ciertas enfermedades (Roque, Jensen et al., 2011).

Sin embargo, la mayor dificultad para la consecución efectiva de la recuperación de aquellos registros médicos electrónicos de relevancia para una consulta, proviene de la necesidad de procesar el lenguaje natural que expresa su información. Además hay que tener en cuenta que el lenguaje natural no solo es complejo en su tratamiento, sino que es muy dependiente del contexto al que esté vinculado. Así por ejemplo, en un contexto genérico y amplio como el de la lengua inglesa, el cual ha sido tratado en numerosos trabajos, es necesaria la utilización de recursos u ontologías que representen dicho dominio de conocimiento, como es WordNet (Fellbaum, 1998). Desgraciadamente estos

recursos no son apropiados o tienen una aplicación limitada en contextos particulares y especializados como es el de la biomedicina. Ya que la información manejada en este dominio es compleja, ambigua y especifica. El lenguaje utilizado en este contexto, carece de construcciones gramaticales correctas, es variable, no estructurado, contiene abreviaturas, se utilizan codificaciones y diversas representaciones de su información contenida. Por tanto, el análisis de la información biomédica requerirá el uso de terminologías especializadas (Friedman, Kra, & Rzhetsky, 2002) y nuevas destrezas en la recuperación de información en esta rama de la tecnología (Alpi, 2005). Para ello, será necesaria la utilización de ciertos recursos especializados, — diccionarios o tesauros como UMLS (McCray, Aronson, Browne, Rindflesch, Razi, & Srinivasan, 1993) —, que den valor semántico a la información almacenada en este entorno.

Cabe reseñar que trabajos previos han mostrado su interés por aplicar sobre contextos generales como la lengua inglesa, métricas que definan el grado de similitud semántica entre términos (Collins, & Loftus, 1975), basándose en la infraestructura del recurso "WordNet" (Meng, Huang, & Gu, 2013). Sin embargo estas aproximaciones no ofrecen unos resultados satisfactorios cuando son aplicados a un dominio especializado como el de la biomedicina. Esto se debe a que el recurso universal de la lengua inglesa "WordNet", tiene una cobertura muy limitada en contextos especializados. *"Only 2% of the domain-specific concepts from UMLS were found in WordNet, but 83% of the domain-specific concepts from WordNet were found in the UMLS"* (Burgun, & Bodenreider, 2001). Trabajos posteriores tratan de resolver este problema, mediante la incorporación de recursos y ontologías específicas (MeSH, y SNOMED CT) en el estudio y adaptación de las distintas métricas de similitud semántica en el campo de la biomedicina, pero siempre desde un punto de vista teórico y sobre entornos controlados. (Caviedes & Cimino, 2004; Al-Mubaid & Nguyen, 2006; Nguyen & Al-Mubaid, 2006; Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Batet, Sánchez, & Valls, 2011).

1.1. Motivación

Sin embargo la confluencia de las líneas de investigación aportadas por el desarrollo del Metatesauro UMLS y de las diferentes métricas de similitud semántica en ámbito de la biomedicina, no han llegado a confluir en su aplicación en contextos reales de recuperación de información biomédica.

En consecuencia, este trabajo trata de cubrir este hueco, ayudando en la mejora de los sistemas de recuperación basados en registros médicos electrónicos, en función de su contenido semántico. Siendo obligado para ello, poder identificar e interpretar las necesidades de información reflejadas en cualquier consulta, para una selección eficiente de aquellos documentos médicos más relevantes en términos de proximidad semántica. Para la consecución de este objetivo, será necesario definir y representar mediante conceptos biomédicos, tanto la información contenida en cada uno de los registros médicos, como de las consultas a las que se enfrentan. De esta manera, se podrá establecer la proximidad semántica existente entre ambos, emulando el proceso humano de enjuiciamiento del grado de similitud semántica existente entre dos términos (Rubenstein, & Goodenough, 1965; Miller, & Charles, 1991).

1.2. Estructura memoria de tesis

Este documento se estructura en nueve capítulos, en los que se describe el ámbito de investigación en el que se ha centrado el trabajo de esta tesis, así como el estudio y análisis de los trabajos de mayor referencia en este entorno. A continuación se describen cada uno de los siguientes capítulos de la tesis.

En el capítulo 2, se describe el ámbito principal de información asociada al problema tratado en esta tesis. Así como la estructura, componentes y características de la base de conocimiento del dominio biomédico utilizada en esta tesis.

Seguidamente, en el capítulo 3, se ha realizado un extenso estudio y análisis de los diferentes trabajos propuestos por la comunidad científica, en el ámbito de la recuperación de información (R.I.). Donde se han identificado dos principales líneas de trabajo, la primera aplicada a la información basada en Corpus y la segunda en Colecciones u Ontologías. Además se ha realizado un estudio de los diferentes trabajos centrados en la definición de métricas de similitud semántica apoyadas en ontologías especializadas o dependientes del dominio biomédico.

Posteriormente, en el capítulo 4, se justifica y define el alcance de la tesis, describiendo cada uno de los objetivos a cumplir para la consecución final de este trabajo. Además se detalla cada una de las principales herramientas utilizadas en el desarrollo de esta tesis.

En el capítulo 5, se ha definido un marco de trabajo para la evaluación de las diferentes métricas de similitud semántica sobre un contexto teórico de información biomédica, donde se evalúa el impacto de la utilización de diferentes relaciones y recursos contenidos en el Metatesauro UMLS, como optimización en el funcionamiento y resultados aportados por las métricas de similitud semántica. Los resultados de este capítulo han sido publicados en (Alonso, Contreras, & Romero, 2012; Alonso, Contreras, & Romero, 2013)

A continuación en el capítulo 6, se da continuidad al trabajo anterior, enfocado en un novedoso estudio experimental para la aplicación y evaluación de las métricas de similitud semántica en un entorno real de información biomédica. Para ello, en este capítulo se describen los pasos realizados en el desarrollo de un sistema de recuperación de información basado en métricas de similitud semántica, donde se evaluará el impacto de las métricas *Path* e *Intrinsic IC-Path* en un contexto real de recuperación de información sobre registros médicos electrónicos obtenidos del TREC Medical Records Track 2011. Se realizará una exhaustiva evaluación de los resultados obtenidos tanto de manera individualizada, como agregada. Obteniendo como principal conclusión de estos resultados, el hecho de que ambas métricas (*Path* e *Intrinsic IC-Path*), muestran un comportamiento similar en un contexto real de información, al contrario de lo ocurrido en un contexto teórico y cerrado. Los resultados de este capítulo han sido aceptados para su publicación en la revista "Expert Systems with Applications" (Alonso, & Contreras, 2015)

Seguidamente, se realiza un resumen de las conclusiones y aportaciones finales de esta tesis, así como de los trabajos futuros que darán continuidad a las investigaciones aquí desarrolladas. Finalizando, con los trabajos publicados durante el desarrollo de esta tesis y las referencias bibliográficas consultadas en este proceso.

Capítulo 2

2. Ámbito del problema

Este capítulo se centra en describir la problemática asociada al entorno de la recuperación de información en el contexto de la información biomédica. Para ello se realiza un resumen de los principales tipos de repositorios en este ámbito. Además de una descripción detallada del principal recurso diseñado como base de conocimiento del dominio biomédico (UMLS).

2.1. Introducción

Actualmente la utilización de las tecnologías de la información y las comunicaciones (TIC) dentro del ámbito de la medicina y la salud, han contribuido a la mejora de la gestión, de los servicios sanitarios, así como a la disponibilidad e intercambio de la información y conocimiento del dominio biomédico (Sittig, et al, 2010). El problema en el tratamiento y recuperación de documentos biomédicos se centra en la complejidad de la información contenida, siendo mayor aún esta dificultad cuando el contexto de búsqueda está compuesto por información clínica. Esto se debe a que los repositorios de documentos clínicos pueden representar la información contenida de diferentes formas para referenciar o describir los mismos o similares conceptos médicos. Además, en estos documentos la información puede estar expresada en texto libre, texto no estructurado asociado a vocabularios controlados del dominio médico, abreviaturas y codificaciones para la descripción de dicha información.

En los siguientes apartados se pretende dar una visión más detallada del entorno del problema y de aquellos trabajos relacionados en el área de la recuperación de información en el ámbito biomédico. Para ello, previamente se ha de distinguir los diferentes documentos digitales tratados por las tecnologías de la información aplicadas en esta área. Estos se pueden dividir básicamente en documentos biomédicos que conforman la literatura médica (artículos, revistas, abstracts, estudios, etc.) y por los documentos clínicos, los cuales contienen los informes o historias clínicas de los pacientes y son sobre los que se focaliza este trabajo de investigación.

2.2. Documentación digital basada en literatura biomédica

Actualmente la documentación biomédica está formada por una ingente cantidad de trabajos elaborados mediante publicaciones de diferentes tipos de documentos, como: artículos, artículos cortos (*letters*), resúmenes (*abstracts*), etc. Donde se reflejan las investigaciones y avances en el mundo de la biomedicina. Dichos documentos, hoy en día están disponibles en formato digital para la comunidad científica. Siendo su accesibilidad no restringida (en algunos casos mediante subscripción), gracias a internet y las tecnologías de la información. Pero ante tanta abundancia y diversidad de documentos e información científica, la dificultad de los usuarios para identificar o extraer aquellos documentos que les servirán como paso previo a la obtención del conocimiento buscado, se hacen evidentes.

Para ello surgen las bases de datos bibliográficas, las cuales pueden agrupar miles o millones de referencias de documentos almacenados electrónicamente y que se pueden buscar y recuperar de manera interactiva a través de métodos de consulta o interrogación. Este es el caso de MEDLINE[1] (Medical Literature Analysis and Retrieval System Online) es una base de datos de referencias bibliográficas a artículos científicos sobre biomedicina y salud.

El sistema de búsqueda PubMed[2], es el sistema de recuperación de información desarrollado por el NCBI (National Center for Biotechnology Information) de la NLM. Este sistema integra 18 millones de referencias provenientes de MEDLINE y de otras fuentes científicas. Estas bases de datos de referencias bibliográficas sobre artículos científicos de biomedicina y salud, han demostrado su eficiencia en la búsqueda de documentos biomédicos (Hersh & Hickam, 1998). Reflejando que los resultados obtenidos por los métodos de búsqueda avanzados son significativamente más eficaces que los métodos de búsqueda simple, basadas solamente en cadenas de texto. Estos estudios determinan que nuevas investigaciones y desarrollos son necesarios para mejorar la utilidad y el rendimiento de los sistemas de recuperación de información en este dominio.

Paralelamente y confirmando las ideas anteriormente expuestas, distintos trabajos observan la complejidad en el tratamiento del lenguaje del domino biomédico, así como la dificultad en resolver las necesidades de información de los usuarios a partir de las premisas iniciales de las consultas. Para ello, estudios como (Stapley & Benoit 2000) aplican al concepto de búsqueda, la coocurrencia de los nombres o alias de genes con funciones biológicas relacionadas. Extrayendo de la colección Medline, las coocurrencias para un tipo de gen en particular e incluir y combinar estas términos en la consulta documental. Además el conjunto de documentos recuperados podrán ser filtrados mediante el uso de términos MeSH (tesauro de títulos y términos).

Repositorios de pruebas basados en literatura biomédica

Actualmente existen ciertas colecciones de prueba basadas en literatura biomédica, extraídas fundamentalmente de artículos recogidos en MEDLINE. Estas colecciones permiten a los investigadores realizar los estudios, pruebas y evaluaciones en el

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[2] http://www.ncbi.nlm.nih.gov/pubmed

tratamiento de documentos biomédicos en diferentes disciplinas como, la recuperación de información, extracción de conocimiento, procesamiento del lenguaje natural y tratamiento de técnicas basadas en pregunta/respuesta.

Algunas de estas colecciones de prueba más ampliamente usadas son:

- ***TREC[3] (Genomics Track[4].)***
  Es uno de los bancos de pruebas más utilizados en las investigaciones de recuperación de información. Las diferentes colecciones de testeo son proporcionadas por el TREC *(Text REtrieval Conference)* y copatrocinadas por el National Institute of Standards and Technology (NIST) y el Departamento de Defensa de EE.UU. Iniciado en 1992 como parte del programa TIPSTER. Constituyen una serie de talleres o *tracks*, con colecciones específicas, para fomentar la investigación en la recuperación de la información y sus aplicaciones relacionadas. Proporcionando una colección de ensayo grande y uniforme, de forma que nuevas investigaciones tengan la oportunidad de probar sus sistemas contra colecciones reales. Pudiendo así, contrastar los resultados propuestos sobre una base sólida y estandarizada.

  Algunas de las líneas de trabajo definidas en el TREC, se centra en el tratamiento y recuperación de documentos basados en artículos de biomedicina, es el TREC Genomics Track Esta colección de testeo se desarrolló entre los años 2003 y 2007. Siendo actualizados los archivos de protocolo y datos a finales del 2008 para reflejar el final de la colección, la cual aún está disponible. Una visión general de la colección se puede obtener en el artículo de Hersh y Voorhees (Hersh & Voorhees, 2009).

- ***OHSUMED[5]***
  Es una colección de prueba formada por un subconjunto de artículos MEDLINE con orientación médica que consta de 348.566 referencias a artículos (de un total de 7 millones) que abarcan 270 revistas médicas entre los años 1987 y 1991. Los únicos campos presentes son el título, el resumen, los términos de indexación MeSH, el autor, la fuente y el tipo de publicación. Esta base de datos de prueba no es útil en la búsqueda de información contenida. Su único propósito, es ser útil en la investigación para la recuperación de información médica, mediante experimentos desarrollados con el sistema de recuperación SMART (Hersh, Buckley, Leone & Hickam 1994).

2.3. Documentación digital basada en registros médicos

En un principio, la utilización de las Tecnologías de la Información (TI) en el área de la Salud y la Medicina estaba ligada casi exclusivamente a las tareas administrativas y contables de los centros sanitarios, ya que más que la automatización de tareas o la ayuda en la gestión, lo que se buscaba era reducir el espacio necesario para el almacenamiento de los expedientes sustituyendo el papel por el formato electrónico. Más adelante, las TI pasaron a ocuparse no solo de tareas administrativas sino también de las tareas relativas a

[3] http://trec.nist.gov/
[4] http://ir.ohsu.edu/genomics/
[5] http://ir.ohsu.edu/ohsumed/

la gestión hospitalaria, dando paso a la informatización de muchos de los procesos relacionados con la atención de pacientes (Kuhn & Giuse, 2001). Actualmente estos avances en la gestión hospitalaria y en el seguimiento en la atención al paciente, se está trasladando como soporte y aplicación en medios digitales. Definiendo un conjunto de tecnologías de la información de la salud, que contribuyen en el ámbito de la atención sanitaria del paciente, mejorando la calidad de sus servicios y la reducción de costes, mediante el uso y tratamiento de la información almacenada en historiales clínicos electrónicos de los pacientes (Hillestad, Bigelow & et al. 2005).

Repositorios de pruebas basados en registros médicos

Actualmente se están realizando grandes esfuerzos para la generación de repositorios de documentos basados en información clínica como apoyo a la investigación.

- ***TREC 2011[6] (Medical Records Track).***
  Dentro de los distintos desafíos propuestos en el TREC 2011, conocidos como "tracks". Hay que destacar el nuevo reto incorporado de manera inaugural en ese año y que se ha definido como *Medical Records Track*. Pretende fomentar la investigación sobre tecnologías que permitan la recuperación de registros médicos electrónicos, basándose en la semántica contenida en sus campos de texto libre. La capacidad para recuperar registros médicos a partir de su contenido semántico, será de gran utilidad en áreas tales como estudios epidemiológicos y ensayos clínicos. El material de entrenamiento y testeo están disponibles en el National Institute of Standards and Technology (NIST[7]), se compone de un repositorio de documentos clínicos (donde los documentos denominados "*reports*" son agrupados en visitas "*visits*"), un conjunto de necesidades de información (donde las consulta de denominan como "*topics*") y unos juicios de relevancia definidos por los expertos para cada uno de los *reports* enfrentados a los diferentes *topics*.

- ***SERP (Stockholm Electronic Patient Record Corpus)***
  Surge como proyecto de investigación del departamento de Ciencias de la Computación y Sistemas de la Universidad Stockholm. En la que se trata de de-identificar los datos asociados al paciente incluido en el documento clínico (Dalianis, Hassel, Henriksson & Skeppstedt 2012). Extraídos de uno de los mayores sistemas de registros electrónicos médicos en Suecia, the Stockholm EPR (Electronic Patient Records) Corpus. El objetivo es crear un repositorio gold estándar denominado, The Stockholm EPR PHI Corpus, que esté disponible para otros investigadores en el futuro.

- ***CLEF[8]. (CLinical E-science Framework)***
  CLEF tiene como objetivo, el desarrollo de un repositorio de información interoperable, derivado de la operativa basada en registros electrónicos del paciente, para permitir el acceso de una manera ética y sencilla a la

[6] http://trec.nist.gov/pubs/call2011.html
[7] http://www.nist.gov/index.html
[8] http://nlp.shef.ac.uk/clef/

información del paciente como apoyo al cuidado clínico y a la investigación biomédica (Roberts, Gaizauskas, et al., 2007). Este corpus está compuesto por narrativa clínica, informes histopatológicos e informes de imágenes de 20 mil pacientes. (Kalpathy-Cramer, et al., 2015).

2.4. Base de Conocimiento del Dominio Biomédico (UMLS)

En este apartado se describe la base de conocimiento más importante del dominio biomédico, reflejándolo desde diferentes aspectos. Si bien existen, un gran número de recursos que integran la información del dominio biomédico, como vocabularios controlados, bases de datos, clasificaciones, codificaciones, etc. Hay que decir que el proyecto de la National Library of Medicine del National Institute of Health de Estados Unidos, denominado UMLS[9] (The Unified Medical Language System), pretende incorporar todo este conocimiento del dominio biomédico bajo una única ontología, que integre y relacione todos los recursos contenidos. Las ontologías según (Bodenreider, & Burgun 2005), juegan un papel crucial en el desarrollo de la informática médica, contribuyendo al procesamiento del lenguaje natural y permitiendo el acceso a recursos de información heterogéneos. De esta forma, las ontologías actúan como un recurso del conocimiento del dominio biomédico para las aplicaciones en dicho entorno.

UMLS es un sistema que permite trabajar con referencias cruzadas entre diversos vocabularios controlados y clasificaciones de terminologías médicas. La mayoría de estas referencias cruzadas se obtienen gracias al análisis léxico de los términos, y es por esto que UMLS se suele incluir dentro de la categoría de sistemas léxicos de clasificación (Bodenreider, & Burgun 2005).

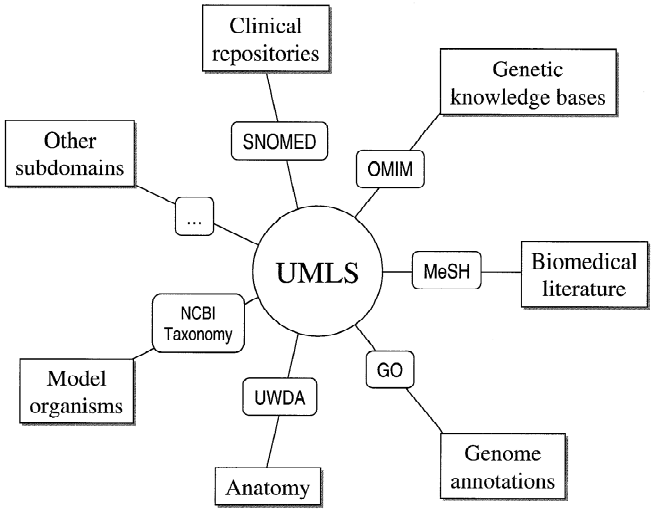

Figura 2.1 Subdominios integrados en UMLS. Fuente: (Bodenreider, 2004)

UMLS es actualmente el sistema que integra la mayor cantidad de vocabularios de terminología médica, lo que permite la integración de gran cantidad de aplicaciones que utilizan los distintos vocabularios contemplados en UMLS (Figura 2.1).

[9] http://www.ncbi.nlm.nih.gov/books/NBK9676/

El sistema está formado por tres componentes fundamentales:

- **Metatesauro**: Contiene más de 1 millón de conceptos (CUIs) biomédicos de más de 100 vocabularios.
- **Red semántica**: Define 135 categorías y 54 relaciones semánticas que categorizan el dominio biomédico.
- **Lexicón especializado y herramientas léxicas**: Contiene información léxica y herramientas para el procesamiento del lenguaje natural.

Metatesauro[10]

Es el núcleo de UMLS. Es una base de datos multi-idioma que intenta reunir la totalidad de los conceptos utilizados en el campo de la biomedicina. Incluye las diferentes descripciones que dichos conceptos pueden tomar en los distintos vocabularios y clasificaciones así como también las relaciones que existen entre ellos. Actualmente contiene más de 5 millones de términos provenientes de más de 100 vocabularios.

Los términos del Metatesauro se agrupan en conceptos de acuerdo a su significado. Cada concepto posee características específicas que definen su significado y lo relacionan con sus correspondientes términos en las distintas ontologías o vocabularios que forman parte de UMLS. Las relaciones entre conceptos son del tipo 'es un', 'parte de', 'causado por', etc. Los conceptos tienen asociado un identificador único CUI (Concept Unique Identifier) a través del cual se relaciona con otros términos y variantes que puedan existir para referirse al mismo concepto.

El Metatesauro organiza sus términos en una estructura de cuatro niveles:

- **CUI (Concept Unique Identifier)**: Código que identifica de manera única a cada concepto dentro del Metatesauro, el cual representa un significado. Agrupando en un código único aquellos términos sinónimos para dicho significado. Comienza por la letra 'C'.
- **LUI (Lexical Unique Identifier)**: Variantes léxicas o términos de un determinado concepto. Comienza por la letra 'L'.
- **SUI (String Unique Identifier)**: Representa cada cadena descriptiva asociada a un término. Son las distintas cadenas de caracteres que pueden encontrarse en los vocabularios del Metatesauro para referirse a un determinado concepto en mismos o diferentes lenguajes. Las variaciones entre mayúsculas y minúsculas, plurales y singulares, o signos de puntuación se consideran cadenas diferentes. Una de ellas será designada como nombre o término preferido. Comienza por la letra 'S'.
- **AUI (Atom Unique Identifier)**: Corresponden a la ocurrencia de cada cadena procedente de un recurso o vocabulario del Metatesauro.

En el siguiente ejemplo, se puede apreciar la estructura de cuatro niveles con la que se representan los conceptos en UMLS (Figura2.2, Figura 2.3).

[10] http://www.ncbi.nlm.nih.gov/books/NBK9684/

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

Figura 2.2 Estructura de representación de conceptos de cuatro niveles. Fuente: www.nlm.nih.gov

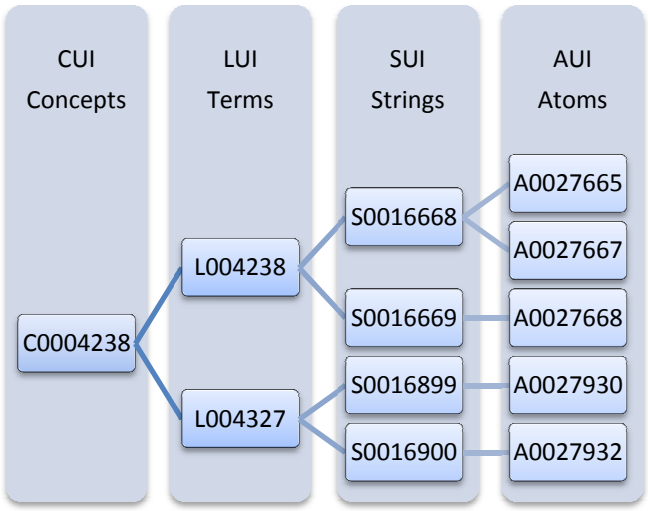

Figura 2.3 Estructura de representación de conceptos.

Es importante destacar que el Metatesauro no es un vocabulario en sí mismo, sino que integra la mayoría de los vocabularios controlados y clasificaciones que se utilizan en el área de la biomedicina y establece relaciones entre sus términos. El Metatesauro recoge y mantiene los significados, atributos, conexiones jerárquicas y relaciones que puedan existir entre los términos de los vocabularios, además de incluir nuevas relaciones entre conceptos y términos de vocabularios distintos.

El 62% de los vocabularios controlados a los que hace referencia el Metatesauro, están construidos bajo terminología inglesa. Sin embargo, el Metatesauro completa esta definición incluyendo, vocabularios con terminología de otros 17 idiomas (español,

francés, alemán, japonés, portugués, etc.) por lo que puede utilizarse como traductor de términos médicos.

Los vocabularios del Metatesauro cubren distintas áreas o especialidades de la biomedicina y se clasifican en categorías[11].

Lexicón especializado y herramientas léxicas[12]

El lexicón especializado es una base de datos con información lexicográfica de términos del dominio biomédico, cuyo objetivo es permitir el procesamiento del lenguaje natural. Actualmente el lexicón contiene 60.000 registros con aproximadamente 120.000 formas. Los términos incluidos en el lexicón se obtienen desde distintas fuentes, actualmente su núcleo está compuesto por más de 20.000 palabras extraídas de las colecciones de artículos de MEDLINE y del Dorland's Illustrated Medical Dictionary. Incluye además las 10.000 palabras más populares del American Heritage Word Frequency Book y 2.000 palabras de las definiciones del Longman's Dictionary of Contemporary English, entre otros.

El lexicón está organizado en entradas, cada una de las cuales representa una palabra o ítem léxico. Las entradas describen las propiedades sintácticas, morfológicas y ortográficas del ítem al que representan. Actualmente el lexicón contiene 60.000 registros con aproximadamente 120.000 formas (McCray, Srinivasan & Browne, 1994).

Además de la base de datos de términos incluye un conjunto de herramientas software desarrolladas en Java que permiten llevar a cabo el procesamiento del lenguaje natural. Estas herramientas, utilizadas junto con el propio lexicón, permiten que los usuarios desarrolladores de aplicaciones puedan implementar programas de procesamiento de lenguaje natural para el domino biomédico. Las herramientas léxicas del sistema son las siguientes:

- **Norm**: Generador de *strings* normalizados.
- **Wordind**: Generador de índices de palabras.
- **LVG** (Lexical Variant Generator): Generador de variantes léxicas.

Red Semántica[13]

La Red Semántica de UMLS, proporciona una clasificación consistente de todos los conceptos incluidos en el Metatesauro y del conjunto de relaciones que existen entre ellos. Está compuesta y estructurada en tipos y relaciones semánticas:

- **Tipos Semánticos**: Pueden verse como categorías de alto nivel que engloban a un conjunto de conceptos.
- **Relaciones Semánticas**: Son las relaciones que existen entre los tipos semánticos de la red y que pueden ser significativas o útiles desde el punto de vista médico. Los tipos semánticos permiten la categorización semántica de un amplio grupo de terminologías en múltiples dominios de especialidad.

[11] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/
[12] http://www.ncbi.nlm.nih.gov/books/NBK9680/
[13] http://www.ncbi.nlm.nih.gov/books/NBK9679/

Cada concepto del Metatesauro tiene asignado al menos un tipo semántico de la red. Los enlaces entre tipos semánticos además de dibujar la estructura de la red, representan las relaciones entre los conceptos que pueden encontrarse en el dominio. Si bien la información específica de cada concepto está almacenada en el Metatesauro, la red facilita información sobre los tipos semánticos básicos que se le han asignado a estos conceptos y define las relaciones que se pueden establecer entre ellos. Permitiendo una expresividad y tratamiento semántico del conocimiento del dominio biomédico (Kashyap, 2003)

Recursos contenidos en UMLS

A continuación, se realiza una descripción de algunos de los recursos más importantes contenidos en el Metatesauro UMLS.

SNOMED-CT

Systematized Nomenclature of Medicine-Clinical Terms[14,15]. Es considerada la terminología clínico sanitaria universal multilingüe más completa y exhaustiva actualmente. Creada originalmente por el Colegio Americano de Patólogos (CAP). A partir de abril de 2007, paso a ser propiedad, mantenida y distribuida por la IHTSDO (International Health Terminology Standards Development Organisation), una asociación sin ánimo de lucro, localizada en Dinamarca. La NLM (National Library of Medicine) en EE.UU., es el miembro de la IHTSO que distribuye SNOMEC CT sin coste, bajo sus acuerdos de licencia. Los nuevos términos de licencia se han incorporado en la propia licencia de uso del Metatesauro UMLS.

SNOMED-CT, es un vocabulario normalizado que permitirá la representación del contenido de los documentos clínicos para su interpretación automática e inequívoca entre diferentes sistemas de forma precisa y en distintos idiomas, facilitando el acceso a la información relevante para la toma de decisiones clínicas.

En España, SNOMED-CT es la terminología clínica de referencia seleccionada para la Historia Clínica Digital del Sistema Nacional de Salud (HCDSNS), lo que supone un paso hacia la interoperabilidad semántica de la HCDSNS (Historia Clínica Digital del Sistema Nacional de Salud). Para poder disponer de esta terminología clínica, en el año 2008, el Ministerio de Sanidad y Política Social inició las gestiones para el ingreso de España como miembro ordinario la IHTSDO. Convirtiendo al Ministerio de Sanidad y Política Social en el organismo que oficialmente puede distribuir SNOMED-CT, junto con las ediciones y extensiones españolas que se desarrollen, tanto a organizaciones públicas como privadas, dentro de nuestro territorio nacional. Se compone de dieciocho jerarquías independientes que reflejan los conceptos en SNOMED-CT y sus frecuencias de distribución.

ICD-9CM /ICD-10CM

ICD-9CM[16] (The International Classification of Diseases - 9 Clinic Modification), se basa en la novena revisión o modificación clínica de la Organización Mundial de la Salud, para la Clasificación Internacional de Enfermedades.

[14] http://www.ihtsdo.org/snomed-ct/
[15] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[16] http://www.cdc.gov/nchs/icd.htm

La Clasificación Internacional de Enfermedades, es el sistema oficial y universal de asignación de códigos, para todos aquellos posibles diagnósticos y procedimientos médicos. Es uno de los sistemas más importante de codificación y clasificación que permite las comparaciones internacionales y la monitorización de los problemas de salud en todos los ámbitos de actuación asistenciales.

La CIE9-MC es la clasificación que se lleva utilizando en España desde hace ya dos décadas para la codificación clínica de los procesos atendidos en hospitales.

ICD-10 – Es la décima revisión, de la Clasificación Internacional de Enfermedades, de la Organización Mundial de la Salud.

MeSH

MeSH (Medical Subject Heading Terms), es el vocabulario controlado del U.S. National Library of Medicine's para el indexado de los artículos biomédicos, que pueden ser consultado desde MEDLINE/PubMed. La terminología MeSH ofrece una manera consistente para la recuperación de información en la que se puede usar diferente terminología para el mismo concepto.

Es un conjunto de términos organizados en una estructura jerárquica en forma de árbol denominados "*trees*"[17], que permite realizar búsquedas con distintos niveles de especificidad. Para facilitar la búsqueda dentro del vocabulario controlado de una forma más sencilla se puede utilizar el MeSH Browser[18].

RxNorm

La RxNorm[19] es una nomenclatura estandarizada de medicamentos clínicos y dispositivos de administración de fármacos. Producido por la National Library of Medicine (NLM). En consulta con la "Food and Drug Administration, the Department of Veterans Affairs", y la organización "HL7". La RxNorm proporciona estándares de denominación para medicamentos clínicos y para las formas de su dosificación y administración.

NDC

(National Drug Code[20]) Código Nacional del Medicamento, establecido originalmente como una parte del programa de reembolso de medicamentos extra hospitalario de Medicare. La NDC, sirve como identificador universal de medicamentos para los seres humanos. La edición actual del Código Nacional de Drogas se limita a los medicamentos recetados y a un grupo selecto de productos OTC.

LOINC

(Logical Observation Identifiers Names and Codes[21]), es un sistema de codificación para mensajes de contenido clínico, como el resultado de pruebas de laboratorio. Facilitan el intercambio y puesta en común de los resultados para la atención clínica, los resultados, la gestión y la investigación. Como por ejemplo diferentes valores, niveles séricos, o signos vitales.

---

[17] http://www.nlm.nih.gov/mesh/trees.html
[18] http://www.nlm.nih.gov/mesh/MBrowser.html
[19] http://www.nlm.nih.gov/research/umls/rxnorm/
[20] http://www.fda.gov/cder/ndc/
[21] http://www.loinc.org/

Capítulo 3

3. Estado del arte

La revisión del estado del arte se ha centrado principalmente en los trabajos relacionados con la expansión de la consulta en el entorno de la recuperación de la información, profundizando en las diferentes técnicas y áreas aplicadas. Complementándolo posteriormente, con una amplia revisión y estudio de las diferentes métricas para el cálculo de la similitud semántica entre términos, aplicadas en el ámbito biomédico.

3.1. Introducción

Las necesidades de información por parte del usuario, se reflejan típicamente en los sistemas de recuperación de información, mediante la petición y aplicación de consultas. Es aquí donde se observa, la dificultad para obtener resultados relevantes (Rose, & Levinson, 2004). Estas consultas son generalmente incompletas o imprecisas, debido a la complejidad del lenguaje y la información almacenada en dichos sistemas. De manera que en ciertos momentos, las consultas no son de utilidad para recuperar la información correcta asociada a las necesidades de información. Estos sistemas podrán responder a la petición de las consultas, pero requerirán más información para completar el significado de dichas consultas.

Para resolver este problema y completar la información necesaria en la consulta, surge la aplicación de técnicas de expansión de la consulta. Esta técnica se basa en la incorporación de nuevos términos relacionados a la consulta original con el fin de completarla. De esta manera, los resultados obtenidos en la expansión de la consulta, mejorarán los obtenidos por la consulta original (Efthimiadis, 1996).

Los métodos para la expansión de la consulta se vienen estudiando desde tiempo atrás. Aunque los primeros resultados no aportaban mejoras sustanciales de importancia, en las investigaciones más recientes, la mayoría de los investigadores han alcanzado el consenso de su utilidad actualmente. Nuevas variantes, pueden usarse de manera consistente en la mejora de recuperación de información en colecciones generales. Es una técnica poco explorada por los sistemas comerciales, haciendo poco uso de sus ventajas.

Aunque la expansión de la consulta está asociada a la idea de adición de términos, existen otras operaciones que se pueden incluir para mejorar la búsqueda, como son la eliminación de términos, sustitución de términos e incluso sustitución de operadores (Moldovan & Mihalcea, 2000) (Moldovan & Mihalcea, 2002). Un caso especial de expansión de la consulta es la reformulación de la consulta, centrada en el reemplazamiento de términos originales con otros términos similares o relacionados (Jones, Rey, Madani & Greiner, 2006).

Las técnicas para la aplicación de la expansión de la consulta cubierta en la literatura, se pueden clasificar desde un amplio punto de vista, en dos principales aproximaciones (Figura 3.1):

- **Técnicas probabilísticas sobre el corpus de búsqueda**

Las aproximaciones para la mejora de expansión de la consulta dentro de este amplio grupo, se podrían dividir a su vez, según (Baeza-Yates, & Ribeiro-Neto, 1999), en tres categorías:

  - La primera, agrupando las aproximaciones para la obtención de los términos para la expansión, en función de la retroalimentación de información aportada por el usuario (***Relevance Feedback***), en la selección de aquellos documentos definidos como relevantes por el usuario del total de documentos recuperados;
  - la segunda categoría, se basa en las aproximaciones para la extracción de términos aplicables a la expansión de la consulta, a partir del análisis del conjunto de documentos recuperados inicialmente como respuesta a una consulta "*q*", denominado como **análisis local de documentos**;
  - la tercera categoría, se centra en la obtención de términos aplicables, mediante procesos de análisis y extracción, a partir de toda la colección del corpus de documentos a consultar, denominándose como **análisis global de documentos**.

- **Técnicas de análisis de colecciones o estructuras de conocimiento**

Por otro lado, las técnicas de más reciente aplicación, se ha definido como **análisis de colecciones**, se basan en **el análisis de estructuras del conocimiento**. Son un enfoque alternativo, en el que se aplica o utiliza el conocimiento almacenado en dichas estructuras, para seleccionar e incorporar los términos más adecuados que mejoren la expansión de la consulta. Estas estructuras u ontologías, pueden ser "**independientes**" o de uso en contextos generales (como la lengua inglesa), por lo que no están relacionadas con un dominio especializado, como es el caso de WordNet[22]. Mientras que otras ontologías son "**dependientes**" o relacionadas con un dominio de aplicación especializado, como MeSH[23], SNOMED CT[24] o UMLS[25] [26].

[22] http://wordnet.princeton.edu/
[23] http://www.nlm.nih.gov/mesh/
[24] http://www.ihtsdo.org/snomed-ct
[25] http://www.nlm.nih.gov/research/umls/
[26] https://uts.nlm.nih.gov/home.html

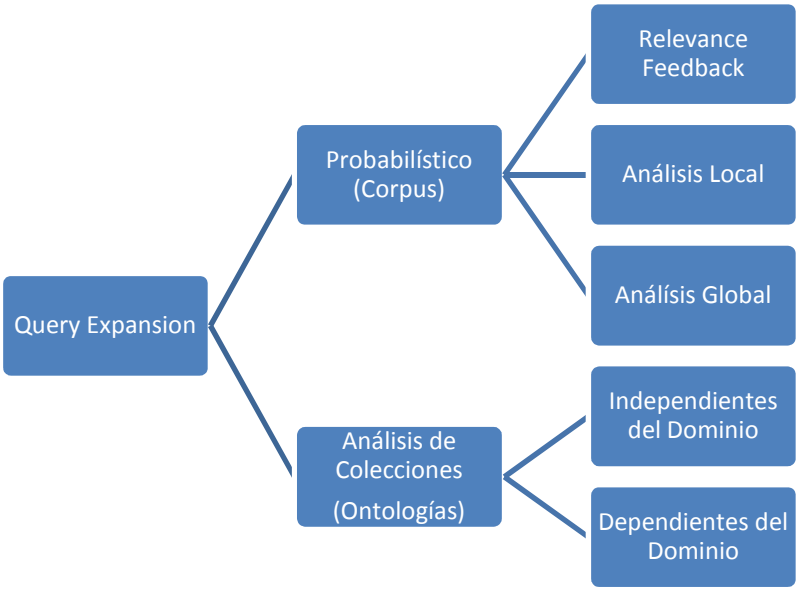

Figura 3.1 Esquema resumen de los ámbitos aplicados a la expansión de la consulta.

3.2. Técnicas de Recuperación de Información basadas en Corpus

Este conjunto de técnicas fueron inicialmente las más extendidas, centrándose en la expansión de la consulta basada en **técnicas de análisis probabilístico sobre los corpus de búsqueda.** Estas técnicas se orientan fundamentalmente en la co-ocurrencia de términos en los documentos de búsqueda, para seleccionar así, los de mayor relación con el término de la consulta inicial. Una buena revisión de las técnicas probabilísticas se puede encontrar en la introducción de (Xu, & Croft, 2000), así como en los trabajos relacionados de (Cui, Wen, Nie, & Ma, 2002) y la segunda sección de (Carpineto, Mori, Romano, & Bigi, 2001)

3.2.1. Expansión basada en Retroalimentación

La expansión basada en la retroalimentación del usuario (*Relevance Feedback*), es una de las primeras y más populares estrategias de expansión de la consulta. El ciclo de retroalimentación se basa en un proceso semiautomático, en el que se examinan y seleccionan por el usuario, los documentos más relevantes recuperados por la herramienta dentro del conjunto de documentos total consultados (Efthimiadis, 2000). En la práctica, el usuario hace una división de documentos relevantes y no relevantes de entre los primeros 10 o 20 resultados obtenidos por el sistema. La idea fundamental consiste en obtener aquellos términos, expresiones o definiciones incluidas en el conjunto de documentos identificados como relevantes por el usuario, mejorando la reformulación de la consulta con dichos términos. De esta manera, los resultados obtenidos por la nueva consulta expandida, son generados por el efecto de dirigir la consulta hacia los documentos relevantes y apartarla de los no relevantes.

Algunos de los primeros experimentos (Salton, 1971), (Salton, 1990), (Robertson, 1976), demostraron la mejora en la precisión en los documentos recuperados cuando se usa la retroalimentación en pequeños conjuntos. Este proceso tiene la ventaja de que es el

usuario, quien juzga y controla la relevancia de los documentos, poniendo mayor énfasis en los relevantes y menor en los no relevantes.

En diferentes trabajos se ha discutido el uso de la retroalimentación desde la aplicación del modelo vectorial y el modelo probabilístico (Selberg, 1997).

La aplicación del modelo vectorial sobre términos ponderados, se basa en el algoritmo de Rocchio (Rocchio1971), siendo uno de los más conocidos en la expansión de la consulta basada en los resultados de búsqueda. La idea principal consiste en aplicar una reformulación de la consulta donde se aproxime a los términos ponderados del vector espacio de los documentos relevantes. De forma, que el objetivo sea obtener un vector definido como Q que maximice la similitud con los documentos relevantes, mientras minimiza la similitud con los documentos no relevantes. Si R es el conjunto de documentos relevante y S el de no relevantes, la consulta óptima sería:

$$Q_{opt} = max_Q\ [sim(Q,R) - min\ (Q,S)\ ] \tag{3.1}$$

Por otro lado, (Robertson, 1976) inicio el estudio sobre la ponderación de términos basados en el modelo probabilístico clásico. El modelo probabilístico, clasifica los documentos similares a la consulta q, de acuerdo con los principios de clasificación probabilística. De manera que la similitud *sim(d_j,q)* de un documento d_j a una consulta q, se define como:

$$sim(d_j, q) = \frac{P(R/\vec{d}_j)}{P(\bar{R}/\vec{d}_j)} \tag{3.2}$$

Siendo R el conjunto de documentos relevantes y $\bar{R}$ el conjunto de documentos no relevantes, entonces $P(R/\vec{d}_j)$ es la probabilidad de que un documento d_j sea relevante a las consulta q y $P(\bar{R}/\vec{d}_j)$ sea la probabilidad de que un documento d_j sea no relevante a la consulta q. En este modelo basado en la retroalimentación del usuario, las probabilidades se conocerán a partir de las estimaciones del usuario.

3.2.2. Expansión de la consulta mediante análisis local

Siguiendo la idea básica, de la expansión de la consulta basada en la retroalimentación. El análisis local se fundamenta en el concepto de expansión de la consulta a partir del análisis del conjunto de documentos principales recuperados a partir de una consulta inicial. Esta aproximación pretende obtener una selección de los documentos relevantes de forma automática. Esto implica la identificación dentro del conjunto inicial de documentos principales recuperados, de aquellos términos que están relacionados concurrentemente con el término de la consulta inicial "*q*". Estos términos identificados, pueden relacionarse con el término inicial, mediante relaciones de sinónima, o por formas derivadas (*stemming*), o por proximidad en el cuerpo del documento. Dichos términos relacionados y

extraídos del conjunto de documentos relevantes a partir de una consulta "*q"*, son los que se utilizarán para la expansión de la nueva consulta basada en los resultados locales de "*q"*.

Existen dos estrategias básicas de análisis local:

- La primera estrategia (Attar, 1997), se basa en el agrupamiento o ***clustering* de términos relacionados** que han sido seleccionados del conjunto de documentos (top-n) obtenidos de la consulta inicial "*q"*, para posteriormente ser aplicados en la expansión de la consulta. Diferentes métodos se han propuesto para la selección del conjunto de términos concurrentes extraídos del conjunto inicial (top-n) de documentos, así como para minimizar el efecto de los documentos irrelevantes devueltos en el conjunto inicial (Mitra, Singhal, & Buckley, 1998).
- La segunda estrategia, más reciente y con mejores resultados (Xu, & Croft, 1996), (Xu, & Croft, 2000), se basa en el **análisis del contexto local**, mostrando las ventajas de la combinación de las técnicas de análisis global y local. Esta aproximación se basa en el uso de grupos de palabras simples o adyacentes, en vez de una palabra clave como concepto del documento. Para la expansión de la consulta, los conceptos son seleccionados del conjunto más relevante de documentos, en función de su coocurrencia con los términos de la consulta, como aplicación del análisis local. Sin embargo, en vez de documentos, son pasajes o ventanas de texto las que son usadas para determinar la coocurrencia, como aplicación del análisis global. En resumen, esta aproximación incluye la recuperación y búsqueda por pasajes y no documentos. Así como, la comparación de similitud de la totalidad de los términos de la consulta como candidatos para su expansión, en vez de individualmente para cada término de la consulta candidata.

3.2.3. Expansión de la consulta mediante análisis global

Si bien, en el caso anterior la expansión de la consulta se centraba en la extracción de información del conjunto de documentos recuperados. Esta nueva aproximación basada en el análisis global, son técnicas que buscan dentro del corpus o colección de documentos, las relaciones que permitan extraer la información necesaria para aplicar a la expansión de la consulta. Así por ejemplo, ciertas técnicas se fundamentan en el agrupamiento mediante vectores de co-ocurrencia (Xu, Zhu, Zhang, Hu, & Song, 2006; Zhu, Wu, Carterette, & Liu, 2014) y de modelos probabilísticos a partir de los términos más relevantes (Qi, & Laquerre, 2012).

Otras técnicas aplicadas al análisis global, se centran en un tesauro de similitud (Qiu, & Frei, 1993). Dicho tesauro es construido considerando las relaciones entre términos, estas relaciones no derivan directamente de la coocurrencia de los términos dentro de los documentos. Es decir, que el tesauro no se deriva de una matriz de coocurrencia como en casos previos, sino que surgen del corpus de documentos, en la que se usan medidas de distancia entre los términos del espacio conceptual. De forma innovadora, en este espacio conceptual, cada término es indexado por los documentos en los que aparece. De esta manera, los términos asumen el papel original de los documentos, mientras que los documentos son interpretados como elementos de indexación.

Esta técnica expandirá la consulta inicial, mediante la elección de términos cercanos conceptualmente dentro de dicho espacio conceptual, con respecto al centroide de la totalidad de la consulta. En vez, de por proximidad de términos con respecto a cada término de la consulta.

Otra técnica asociada a este entorno, relaciona los términos con la agrupación de documentos en conjuntos estrechos de agrupamiento (Crouch, & Yang, 1992), para luego elegir los términos de menor frecuencia y por tanto con alta discriminación en los grupos. Debido a la dificultad de poder realizar un agrupamiento por baja frecuencia de términos, ya que conforman un pequeño grupo. Este problema se resolverá, mediante el agrupamiento de los documentos en clases en vez de por baja frecuencia de términos en los documentos. Definiendo un tesauro de clases, las cuales serán pequeñas y ajustadas.

De forma esquemática, el algoritmo de agrupamiento funcionaria de la siguiente manera:

- Inicialmente, se sitúa cada documento en un grupo (*cluster*).
- Se comprueba la similitud entre pares de *clusters*.
- Se determina que pares de *clusters* tienen entre si la mayor similitud
- Se combinan los pares de *clusters* con mayor similitud.
- El proceso se repetirá hasta llegar a un criterio de parado en la similitud entre *clusters*.

Otras técnicas bien conocidas y aplicadas en el análisis global, son el índice latente semántico (*Latent Semantic Index*) (Deerwester, et al., 1990), también denominado análisis latente semántico. Estas técnicas aplican diferentes métodos para la creación de una matriz de similitud de términos que permita seleccionar aquellos con mayor frecuencia a los términos de la consulta inicial.

Trabajos más recientes se fundamentan en la aplicación de modelos de distribución semánticos sobre elementos lingüísticos existentes en dichas colecciones, con el fin de extraer sinónimos y abreviaturas de forma automática (Henriksson, Moen, Skeppstedt, Daudaravicius, & Duneld, 2014; Zeng, Redd, Rindflesch, & Nebeker, 2012). Por último, otros trabajos dentro de este grupo analizan la aplicación de especificas métricas de similitud semántica sobre estructuras previamente definidas como contenedoras de los elementos a evaluar semánticamente, tales como el *Vector Space Model* (Turney, & Pantel, 2010) y la comparación de distancias entre histogramas o "*cross-bin distances*" (Kurtz, Beaulieu, Napel, & Rubin, 2014).

3.3. Técnicas de Recuperación de Información basadas en Colecciones

Otras estrategias aplicadas a la expansión de la consulta, son aquellas que se apoyan en recursos externos como estructuras de conocimiento del entorno que representan. Las cuales pueden ser definidas a partir de diccionarios, hasta complejas estructuras de conocimiento como son los tesauros u ontologías.

Se podría decir que ciertos tesauros de generación automática, como los obtenidos en las técnicas de expansión mediante el análisis global, comentadas en el punto anterior. Dan lugar a un conjunto de términos relacionados, extraídos a partir de la colección de documentos y basada en su propia estructura (Mandala, Tokunaga, & Tanaka, 2000). Mientras que las ontologías, son estructuras externas donde la asociación entre términos refleja las relaciones semánticas existentes entre ellos. Siendo utilizadas como recursos de apoyo en la recuperación de información, aportando la representación del conocimiento de un dominio específico (Gruber, 1993).

La aplicación de ontologías como técnica para la expansión de la consulta, se viene estudiando a partir de los primeros trabajos relacionados con esta idea (Voorhees, 1994). Aquí se describe la utilización de ontologías en la expansión de la consulta que posteriormente será adaptada en nuevos trabajos. Para ello, inicialmente los términos de la consulta deben ser desambiguados de forma que apunten a un único concepto en la ontología. De esta forma, los términos relacionados en la ontología con el término desambiguado de la consulta, serán añadidos a la expansión de la consulta.

Según (Bhogal, Macfarlane, & Smith, 2007), se han de cumplir tres aspectos para asegurar la efectividad de una ontología a la hora de apoyarse en ella, como técnica para la expansión de la consulta.

- Calidad del modelo de conocimiento: La ontologías deben representar el dominio de conocimiento al que hacen referencia, de una manera precisa, estable, completa y actualizada. Esto es necesario para representar la información de la forma menos ambigua, en cada tarea donde sea necesario aplicar el uso de ontologías.
- Familiaridad del modelo de conocimiento: Si los usuarios conocen el modelo de conocimiento representado en la ontología o son capaces de poder manejar la ontología de forma simple, las posibilidades de expansión serán mayores y de mejor calidad.
- Navegabilidad del modelo de conocimiento: Es necesario que el modelo que represente la ontología del dominio sea de fácil acceso. De manera que sea sencillo navegar en la ontología, donde el modelo de conocimiento es representado. De esta manera, cualquier tarea para la obtención de los términos más relevantes asociados a un concepto será capaz de realizarse de manera eficiente tanto por el usuario, como por un sistema automatizado.

Las ontologías, se podrían agrupar en dos conjuntos básicos.

- El primer grupo se podría incluir aquellas ontologías que se definen como **independientes del dominio de conocimiento** o de uso general. Algunas de estas ontologías son, WordNet[27] o EuroWordnet[28]. WordNet es una gran base de datos léxica del idioma inglés (Miller, 1995), (Fellbaum, 1998). Donde distintos elementos del lenguaje como, sustantivos, adjetivos, verbos, adverbios se agrupan en conjuntos de sinónimos, denominados *synsets*, representando un concepto. Los distintos *synsets* se relacionan ente si, de forma que muestran las relaciones semánticas y léxicas del lenguaje. Estas estructuras de representación del lenguaje son de gran utilidad como herramienta computacional, en el procesamiento del lenguaje natural (Hotho, Staab, & Stumme, 2003), en la indexación de documentos (Gonzalo, Verdejo, Chugur, & Cigarran, 1998) y en la expansión de la consulta (Hsu, Tsai, & Chen, 2006). EuroWornet es una gran base de datos multilingüe formada por distintos WordNets de distintas lenguas europeas (Holandés, Italiano, Español, Alemán, Francés, Checo y Estonio), siguiendo la estructura basada en *synsets*. Estos WordNets están enlazados mediante un índice (ILI) Índice Inter-Lingual, que permite la traducción de términos de un lenguaje a otro.

- Mientras que un segundo grupo, lo integran aquellas ontologías **dependientes de un dominio específico o especializado de conocimiento**. Como pueden ser el entorno jurídico (Benjamins, et al., 2004) o el biomédico (Aronson, Rindflesch, & Browne, 1994). Estas ontologías pretenden representar y permitir el acceso al conocimiento de un dominio, mediante relaciones semánticas entre los conceptos de dicho dominio.

Uno de los proyectos más representativos y que abordan esta problemática desde el punto de vista biomédico, es el Metatesauro UMLS[29] (The Unified Medical Language System). Este proyecto trata de resolver el acceso al conocimiento del domino biomédico tanto para usuarios, como sistemas automatizados (McCray, et al., 1993). Consiguiendo que a pesar de la dificultad del dominio, la gran variedad de accesos y tipos de información contenida. Se pueda interactuar con las distintas bases de datos, vocabularios controlados y sistemas de codificación que integran el Metatesauro, para su aplicación como herramienta de extracción y representación del conocimiento de dicho dominio.

UMLS se ha mostrado como un recurso de gran utilidad en las tareas de recuperación de información mediante su representación de estructuras conceptuales. La variedad y profundidad del conocimiento biomédico expresado por este recurso, ofrece una base sólida para la experimentación en el tratamiento de la información y el lenguaje natural en este contexto.

[27] http://wordnet.princeton.edu
[28] http://www.illc.uva.nl/EuroWordNet/
[29] http://www.nlm.nih.gov/research/umls/

3.3.1. Expansión de la consulta aplicada al dominio biomédico

Tras la revisión del estado del arte relacionado con las diferentes estrategias de recuperación de información basada en la expansión de la consulta y otras técnicas asociadas, se puede observar que los avances obtenidos en las últimas décadas han sido significativos. Aunque si bien hay que decir que el problema en la extracción de los documentos más relevantes con respecto a una consulta imprecisa o compleja no está completamente resuelto. Especialmente, en el dominio médico y más concretamente en el contexto relacionado con la información biomédica. La recuperación de información en este entono, se hace más difícil de resolver debido a la complejidad de la información tratada y de la ambigüedad de la terminología médica (Roth, & Hole, 2000).

Cada vez los recursos de información en el ámbito médico son mayores y más numerosos y las necesidades de información en dicho ámbito son más importantes en cuanto cantidad y calidad de información requería.

Actualmente diferentes estudios han demostrado la eficacia de los sistemas de recuperación de información, sin embargo en los sistemas centrados en el dominio biomédico, los resultados no son tan impactantes debido a la dificultad de la información manejada y los requisitos o necesidades de información específicas de los usuarios. Especialmente en cuanto al tratamiento de información clínica y diagnóstica se refiere, a partir de registros electrónicos médicos (Natarajan, et al., 2009).

En el entorno médico y sanitario, la utilización de registros médicos electrónicos o textos clínicos electrónicos, es cada vez mayor y dan lugar a un gran conjunto de información de características específicas y muy diferentes. Dichos registros o textos biomédicos, centran la información en la descripción de los pacientes, enfermedades, patologías, síntomas, medicamentos, vacunaciones, historias médicas, etc. (Meystre, Huang, & Gu, 2008). Junto con estas características y el hecho de que los textos clínicos sean de difícil acceso, en contraste con la literatura biomédica, da lugar a que estos corpus de información sean pequeños y poco frecuentes en la investigación relacionada con la RI.

Las colecciones médicas son conjuntos de información que presenta otros tipos de retos diferentes a la literatura biomédica, debido a las a las características diferenciadoras entre ellos, como por ejemplo:

- Los textos biomédicos, no se componen de construcciones gramaticales completas y correctas, sino de composiciones cortas, telegráficas y descriptivas. Donde la gramática no es de importancia con respecto al contenido clínico descriptivo de la información, permitiendo una comunicación clara y sencilla.

- Los textos biomédicos están repletos de abreviaturas y acrónimos, utilizados de forma frecuente y de una manera no estandarizada, pero que permite la representación de conceptos clínicos en los documentos de una forma breve y sencilla. Según (Liu, Lussier, & Friedman, 2001), se estima que pueden estar sobrecargados en un 33% de las veces y en muchos casos son altamente ambiguos.

- Pueden aparecer faltas de ortografía o errores gramaticales, especialmente en textos de apoyo clínico. Igualmente se pueden encontrar abreviaturas o acrónimos con errores.

- Estos textos médicos se incluye información no textual, como valores clínicos de laboratorio o constantes vitales. Así como la descripción de conceptos mediante el uso de clasificaciones o codificaciones médicas.

- La información suele estar localizada en estructuras dentro del documento médico, siendo identificables dentro del mismo. Igualmente la información o el lenguaje utilizado en la descripción de los conceptos médicos, está aplicada desde lenguajes estructurados y definidos.

Debido a estas características especiales, nuevas estrategias en la expansión de la consulta pueden ser necesarias en su aplicación. Las técnicas para la expansión de la consulta sobre documentación biomédica no tienen por qué ser poco útiles en este dominio, pero sí que deberán ser adaptadas, modificadas y mejoradas.

La expansión de la consulta aplicada al dominio médico, no es una tarea nueva. Pero si es cierto que la mayor parte de las investigaciones realizadas se han llevado a cabo en el tratamiento de colecciones de literatura biomédica, compuesta por recursos como artículos o *abstracts* localizados en la base de datos MEDLINE. Diferentes estudios (Lu, Kim, & Wilbur, 2009), (Díaz-Galiano, Martín-Valdivia, & Ureña-López, 2009) y (Shin, Han, Gelbukh, & Park, 2004), han demostrado la efectividad de los resultados obtenidos sobre dichas colecciones de testeo, mediante la aplicación de términos MeSH en la expansión e indexación de documentos.

Así por ejemplo, uno de los primeros trabajos (Srinivasan, 1996), muestra la efectividad en distintas estrategias de expansión de la consulta aplicada sobre MEDLINE, utilizando el sistema de recuperación SMART de la Universidad de Cornell. Posteriormente, Aronson y Rindflesch, en diferentes estudios aportan ciertas mejoras a los anteriores resultados, mediante el apoyo del Metatesauro UMLS, aplicado a las diferentes estrategias de expansión. Los primeros trabajos (Rindflesch, & Aronson, 1994), se centraron en la resolución de la desambiguación en el enlazado o "mapeo" del texto libre con los conceptos relacionados en el Metatesauro. Para posteriormente (Aronson, & Rindflesch, 1997), estudiar la asociación de términos del Metatesauro UMLS con la consulta inicial del usuario para su expansión, mediante la aplicación del programa MetaMap. En dichos trabajos, hicieron uso del sistema de recuperación de información INQUERY.

En los estudios realizados por Srinivasan y posteriormente por Aronson, se muestran las mejoras en los resultados obtenidos con la utilización de los términos MeSH asignados en la indexación de los documentos de la colección y la expansión de la consulta basada en tesauros estadísticos y en el Metatesauro UMLS, respectivamente.

Posteriormente, Hersh et al. (Hersh, Price & Donohoe 2000) realizan una evaluación de la expansión de la consulta basada en tesauros, utilizando UMLS. Aplicando en la expansión de la consulta, tres tipos de relaciones existentes en los tesauros:

- Relaciones de Sinonimia (*Synonym*).
- Relaciones de Jerarquía (*Hierarchical*).
- Relaciones de Afinidad (*Related*).

En los resultados obtenidos de las pruebas realizadas, se observó que las expansiones con relaciones jerárquicas no mejoraban el rendimiento de la consulta, mientras que la expansión por palabra o término mejoraba notablemente los resultados de la consulta. Según los autores, existen ciertas limitaciones en dicho estudio que justifica una mayor investigación. Una de las limitaciones se centra en la colección de testeo, ya que se basan en un único tipo de documento, apoyado en referencias MEDLINE con terminología MeSH para su indexación. Por lo tanto, otras colecciones con otras características en su contenido, como texto completo, pueden ofrecer diferentes resultados con respecto a las bases de datos bibliográficas como MEDLINE. Otra de las limitaciones, es que la evaluación se ha realizado con un único sistema de recuperación de información, en este caso un sistema estadístico de palabras.

Posteriores líneas de investigación, continúan estos trabajos, tratando de evaluar y obtener mejoras en la búsqueda de documentos relevantes dentro de colecciones de literatura biomédica (Zhu, et al., 2006) y (Xu, et al., 2006). Estos trabajos evalúan y comparan los resultados obtenidos mediante la aplicación de tres estrategias de expansión: El análisis local, el análisis global y la ponderación de términos basados en ontologías, a través de los motores de búsqueda LUCENE[30] y LEMUR[31]. Donde se observa que los mejores resultados los ofrece la estrategia de expansión basada en términos ponderados de ontologías. Para ello, se reformula la consulta con una selección de términos claves originales, junto con un conjunto de términos ponderados extraídos de los sinónimos asociados al termino original en UMLS. Los experimentos demuestran que la precisión media mejoran en un 20.3% y un 12.1%, con LUCENE y LEMUR respectivamente.

Por ello, el uso y análisis de estructuras basadas en el conocimiento del dominio biomédico, como es UMLS, ofrece un gran interés en el contexto de la biomedicina. Sin embargo, la aplicación de estos recursos como apoyo en expansión de la consulta, requiere la desambiguación de los términos originales de la consulta, de forma que estos apunten a conceptos únicos en la ontología (Bhogal, Macfarlane, & Smith, 2007; Voorhees, 1994). De manera que los conceptos relacionados con el término inicial de la consulta podrán ser incluidos en su expansión.

Una herramienta que junto con el Metatesauro UMLS permite la identificación de conceptos referidos en un texto, es MetaMap (Aronson, 2001; Aronson, & Lang, 2010). Esta herramienta proporciona la base para el desarrollo de diferentes estrategias de la expansión de la consulta (Aronson, & Rindflesch, 1997), mediante la explotación de la semántica contenida en UMLS. De esta manera, diferentes trabajos usan las estructuras definidas en UMLS, para obtener soluciones fundamentadas por ejemplo en: la representación del texto de documentos médicos mediante grafos semánticos basados en conceptos y relaciones (Plaza, & Díaz, 2010); la expansión de una consulta mediante "*random walks*" basados en la estructura de UMLS (Martínez, Otegi, Soroa, & Agirre, 2014);

[30] http://lucene.apache.org/
[31] http://www.lemurproject.org/

la ampliación de una consulta, mediante la creación de una ontología de la propia consulta, asociada a conceptos estrechamente relacionados (Babashzadeh, Huang, & Daoud, 2013); o el uso de las relaciones entre conceptos, para reflejar la distancia semántica entre pacientes a partir de su información almacenada (Melton, Parsons, Morrison, Rothschild, Markatou, & Hripcsak, 2006).

Como consecuencia de las necesidades de mejora en la explotación de la semántica contenida en los diferentes recursos del dominio biomédico, surgen diversos trabajos enfocados en valuar la similitud semántica entre conceptos del dominio biomédico.

3.4. Similitud Semántica basada en Ontologías

La búsqueda de la relación entre términos, es una tarea que tiene sus inicios en trabajos previos centrados en la psicolingüística (Collins, & Loftus, 1975). Estos trabajos centraron sus esfuerzos en la identificación y estudio de dos tipos de relaciones, las relaciones asociativas y las semánticas. Las primeras, hacen referencia a la probabilidad de que una palabra trajera a la mente otra asociada a ésta, por ejemplo "aguja-hilo". Mientras que la segunda relación reflejaba el grado de solapamiento semántico entre palabras, por ejemplo "ballena–delfín". Otros estudios basados en esta problemática, demuestran que los humanos tienen una alta concordancia en el enjuiciamiento del grado de relación que puede existir entre dos términos (Rubenstein, & Goodenough, 1965; Miller, & Charles, 1991).

Dentro de las relaciones semánticas entre términos, es necesario diferenciar entre "similitud" y "relación" semántica. Según (Nguyen & Al-Mubaid, 2006), dos términos semánticamente similares son aquellos cuyo significado semántico es aproximado, mientras que dos términos relacionados pueden no ser semánticamente similares. Así por ejemplo, "coche" y "conductor" están relacionados, pero no son semánticamente similares, mientras que "coche" y "vehículo" si lo son. Por tanto, se puede decir que la relación semántica entre conceptos es una noción más amplia que la similitud semántica, ya que ésta se establece por la semejanza entre conceptos. Es por ello que la similitud semántica se entiende de forma generalizada, como el grado de proximidad taxonómica entre dos términos.

Siguiendo estas ideas, diferentes trabajos han tratado de reflejar el interés por medir la similitud y las relaciones semánticas entre dos términos dentro de **dominios independientes o genéricos**, como en el procesamiento del lenguaje natural en la recuperación de información y/o documentos (Resnik, 1995), (Lin, 1998), (Jiang, & Conrath, 1997), (Patwardhan, & Perdersen, 2006). Muchos de estos trabajos se han basado en el tratamiento de conceptos de la lengua inglesa, mediante el uso de WordNet (Fellbaum, 1998).

Los trabajos anteriores, asociados al procesamiento del lenguaje natural en dominios genéricos, pierden su eficacia en **dominios especializados**, como el de la biomedicina (Burgun, & Bodenreider, 2001). Sin embargo existen un gran número de recursos, como vocabularios controlados, ontologías, clasificaciones, codificaciones, (algunos de ellos son: SNOMED-CT, MeSH, ICD-9, etc.) en continua actualización que aportan el conocimiento del

domino biomédico permitiendo su aplicación en procesos automatizados. Todos ellos se recogen e interrelacionan en el Metatesauro UMLS[32] (Unified Medical Language System).

Estos recursos, pueden mejorar los sistemas de recuperación de información basados en documentos médicos, mediante la aplicación de sinónimos o palabras similares en la optimización de las consultas. Así por ejemplo, la consulta o recuperación basada en historias clínicas electrónicas para controles epidemiológicos de la población, tienen el objetivo de extraer aquellos pacientes asociados una patología o síndrome. Para ello requieren el uso de una gran variedad de términos que no solo se reflejan en los diagnósticos, sino también en síntomas, tratamientos, enfermedades y diferentes conceptos que pueden estar estrechamente relacionados o ser similares con la enfermedad o síndrome original.

Por tanto, para establecer la similitud semántica entre dos conceptos (Pedersen, Pakhomov, Patwardhan, & Chute, 2007), es necesaria la aplicación de diferentes métricas, basadas en el uso de recursos del dominio biomédico. Dichas métricas, toman como entrada dos conceptos, para devolver un único valor dentro de un rango definido que representa como de similares o parecidos son entre sí. Estos valores deberán ser comparados con el grado de similitud semántica definido por los expertos para ambos conceptos y así poder establecer el grado de concordancia de los resultados obtenidos por las diferentes métricas. Típicamente las métricas, se basan en las relaciones jerárquicas taxonómicas de tipo "*is-a*" entre términos, para el cálculo de la similitud semántica. Existen otros tipos de relaciones jerárquicas, que no han sido ampliamente exploradas en dichas métricas.

3.4.1. Métricas de Similitud Semántica en el dominio biomédico

Desde un punto de vista semántico es posible determinar un grado de similitud entre dos conceptos, basándose en recursos definidos como estructuras de representación del conocimiento. A partir de dichos recursos, como taxonomías y ontologías basados en el conocimiento del domino, se puede cuantificar como de parecidos son dos conceptos por su proximidad jerárquica en la taxonomía. Estos recursos son considerados como un grafo en el que se representan y modelan las relaciones semánticas entre conceptos.

Antes de revisar las diferentes métricas definidas para establecer la similitud semántica entre términos, conviene recordar la diferencia entre similitud semántica y relación semántica. Cuando dos conceptos se relacionan a partir de la semejanza entre ellos, se puede decir que son semánticamente similares. Mientras que dos conceptos semánticamente relacionados pueden no ser similares y tener una cierta relación que los asocia. De esta forma, se considera la relación semántica un concepto más amplio que la similitud semántica.

Típicamente las métricas para determinar la similitud semántica entre conceptos se basan en enlaces jerárquicos de tipo "*is-a*" que los relacionan de manera directa o indirecta. Obteniendo la medida de similitud a partir de los caminos entre conceptos o mediante la aplicación combinada del camino más corto junto a un corpus que aporte cierta información empírica. Mientras que las métricas para establecer la relación

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

semántica entre conceptos, incluyen información sobre otras relaciones más amplias y definiciones extendidas de los conceptos o se basan en información estadística de coocurrencia de términos a partir de un corpus.

- ***Métricas basadas en el camino entre conceptos***

Para la representación del conocimiento del domino, los conceptos son organizados mediante estructuras jerárquicas en forma de árbol invertido, como son las taxonomías u ontologías. Dichos conceptos son más genéricos cuanto más cercanos estén de la raíz de la taxonomía, mientras que son más específicos cuanto más próximos a las hojas de la jerarquía se sitúen. Junto con estas características, diferentes métricas obtienen una relación de similitud entre conceptos en función de la longitud del camino entre ambos términos. Algunas de estas métricas han sido estudiadas previamente en dominios independientes para el procesamiento del lenguaje natural (NLP) y han sido adaptadas al domino biomédico.

- **Conceptual Distance (CDist):**

Esta métrica, se basa en el cálculo del camino más corto entre dos conceptos. Para ello se contabiliza las distancias en función del número de nodos entre dos conceptos, incluyendo los extremos. Una primera métrica basada en esta idea, es la definida por Rada (Rada, Mili, Bicknell, & Blettner, 1989) en la que se calcula la similitud entre términos a partir de la relaciones de tipo RB/RN entre términos del vocabulario controlado MeSH (Medical Subject Headings). Posteriormente Caviedes y Cimino, desarrollan la métrica definida como CDist (Caviedes, & Cimino, 2004). Esta se basa en la aplicación del cálculo del camino más corto entre dos conceptos, pero aplicado a relaciones de tipo PAR/CHD, para un subconjunto de recursos del dominio médico contenidos en UMLS (estos son, MeSH, ICD-9-CM y SNOMED-CT). Estos enfoques consideran las ontologías como un grafo en el que las interrelaciones semánticas son modeladas como enlaces entre conceptos. Representando la similitud de conceptos como la distancia mínima de interrelación entre conceptos, en la estructura taxonómica (s*p*).

$$sim_{CDist}(c_1, c_2) = sp(c_1, c_2) \tag{3.3}$$

$$where\ "sp" - is\ the\ "shortest\ path"\ between\ c_1, c_2$$

El resultado obtenido por esta métrica es igual a 1, cuando se obtiene la mayor similitud semántica entre términos, es decir cuando ambos conceptos son el mismo. Mientras que cuanto menor es la similitud entre conceptos, mayor es el resultado obtenido por CDist ($sim_{cdist} \geq 1$). Dicho de otra forma, la similitud es menor, cuanto mayor es la distancia entre conceptos en la jerarquía.

- **Path Length:**

Es una variación normalizada de la métrica anterior, para el cálculo de la distancia mínima entre dos conceptos. Al contrario de *CDist*, la similitud es inversamente proporcional a la distancia entre ambos conceptos. Por lo tanto, los posibles valores que puede devolver esta métrica estarían entre $(0 \leq sim_{path}(c_1, c_2) \leq 1)$.

$$sim_{Path}(c_1, c_2) = 1/sp(c_1, c_2) \tag{3.4}$$

Pedersen et al., definen esta métrica para el cálculo de la similitud semántica entre conceptos a partir de relaciones "*is-a*" definidas en SNOMED-CT (Pedersen, Pakhomov, Patwardhan, & Chute, 2007).

- **Leacock y Chodorow:**

Leacock y Chodorow, proponen una métrica que considera el camino más corto *(sp)* entre dos conceptos *(c_1, c_2)*, escalándolo de forma constante con respecto a la profundidad *(Depth)* de la taxonomía en la que se produce (Leacock, & Chodorow, 1998). Una propuesta para normalizar esta métrica en el intervalo de la unidad se define en Garla (Garla, & Brandt, 2012).

$$sim_{lch}(c_1, c_2) = 1 - \log(sp)/\log 2Depth \tag{3.5}$$

En esta métrica, cuanto mayor es la profundidad de la taxonomía, es decir más compleja y completa es, mayor es el valor relativo de la similitud semántica entre dos conceptos. Por lo tanto, los casos que dan lugar a la mayor similitud semántica posible, serán aquellos en los que la profundidad de la taxonomía sea la más alta posible y la distancia entre conceptos la más corta.

Es interesante observar que en esta métrica, no se tiene en cuenta en que parte de la taxonomía se encuentra el camino más corto *(sp)* entre los conceptos *(c_1, c_2)*. Sin embargo, la similitud semántica entre ellos, puede verse afectada por el lugar en la que se encuentren dentro de la taxonomía (Figura 3.4). Por ejemplo, el camino más corto entre dos conceptos muy próximos a la raíz de la taxonomía, implica baja similitud semántica ya que son conceptos muy genéricos. Mientras que si el camino obtenido se encuentra entre conceptos muy próximos a las hojas de la taxonomía, es decir entre conceptos muy específicos, implica una similitud muy alta.

- **Wu y Palmer:**

Las ontologías han evolucionado con la intención de formalizar el conocimiento de un domino, mediante diferentes relaciones y descripciones lógicas entre conceptos. Éstas representan el conocimiento mediante jerarquías de subsunción. La métrica propuesta (Wu, & Palmer, 1994), trata de obtener la similitud entre dos términos a partir del

concepto antecesor común más próximo que los engloba (LCA - *Least Common Ancestor*), o también conocido como (LCS - *Least Common Subsumer*), respecto a la profundidad de ambos conceptos en la taxonomía a partir de su elemento común antecesor (Figura 3.4). Wu y Palmer aplicaron esta métrica sobre un dominio general e independiente como WordNet, para la lengua inglesa. Existen adaptaciones de esta métrica a dominios específicos como SNOMEDCT (Pedersen, Pakhomov, Patwardhan, & Chute, 2007).

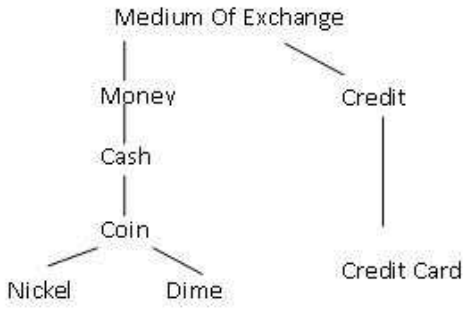

Figura 3.2 Ejemplo taxonomía WordNet.

Esta métrica de calcula de la siguiente forma: N_1 y N_2 son los caminos más cortos formados por enlaces "is-a" desde c_1 y c_2 respectivamente al LCS c, y N_3 es el camino más corto desde el LCS c a la raíz ρ de la ontología.

$$sim_{wup}(c_1, c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \tag{3.6}$$

Así para el ejemplo de la Figura 3.2, entre los conceptos c_1= "Nickel" y c_2="Dime", el LCS es "*Coin*". La similitud semántica en este caso, es mucho mayor que entre, c_1= "Nickel" y c_2="*Credit Card*", donde el LCS es "Medium of Exchange", el cual está a una distancia mucho más lejana de los conceptos subsumidos y por tanto la similitud semántica es menor. Por lo tanto, se puede decir que cuanto mayor es la distancia de los conceptos subsumidos (c_1, c_2) con respecto al LCS, menor es la similitud semántica entre ellos. Si ambos conceptos son el mismo y coinciden con el LCS, la similitud es 1. Los valores posibles para esta métrica están entre $(0 \leq sim_{wup}(c_1, c_2) \leq 1)$.

Otra forma de representar, esta métrica es:

$$sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs, root)}{depth(c_1, lcs) + depth(lcs, root) + depth(c_2, lcs) + depth(lcs, root)} \tag{3.7}$$

De manera que simplificando, se obtiene:

$$sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs, root)}{depth(c_1, root) + depth(c_2, root)} \tag{3.8}$$

En las dos métricas anteriores, el camino más corto entre dos conceptos es escalado de alguna forma con la profundidad de la taxonomía, para evitar la dependencia absoluta de la métrica con respecto a la longitud del camino más corto. Pero no se identifica en qué lugar dentro de la taxonomía, se encuentran estos caminos más cortos, lo cual puede afectar a la similitud semántica como se comentó anteriormente.

Una variación normalizada es introducida en (Garla, & Brandt, 2012), incluyendo el camino más corto entre dos conceptos *(sp)*, para evitar que en el caso de que *(c_1=c_2)*, el cual resultaría con una similitud menor a 1.

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth\big(LCS(c_1, c_2)\big)}{sp(c_1, c_2) - 1 + \ 2 \ \times \ depth\big(LCS\ (c_1, c_2)\big)} \tag{3.9}$$

- **Nguyen y Al-Mubaid:**

Esta métrica pretende resolver el inconveniente de los dos casos anteriores. Ya que junto con la importancia del camino más corto entre dos conceptos, es relevante la localización de dicho camino dentro de profundidad de la taxonomía. Ya que a mayor profundidad se encuentre éste, mayor relevancia tendrá la similitud entre conceptos al ser estos más específicos.

Para ello, Nguyen y Al-Mubaid definen una nueva métrica (Nguyen & Al-Mubaid, 2006), que incorpora la importancia de la localización del camino más corto entre conceptos, dentro de la taxonomía.

$$sim_{nam}(c_1, c_2) = \log_2\Big((sp(c_1, c_2) - 1) * \big(Depth - depth(lcs, root)\big) + 2\Big) \tag{3.10}$$

Analizando esta métrica podemos observar que el camino más corto entre dos conceptos $(sp(c_1, c_2) - 1)$, se ve afectado por la profundidad, en la que se encuentre el LCS de ambos conceptos en la taxonomía $\big(Depth - depth(lcs, root)\big)$. Por tanto, la mayor similitud semántica entre conceptos se obtiene, cuando el LCS se sitúa en la parte más próxima a las "hojas" de la taxonomía y cuanto más corto es el camino entre los dos conceptos. Obteniendo una mayor similitud semántica cuanto menor es el resultado obtenido por la métrica y más se aproxime a 1.

Por tanto, los valores posibles para esta métrica son $(sim_{nam}(c_1, c_2) \geq 1)$. De manera, que cuanto mayor es el resultado obtenido por la métrica, menor es la similitud semántica entre conceptos y viceversa.

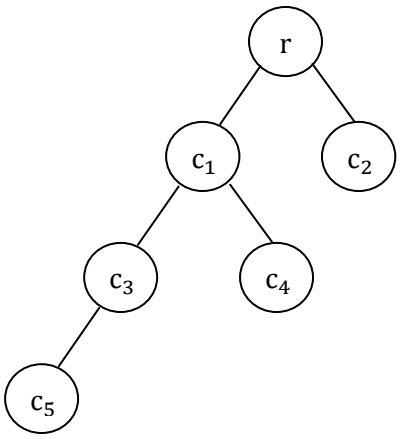

Figura 3.3 Jerarquía entre conceptos.

Reflejándolo en el ejemplo de la Figura 3.3, para las métricas anteriores, basadas en el camino más corto (*path, lch* y *wup*), la similitud semántica entre $sim(c_1, c_2)$ y $sim(c_3, c_4)$ es la misma. Sin embargo en (*nam*), la profundidad de localización de estos conceptos dentro de la jerarquía de la taxonomía, determina que la similitud de $sim(c_3, c_4)$ es mayor que la de $sim(c_1, c_2)$. Esto es debido a que los conceptos c_3 y c_4 son más específicos y comparten más información que los conceptos c_1 y c_2 que están por encima de ellos en la taxonomía y más próximos a la raíz.

$$sim_{nam}(c_1, c_2) = \big(log(2 + sp(c_1, c_2) - 1)\big) \times (D - depth(LCS(c_1, c_2))) \quad (3.11)$$

Esta métrica y las anteriores han sido adaptadas e implementadas en un *framework open-source* basado en un conjunto de módulos Perl denominado UMLS-Similarity[33] (McInnes, Pedersen, & Pakhomov, 2009; McInnes & Pedersen 2013). Para el cálculo de las distintas métricas sobre un *benchmark* de pares de conceptos (Pedersen, Pakhomov, Patwardhan, & Chute, 2007) definidos por de expertos, con el objeto de evaluar la proximidad de la distintas métricas en el cálculo de la similitud semántica entre términos.

[33] http://search.cpan.org/dist/UMLS-Similarity/

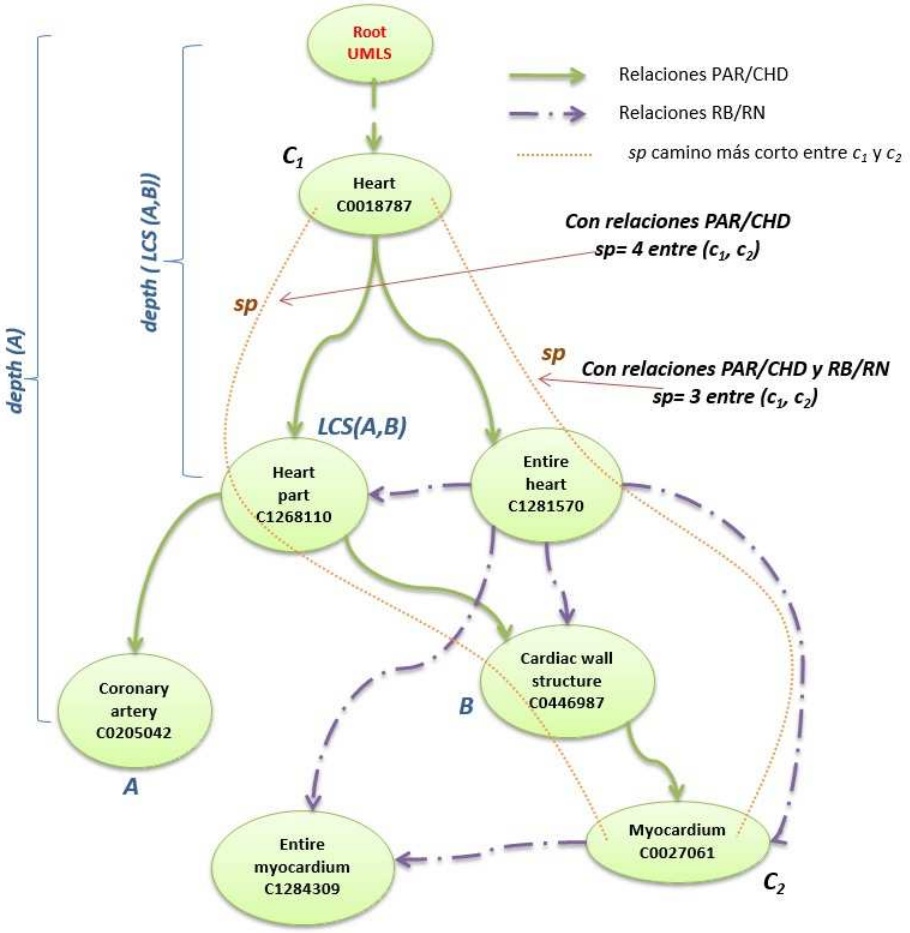

Figura 3.4 Ejemplo de relaciones jerárquicas entre conceptos del Metatesauro UMLS. Los términos *depth*, *path* y *LCS* son representados en la figura.

- ***Métricas basadas en Información Contenida (IC)***

Otros enfoques, se basan en la Teoría de la Información de Shannon (Shannon, 2001) para el cálculo de la similitud semántica entre términos o conceptos. En este caso, serán evaluados de acuerdo a la cantidad de información contenida que éstos proporcionan.

La aplicación de la idea basada en la Información Contenida (IC), puede ser definida de dos diferentes maneras. La primera se denomina (Corpus IC) (Resnik, 1995) y se centra en la distribución de un término o concepto en el corpus de un texto, con la posibilidad de apoyarse sobre una taxonomía. Mientras que la segunda, denominada (Intrinsic IC) (Seco, Veale, & Hayes, 2004; Zhou, Wang, & Gu, 2008; Sánchez, Batet, & Isern 2011), estima dicha distribución a partir únicamente de la estructura definida en una taxonomía dada.

- **Corpus IC:**

Se puede decir que el **Corpus IC** de un concepto "*c*", es definido como el inverso logarítmico de su frecuencia, donde la frecuencia de ese concepto es la probabilidad de que este suceda un número de veces en un corpus "*C*", *fq(c,C)*. Además, se añadirá el número veces que ocurran sus hijos *(c_s).* De manera que, cuanto más ocurra un concepto, menor información aporta (Resnik, 1995)

$$IC_{Corpus}(c) = -\log\big(fq(c)\big)$$
$$fq(c) = fq(c, C) + \sum_{c_s \in children(c)} fq(c_s) \tag{3.12}$$

- **Intrinsic - IC:**

Por otro lado, **Intrinsic IC** de un concepto "*c*", es una propuesta también definida en diversos trabajos (Seco, Veale, & Hayes, 2004; Zhou, Wang, & Gu, 2008) y que ha sido adaptada al contexto biomédico (Batet, Sánchez, & Valls, 2011). Se puede definir Intrinsic IC de un concepto "*c*", como el ratio entre el número de sus conceptos terminales "*leaves(c)"* y la cantidad de sus ancestros asociados "*subsumers(c)"* (Sánchez, & Batet, 2011). Este ratio se normalizará para el intervalo [0-1] por el número total de hojas en la taxonomía "*max_leaves"*. Así, cuantos más elementos terminales tenga un concepto en relación al número de sus ancestros, menos información aporta.

$$IC_{Intrinsic}(c) = -log\left(\frac{\frac{leaves(c)}{subsumers(c)} + 1}{\max\_leaves + 1}\right) \tag{3.13}$$

Los anteriores enfoques de Información Contenida, darán lugar a las diferentes medidas de similitud semántica definidas por Lin (Lin, 1998) y Jian y Conrath (Jiang, & Conrath, 1997). La medida de similitud definida por Lin propone el ratio entre la información contenida común para un par de conceptos "*IC(LCS(c_1,c_2))"* y la información contenida que describe a cada concepto "*IC(c_i)"*. Por ello, cuanto mayor es el valor de IC del LCS (es decir, más específico), mayor es la similitud entre los conceptos comparados

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC\left(LCS\,(c_1, c_2)\right)}{IC(c_1) + IC(c_2)} \tag{3.14}$$

Por otro lado, Jian y Conrath proponen una medida análoga (opuesta a la similitud), basada en la distancia de conceptos como la diferencia entre la información contenida de cada concepto "*IC(c_i)"* y la información contenida de su ancestro común "*IC(LCS(c_1,c_2))".*

$$Dist_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC\big(LCS(c_1, c_2)\big) \tag{3.15}$$

Las métricas basadas en "*Path Finding*" fueron redefinidas (Batet, Sánchez, & Valls, 2010) en términos de Información Contenida e implementadas para su evaluación (Garla, & Brandt, 2012). Para ello, se redefine el camino más corto entre dos conceptos *(sp)* como

la distancia semántica según Jian y Conrath y la máxima profundidad *(depth)*, como la máxima Información Contenida de cualquier concepto *(ic_{max})*.

Con todo ello, la métrica (*lch)* basada en Intrinsic IC ***(Intrinsic IC-lch)***, se redefine como:

$$sim_{\substack{Intrinsic \\ IC\_lch}}(c_1, c_2) = 1 - \frac{\left(log\left(Dist_{JC}(c_1, c_2) + 1\right)\right)}{log(2 \times ic_{max} + 1)} \tag{3.16}$$

Mientras que la métrica (*Path)*, basada en Intrinsic IC (***Intrinsic IC-Path***), como:

$$sim_{\substack{Intrinsic \\ IC\_Path}}(c_1, c_2) = \frac{1}{Dist_{JC}(c_1, c_2) + 1} \tag{3.17}$$

Estas métricas han sido evaluadas en diversos trabajos sobre diferentes *benchmark* de ensayo (Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Batet, Sánchez, & Isern, 2011; Garla, & Brandt, 2012). Estos trabajos revelan un cierto mejor comportamiento de las métricas basadas en "*Intrinsic_IC*" frente a las basadas en "*Path Finding*" (Garla, & Brandt, 2012).

3.4.2. Métricas de Relación Semántica

La relación semántica es una representación más general que la similitud. En este caso (Patwardhan, & Pedersen 2006), trata de representar la relación semántica de un concepto a partir de un vector de contexto (*Context Vector*), y donde la fuente de información del vector es un corpus de texto y no los caminos entre conceptos. La métrica del vector de contexto, funciona a partir de la formación de un vector de ocurrencias de segundo orden a partir de las definiciones extendidas de un concepto en un corpus, como el Mayo Clinic Corpus (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). La relación entre los dos conceptos se determina como el coseno del ángulo entre los dos vectores, siendo v_1 el vector para el concepto c_1 y v_2 para c_2.

$$rel_{vector}(c_1, c_2) = \frac{\vec{v1} \cdot \vec{v2}}{|v_1| . |v_2|} \tag{3.18}$$

Existen numerosas métricas de relación semántica derivadas e inspiradas en el modelo "*Context Vector*", (Schütze, 1998; Banerjee, et al., 2002). Pero en este caso, no se profundizará en dichas métricas al no ser el objeto de esta tesis.

Capítulo 4

4. Objetivos y alcance de la tesis

En las secciones previas se ha realizado un estudio detallado entorno a la dificultad de acceso y tratamiento de la documentación en el contexto biomédico, se han revisado los trabajos que proponen las diferentes técnicas de recuperación de información en dicho contexto y se han analizado las métricas de similitud semántica aplicadas a bancos de prueba teóricos. En esta sección por lo tanto se plantea la problemática encontrada en este entorno, así como el nexo de partida para definir el alcance de esta tesis y los objetivos necesarios a cumplir para su consecución

4.1. Justificación

Tal y como se vio en la introducción de este trabajo, hoy en día el volumen de datos digitales crece de una manera exponencial en diversos y variados entornos de aplicación. En concreto, en el contexto de la documentación médica continuamente se generan grandes volúmenes de datos digitales, denominados historias clínicas electrónicas (EHR). Sus objetivos iniciales son estandarizar y mejorar la calidad asistencial del paciente (Middleton, et al., 2013; Hoffman, et al., 2011). Sin embargo, esta información ofrece otras oportunidades en la investigación científica, como es la identificación de grupos de cohortes en pacientes, el control y seguimiento de epidemias, la evaluación de tratamientos, el seguimiento y anticipación de ciertas enfermedades, etc.

Para estudiar y analizar dicha información, son necesarias infraestructuras capaces de modelar dichos datos (Howe, et al., 2008). Surgiendo de esta manera, estructuras que recogen el conocimiento del dominio biomédico, como el Metatesauro UMLS y estrategias que se apoyen en ellas para discernir cierta información, como es la similitud semántica entre conceptos.

Es por ello que tras realizar una completa revisión del estado del arte y analizar las diferentes métricas de similitud semántica entre conceptos propuestas en el entorno biomédico, se ha observado la necesidad de definir un marco de evaluación estable y confiable que permita replicar los resultados de los diferentes trabajos previos en contextos teóricos y poder así analizarlos en igualdad de condiciones. De esta forma se podrán interpretar los diversos resultados de forma conjunta y desde diferentes puntos de

vista, al mismo tiempo que se podrá optimizar los resultados de dichas métricas, mediante el uso correcto de recursos y relaciones contenidas en el Metatesauro UMLS.

No obstante, estas métricas deberán ser evaluadas en un contexto de información más amplio y abierto, que permita valorar su rendimiento real aplicado a una finalidad concreta, como es la recuperación automática de registros médicos.

4.2. Objetivos

Por ello, se definen dos principales objetivos a alcanzar en el desarrollo de esta tesis:

- Clarificar y unificar los criterios de aplicación de las métricas de similitud semántica en los diferentes contextos teóricos de información propuestos en trabajos anteriores.
- Proponer un novedoso sistema de recuperación de información aplicado a contextos reales de documentación médica que utilice las principales métricas de similitud semántica sobre el recurso UMLS, el cual permitirá valorar el comportamiento de dichas métricas en nuevos escenarios.

Para la consecución de dichos objetivos, será necesaria la realización de diferentes tareas.

- Configuración de software, herramientas y recursos necesarios:
  - Recursos del conocimiento biomédico (UMLS).
  - Herramientas necesarias para la aplicación de las métricas de similitud semántica entre conceptos (UMLS::Similarity e Ytex).
  - Herramienta de extracción de conceptos en texto (Metamap).
- Obtener un marco de trabajo común sobre el Metatesauro UMLS que permita evaluar de forma fiable los resultados de las distintas métricas de similitud semántica en contextos teóricos de información. Así como validar la influencia de la infraestructura UMLS, en los resultados aportados por dichas métricas.
  - Evaluar y definir el impacto de las diferentes relaciones y recursos del Metatesauro UMLS en los resultados de las distintas métricas de similitud semántica.
  - Evaluar la importancia de los cambios de versiones entre recursos, en la aplicación de las métricas de similitud semántica.
  - Identificar y clarificar el impacto de las distintas correlaciones utilizadas en trabajos previos en la interpretación de los resultados.
- Proponer un nuevo sistema de recuperación de información basada en documentación médica que integre el uso del Metatesauro UMLS aplicado a las métricas de similitud semántica. Este sistema permitirá validar y comparar las principales métricas de similitud semántica en un contexto real de información, realizando las siguientes tareas:
  - Preprocesado y limpieza de documentos médicos.
  - Extracción de conceptos en documentos y consultas (uso herramienta Metamap).

- Procesado de documentos y consultas (generación de variaciones de frases y subfrases en base a conceptos).
- Filtrado de la consulta por tipos semánticos.
- Generación de matriz de máxima similitud (*report-topic).*
- Cálculo de Relevancia semántica de la matriz (*report-topic).*

4.3. Herramientas

En el siguiente apartado, se describen cada una de las herramientas software y los recursos de información biomédica que se han utilizado en el desarrollo de esta tesis.

4.3.1. Text REtrieval Conference[34] (TREC-Medical Records Track)

El TREC 2011 es el año inaugural para el desafío del Medical Records Track. Este pretende fomentar la investigación sobre tecnologías que permitan la recuperación de registros médicos electrónicos, basándose en la semántica contenida en sus campos de texto libre. La capacidad para recuperar registros haciendo coincidir su contenido semántico, permitirá que los registros médicos sean de gran utilidad en áreas tales como estudios epidemiológicos y ensayos clínicos. Para ello, se propone una **Colección de Prueba** que se compone de tres elementos diferenciales:

- Un conjunto de **Registros Médicos** agrupados en visitas o episodios médicos.
- Unas **Necesidades de Información**, recogidas a través de un conjunto de diferentes consultas a evaluar.
- Los **Juicios de Relevancia**, que han sido establecidos por expertos médicos, determinando que visitas (*reports*) son relevantes o no, para una determinada necesidad de información.

Cuando se diseña un estudio clínico, un investigador médico desarrollará "criterios de inclusión" que describen los tipos de pacientes requeridos para dicho estudio. Estos criterios incluyen atributos del tipo como enfermedades presentes, tratamientos, grupos de edad, género o etnia, etc.

Las tareas de recuperación de dicho desafío, pretenden que las sentencias que forman los temas de búsqueda (*topics*) se modelen para incluir dichos criterios. De forma que los sistemas devuelvan una lista de visitas o episodios ordenados (cada visita puede estar asociado a uno o más registros médicos), según la probabilidad de que estos satisfagan los criterios de inclusión. Por tanto, los *topics* reflejan los tipos de consultas que se pueden utilizar para la identificación de cohortes en la investigación comparativa efectiva. Dichas tareas de recuperación se realizarán sobre un conjunto de-identificado de documentos clínicos obtenidos de diferentes hospitales.

Hoy en día la mayoría de los sistemas de basan en el acceso a información estructurada, sin embargo el contenido de los registros médicos se componen de notas y otros campos de texto libre, los cuales no son estructurados. Es por esto, que las técnicas

[34] http://trec.nist.gov/

estándar de procesamiento del texto no funcionan de manera correcta en estos campos. Éstos rara vez contienen sentencias bien formadas y gramaticalmente correctas, el vocabulario es altamente especializado, con términos especiales (abreviaturas, medidas, símbolos, etc.) y donde los textos se refieren de manera implícita a otras partes del registro.

A continuación, se describen más ampliamente cada uno de los componentes de la **Colección de Prueba** del TREC 2011:

- **Registros Médicos (*reports* y *visitas*)**

Es el conjunto de documentos médicos utilizado en el desafío del *Medical Records Track,* es un conjunto de informes clínicos de-identificados disponibles por mediación de la Universidad de Pittsburgh (*NLP Repository*). Este repositorio incluye 9 tipos de informes: Informes Radiológicos (*Radiology Reports*), Historia y exámenes físicos (*History and Physicals*), Informes de Consultas (*Consultation Reports*), Informes de Urgencias (*Emergency Department Reports*), Notas de progreso (*Progress Notes*), Informes de alta (*Discharge Summaries*), Informes operativos (*Operative Reports*), Informes de patologías quirúrgicas (*Surgical Pathology Reports*), e Informes de Cardiología (*Cardiology Reports*).

Un informe (*report*) es enlazado a una visita (*visit*). Aunque la visita representa una estancia o episodio único de un paciente en un hospital, la visita de un paciente, puede componerse de uno a varios informes (las diferentes visitas de un mismo paciente han sido intencionadamente desvinculadas como proceso de la de-identificación). El mapeo entre un informe y una visita, es codificado a través de una tabla que da el correspondiente identificador de visita para cada informe (*ReportMappingToVisit-3.txt*).

Cada informe está compuesto de códigos de diagnóstico (ICD9-CM) asociados a la visita y de texto libre.

El repositorio del Medical Records Track, se compone de:

- 101.712 informes, enlazados a una de las 17.265 visitas.
- La distribución de la mayoría de las visitas se asocia al menos con 5 informes. El tamaño máximo de una visita se compone de 415 informes.
- La unidad o "documento" de recuperación y evaluación en TREC es la visita. Si bien cada visita, se compone de uno o varios informes, la unión de estos informes formará el "documento" de recuperación.

Seguidamente, se describen los restantes elementos necesarios para las tareas de recuperación (los temas de búsqueda o *topics* y los juicios de relevancia de los documentos). Estos fueron definidos por facultativos que también eran estudiantes del programa de bioinformática de la Oregon Health and Science University (OHSU). A estos asesores se les dio una lista de las áreas de investigación del U.S. Institute of Medicine, consideradas decisivas para la investigación comparativa efectiva, las cuales fueron usadas como pautas para dichos *topics*.

- **Necesidades de Información (*topics*)**

Las necesidades de información, se representarán mediante sentencias que pueden incluir desde enfermedades, tratamientos, grupos de edad, géneros, etnias, etc. Estas sentencias incluyen los criterios que deben satisfacer las *visitas* (por tanto sus *reports* contenidos) más relevantes para su inclusión en el resultado.

Las sentencias que definen las necesidades de información (*topics*), deberán ser transformadas a estructuras, basadas en lenguaje natural, que se entregarán a los sistemas para su ejecución (*query*). La creación y ejecución de las consultas pueden realizarse de manera completamente automática, de forma que el sistema tome como entrada la sentencia *topic* y produzca una lista ordenada de identificadores de visitas, sin intervención humana. Mientras que las consultas manuales pueden ser de cualquier otra manera (esta definición es muy abierta, ya que va desde ajustes de consultas automáticas, a construcciones manuales de la consulta inicial, a reformulaciones de nuevas consultas basadas en los conjuntos recuperados, etc.).

Los temas de búsqueda (*topics*) son los siguientes:

- Un conjunto de prueba formado por 35 sentencias. Estos temas de búsqueda están identificados desde el 101 al 135. (*topics101-135.txt*).

- Un conjunto ejemplo de entrenamiento formado por 4 sentencias. (*topics-sample.txt*)

- **Juicios de Relevancia**

Los juicios de relevancia han sido establecidos por los asesores de la OSHU. Para cada *topic*, se entregó a los asesores una media de 250 *visitas* para ser juzgadas. Los asesores debían estudiar los documentos médicos asociados con una *visita*, para juzgar si la visita era "no relevante - 0", "parcialmente relevante - 1" o "relevante - 2", con respecto al *topic*.

El conjunto de juicios de relevancia final para cada *topic* se recoge en el documento "*Qrels.txt*". El cual contiene por cada fila: El número de ***topic***, el **identificador de la *visita* evaluada** y **la *relevancia*** de dicho documento con respecto al *topic*. Finalmente, en el desafio de TREC 2011, las visitas juzgadas como "parcialmente relevantes" y "relevantes" por los expertos, fueron definidas como documentos relevantes.

4.3.2. MetaMap[35] release 2013

MetaMap es una herramienta altamente configurable, desarrollada por Alan Aronson (Aronson, & Lang, 2010), como un proyecto de la National Library of Medicine (NLM).

MetaMap tiene como objetivo, determinar y mapear de manera automática los conceptos del Metatesauro UMLS referenciados en el texto. Este es un prerrequisito primordial para ciertas aplicaciones relacionadas con la recuperación de información, minería de textos, categorización, clasificación, resúmenes, consulta-respuesta, extracción del conocimiento y otras tareas del procesamiento del lenguaje natural. Es por ello, que

[35] http://metamap.nlm.nih.gov/

MetaMap utiliza aproximaciones basadas en el procesamiento del lenguaje natural y técnicas lingüísticas computacionales.

MetaMap fue originalmente desarrollado para la mejora en la recuperación de información de citas relevantes en MEDLINE. El cuál es la base de del indexador de texto médico automático (MTI)[36] , aplicado a la literatura biomédica de la NLM.

Los conceptos buscados en el texto son aquellos contenidos en el Metatesauro UMLS. La tarea comienza con el mapeo del texto frente al Metatesauro. El resultado del mapeo del texto contenido en frases se puede clasificar en una de las **cuatro categorías** en función de su adecuado ajuste al Metatesauro UMLS.

***Simple match*:** El sintagma nominal se mapea exactamente a una cadena del Metatesauro UMLS.

Por ejemplo "prueba1.txt": **`intensive care unit`**

```
Processing 00000000.tx.1: intensive care unit

Phrase: "intensive care unit"
Meta Candidates (Total=12; Excluded=1; Pruned=0; Remaining=11)
  1000   C0021708:intensive care unit [Health Care Related Organization,Manufactured Object]
  1000   C1549475:Intensive care unit (Room type - Intensive care unit) [Idea or Concept]
  1000   C1610734:Intensive care unit (Intensive care unit - RoleCode) [Intellectual Product]
   901   C0085559:intensive care [Health Care Activity]
   901   C1548354:Intensive care (Specialty Type - Intensive care) [Biomedical Occupation or
Discipline]
   827   C0439148:Unit [Quantitative Concept]
   827   C1519795:Unit (Unit of Measure) [Quantitative Concept]
   827   C1704434:UNIT (Unit Device Component) [Medical Device]
   827   C1704753:Unit (Storage Unit) [Manufactured Object]
   827   C1880519:Unit (Enzyme Unit) [Quantitative Concept]
   827   C1947933:Care (care activity) [Activity]
   793 E C0150499:Caring [Individual Behavior]
Meta Mapping (1000):
  1000   C1610734:Intensive care unit (Intensive care unit - RoleCode) [Intellectual Product]
Meta Mapping (1000):
  1000   C1549475:Intensive care unit (Room type - Intensive care unit) [Idea or Concept]
Meta Mapping (1000):
  1000   C0021708:intensive care unit [Health Care Related Organization,Manufactured Object]
```

***Complex match*:** Existe una partición de las palabras en sintagmas nominales, donde cada elemento particionado refleja un "simple match" con el Metatesauro.

Por ejemplo "prueba2.txt": `intensive care medicine`

```
Processing 00000000.tx.1: intensive care medicine

Phrase: "intensive care medicine"
Meta Candidates (Total=7; Excluded=2; Pruned=0; Remaining=5)
   827   C0013227:Medicine, NOS (Pharmaceutical Preparations) [Pharmacologic Substance]
   827   C0025118:Medicine [Biomedical Occupation or Discipline]
   755 E C0205476:Medical [Functional Concept]
   734   C0085559:intensive care [Health Care Activity]
   734   C1548354:Intensive care (Specialty Type - Intensive care) [Biomedical Occupation or
Discipline]
   660   C1947933:Care (care activity) [Activity]
   627 E C0150499:Caring [Individual Behavior]
Meta Mapping (901):
   734   C1548354:Intensive care (Specialty Type - Intensive care) [Biomedical Occupation or
Discipline]
```

[36] http://ii.nlm.nih.gov/MTI/

```
   827   C0025118:Medicine [Biomedical Occupation or Discipline]
Meta Mapping (901):
   734    C1548354:Intensive care (Specialty Type - Intensive care) [Biomedical Occupation or
Discipline]
   827   C0013227:Medicine, NOS (Pharmaceutical Preparations) [Pharmacologic Substance]
Meta Mapping (901):
   734   C0085559:intensive care [Health Care Activity]
   827   C0025118:Medicine [Biomedical Occupation or Discipline]
Meta Mapping (901):
   734   C0085559:intensive care [Health Care Activity]
   827   C0013227:Medicine, NOS (Pharmaceutical Preparations) [Pharmacologic Substance]
```

***Partial match*:** El sintagma nominal mapea con respecto al Metatesauro de manera que al menos una palabra del sintagma o de la cadena del Metatesauro no participa en el mapeo.

Por ejemplo "prueba3.1.txt": `cochlear implant subjects`

```
Processing 00000000.tx.1: cochlear implant subjects

Phrase: "cochlear implant"
Meta Candidates (Total=9; Excluded=2; Pruned=0; Remaining=7)
  1000   C0009199:Cochlear implant (Cochlear Implants) [Medical Device]
  1000    C0302559:cochlear implant (Cochlear implant procedure) [Therapeutic or Preventive
Procedure]
   861   C0009195:Cochlear (Cochlear structure) [Body Part, Organ, or Organ Component]
   861   C0021102:Implant (Implants) [Medical Device]
   861   C0021107:implant (Implantation procedure) [Therapeutic or Preventive Procedure]
   861   C0332837:Implant (Traumatic implants) [Injury or Poisoning]
   861   C1711357:Implant (Administration via Implantation) [Functional Concept]
   827  E  C2926601:Implants  (Procedure  implants:Finding:Point  in  time:^Patient:Narrative)
[Clinical Attribute]
   789 E C1278895:Cochlea (Entire cochlea) [Body Part, Organ, or Organ Component]
Meta Mapping (1000):
  1000   C0009199:Cochlear implant (Cochlear Implants) [Medical Device]
Meta Mapping (1000):
  1000    C0302559:cochlear implant (Cochlear implant procedure) [Therapeutic or Preventive
Procedure]

Phrase: "subjects"
Meta Candidates (Total=5; Excluded=0; Pruned=0; Remaining=5)
   966   C0681850:Subject (Study Subject) [Group]
   966   C1550501:{Subject} (Subject -direct target) [Idea or Concept]
   966   C1706203:Subject (Subject - topic) [Idea or Concept]
   966   C2349001:Subject (Human Study Subject) [Human]
   966   C2697811:Subject (Investigative Subject) [Functional Concept]
Meta Mapping (966):
   966   C2349001:Subject (Human Study Subject) [Human]
Meta Mapping (966):
   966   C2697811:Subject (Investigative Subject) [Functional Concept]
Meta Mapping (966):
   966   C0681850:Subject (Study Subject) [Group]
Meta Mapping (966):
   966   C1706203:Subject (Subject - topic) [Idea or Concept]
Meta Mapping (966):
   966   C1550501:{Subject} (Subject -direct target) [Idea or Concept]
```

***No match*:** Ninguna parte del sintagma nominal tiene mapeo con ninguna cadena del Metatesauro.

Las categorías anteriores se ordenan por la fuerza de su mapeo, siendo "*simple match*" la más fuerte. Cabe destacar sin embargo, que la calidad semántica o conceptual del mapeo varía ampliamente.

- **Estrategia de Mapeo**

Los pasos a seguir en la estrategia de mapeo de cada expresión textual son:

1. Analizar el texto en sintagmas o frases nominales y realizar los pasos siguientes para cada frase.
2. Generar variaciones para cada frase nominal (consistentes en una o más frases con variaciones, abreviaciones, acrónimos, sinónimos, inflexiones y principales combinaciones.)

   El algoritmo para la generación de variaciones se basa en el conocimiento, extraído de los siguientes recursos:
   - El lexicón SPECIALIST y la tabla de formas canonícas derivadas de él.
   - La base de conocimiento SPECIALIST de acrónimos y abreviaturas.
   - La base de conocimiento SPECIALIST de reglas de morfología derivacional y de sinónimos.
3. Formar el conjunto de candidatos, conteniendo cada una de las variaciones. (**Meta Candidates**)
4. Para cada candidato calcular su fuerza mediante funciones de evaluación.
5. Seleccionar y combinar aquellos candidatos que más y mejor se ajustan al mapeo de la frase original. (**Meta Mapping**).

Los **Meta Mappings** serán los elementos finales que definen las diferentes variaciones de conceptos asociada a una representación textual.

- **Ejecución**

La ejecución de Metamap se realiza mediante el comando metamap12 (12 corresponde a la última versión) y su comportamiento se controla mediante opciones. Las opciones tienen una versión larga y corta, por ejemplo para mostrar los CUIs se utiliza (--show_cuis) o (-I).

Su uso es: metamap12 [Options] [InputFile] [OutputFile]

Hay dos maneras de ejecutar MetaMap:

- Pasando directamente la cadena como entrada a MetaMap y obteniendo los resultados por pantalla. Con esta opción se visualizan las opciones de control seleccionadas. (directorio "public_mm" de la instalación de MetaMap para su ejecución.)

  Echo "texto a analizar" | ./bin/metamap12 –I –n

- Tomando como entrada un fichero de texto y recogiendo el resultado de la salida en otro fichero, es necesario estar en el directorio que contiene dicho fichero. Con esta opción no se recogen y visualizan las opciones de control en el fichero de salida.

  /metamap12 –I –n fichero.txt fichero.out

La parametrización aplicada para la ejecución de MetaMap, en el trabajo desarrollado en esta tesis, con el objeto de extraer la clasificación de conceptos representativos de los documentos médicos propuestos por el TREC, es la siguiente:

Metamap12 –f –Q 3 –u –E –prune 25 –negex_st_add all –XMLf

Cada uno de los parámetros aplicados tiene las siguientes funciones:

- **-f (--number_the_mappings)**

Numera los resultados de mappings obtenidos.

- **-Q <integer> (--composite_phrases <integer>)**

Esta opción permite a Metamap construir frases compuestas más largas, a partir de frases simples producidas por el analizador. Una frase compuesta es un sintagma nominal seguido de cualquier frase preposicional y opcionalmente seguidas por una o más frases preposicionales introducidas por "of". En resumen, una frase compuesta consiste en:

- un nombre seguido de
- cualquier frase preposicional, opcionalmente seguida por
- una o más frases preposicionales introducidas por "of"

Un ejemplo es "*pain on the left side of the chest*", el cual enlazará con ("*left sided chest pain*") con la opción de frase compuesta activa, pero enlazará con conceptos separados si esta opción no está activa ("*pain*"; "*left side*"; "*chest*").

El máximo número de frases preposicionales está bajo control del usuario y se puede especificar proporcionando obligatoriamente un argumento entero a esta opción, esta opción permite controlar cuantas frases preposicionales se permiten aprovechar sobre el sintagma nominal. Los experimentos sugieren que 3 o 4 son valores razonables. Este comando invoca automáticamente a --ignore_word_order y --term_processing.

- **-u (--unique_acros_abbrs_only);**

Restringe la generación de variantes acrónimo/abreviatura para aquellas formas con expansiones únicas. Los experimentos han revelado que esta opción generalmente produce mejores resultados que permitir *todas* las formas de variantes acrónimo/abreviatura.

- **-E (--indicate_citation_end)**

Escribe 'EOT' (End Of Transmission) cuando se finaliza el proceso de cada unidad o fichero de entrada

- **`--negex (no short form)`**

La versión del 2013, contiene una extensión de la implementación del algoritmo de negación "*negex*", el cual permite identificar la negación en el texto contenido, aplicados a todos los tipos semánticos, así como añadir o eliminar tipos semánticos al conjunto establecidos por defecto.

- **`--XMLf (no short form)`**

Genera una salida XML, que permite su procesamiento posterior de una manera automatizada. Para su procesamiento, han sido necesario el desarrollo de diversos procesos Perl y scripts de lenguaje Shell Linux y awk.

4.3.3. UMLS –Similarity[37] / Ytex[38]

UMLS::Similarity es un paquete Perl, formado por un conjunto de módulos que implementan diferentes métricas de similitud semántica con el objeto de calcular la similitud entre dos conceptos UMLS. Algunas de estas métricas implementadas son:

- Path Length (Path)
- Conceptual Distance (cdist)
- Leacock & Chodorow (lch)
- Nguyen & Al-Mubaid (nam)
- Wu & Palmer (wup)
- Lin (lin)
- Jiang & Conrath (jcn)
- Resnik (res)

Estas métricas hacen uso del módulo Perl UMLS-Interface, para el acceso a los recursos UMLS que se encuentren cargados en la bases de datos mysql.

Ytext es parte del proyecto cTAKES y provee una funcionalidad que ha sido empleada también en los trabajos de esta tesis. Esta funcionalidad se denomina **Semantic Similarity**, y se compone de un entorno para el cálculo de la similitud semántica entre pares de conceptos UMLS. Este entorno está integrado y relacionado con otros componentes como son clinical NLP, Data Mining, y herramientas Feature Engineering.

- Path-Finding Measures

  WUPALMER: Wu & Palmer
  LCH: Leacock & Chodorow
  PATH: Path
  RADA: Rada

- Corpus IC Based Measures:

  LIN: Lin

[37] http://search.cpan.org/dist/UMLS-Similarity/
[38] https://code.google.com/p/ytex/

- Intrinsic IC Based Measures:

  INTRINSIC_LIN: Intrinsic IC based Lin
  INTRINSIC_LCH: Intrinsic IC based Leacock & Chodorow
  INTRINSIC_PATH: Intrinsic IC based Path, idéntico a Jiang & Conrath
  INTRINSIC_RADA: Intrinsic IC based Rada
  JACCARD: Intrinsic IC based Jaccard
  SOKAL: Intrinsic IC based Sokal & Sneath

4.3.4. Perl, awk y utilidades shell linux

Para la consecución del trabajo y objetivos del capítulo 6, ha sido necesaria la utilización de diferentes lenguajes de programación y utilidades para:

- El procesado previo de los registros médicos (reports) y las consultas (consultas).
- El desarrollo de un sistema de recuperación de información para un entorno de información biomédica, basada en conceptos UMLS y métricas de similitud semántica.
- El estudio experimental del rendimiento, evaluación y análisis de las dos principales métricas (*Path* e *Intrinsic IC-Path*) en un contexto real de recuperación de información.

Capítulo 5

5. Evaluación UMLS y métricas de similitud semántica en contexto teórico

En este capítulo de la tesis, se analizarán las características existentes en la herramienta base utilizada en este trabajo, el Metatesauro UMLS, que permitan mejorar el cálculo de la similitud semántica. Estas características a analizar, serán la evaluación de los recursos y tipos de relaciones existentes entre conceptos en las versiones de UMLS empleadas.

Este estudio se centra en las métricas más importantes de similitud semántica entre dos conceptos asociados a recursos del dominio biomédico. El objetivo que se define es evaluar los resultados obtenidos por dichas métricas para el conjunto de prueba propuesto formado por 29 pares de conceptos (McInnes, Pedersen, & Pakhomov, 2009; Pedersen, Pakhomov, Patwardhan, & Chute, 2007), con la actualización 2010AB del Metatesauro UMLS. En estos trabajos referenciados se ha utilizado la versión UMLS 2008AB para los recursos SNOMED-CT y MeSH sobre relaciones PAR/CHD.

Además, se evaluará cómo afecta a los resultados la utilización (independiente o combinada) de diferentes recursos del dominio y la incorporación específica de relaciones jerárquicas indirectas (RB/RN). También, se analizará globalmente el impacto de todas las relaciones existentes en el Metatesauro UMLS.

Por último, la falta de normalización en la utilización de los coeficientes de correlación en los distintos trabajos arroja resultados con grandes diferencias entre ellos (Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Batet, Sánchez, & Valls, 2011). Por este motivo, en este trabajo se justifica el uso del coeficiente de correlación de Pearson como medida estándar de utilización de dichos trabajos.

Todas las métricas expuestas en la revisión del estado del arte han sido desarrolladas bajo diferentes plataformas y aplicadas a conjuntos de pruebas no equivalentes. Para poder evaluarlas, de una forma coherente y sistemática, se definió un conjunto de prueba con la ayuda de 9 codificadores médicos y 3 facultativos de la clínica Mayo ellos (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). Este conjunto se formó inicialmente a partir de 120 parejas de conceptos que los expertos evaluaron, asignando una puntuación que

representa la similitud semántica existente entre ellos. Esta puntuación se establece en una escala de 4 a 0 en el que se pretendía definir si dos conceptos eran prácticamente sinónimos (4.0), relacionados (3.0), ligeramente relacionados (2.0) y no relacionados (1.0). (Estas escalas posteriormente han sido normalizadas en el rango de valores [0;1], para su mejor interpretación. Siendo el valor "1" indicativo de la mayor similitud y "0" ninguna similitud). Tras la evaluación, se extrajo un conjunto de prueba confiable formado por 30 pares de conceptos ordenados por grado de similitud descendente, cuya concordancia era alta. Finalmente, se excluyó del conjunto el término "*lung infiltrates*", ya que no se encontraba en la terminología SNOMED-CT de la versión de 2004. El conjunto de prueba final está formado por 29 pares de conceptos y el coeficiente de correlación de Pearson obtenido de la valoración de similitud otorgado por el grupo de expertos a estos conceptos es de 0.85.

5.1. Métricas sobre recursos MeSH y Snomed-CT en versiones de UMLS

En primer lugar se evalúan las distintas métricas sobre el recurso MeSH para las relaciones de tipo PAR/CHD, con el objetivo de contrastar los resultados aquí obtenidos, con los aportados por McInnes y Pedersen (McInnes, Pedersen, & Pakhomov, 2009; Pedersen, Pakhomov, Patwardhan, & Chute, 2007). Los valores obtenidos para la nueva versión MeSH (UMLS 2010AB) reflejaron que 11 pares de conceptos del conjunto de pruebas, no devolvieron ningún resultado. El resto de valores obtenidos, fueron muy próximos a los recogidos por McInnes y Pedersen. Debido al alto número de conceptos excluidos, se optó por no realizar el cálculo de correlación de estos valores con los definidos por el grupo de expertos, ya que se tratan de conjuntos de evaluaciones diferentes y no equiparables.

Por este motivo, se han replicado los resultados obtenidos por Pedersen (Pedersen, Pakhomov, Patwardhan, & Chute, 2007) y McInnes (McInnes, Pedersen, & Pakhomov, 2009) sobre el recurso SNOMED-CT de la versión 2008AB de UMLS (utilizada en su trabajo) y la versión 2010AB (Tabla 5.1).

Esta tabla refleja los coeficientes de correlación Spearman de la métrica *Path* frente a las estimaciones de los expertos facultativos y *coders*. Por un lado, se muestran los resultados del coeficiente de correlación basado en valores mínimos del ranking de ordenación para grupos de valores de similitud con repetición utilizado en (McInnes, Pedersen, & Pakhomov, 2009; Pedersen, Pakhomov, Patwardhan, & Chute, 2007) y, por otro, los resultados de correlación basados en valores medios de dicho ranking (empleado en este trabajo por considerarlo el cálculo más adecuado en este contexto).

	SNOMED-CT 2008AB		SNOMED-CT 2010AB	
	Minimum values	**Average values**	**Minimum values**	**Average values**
Physicians	0.3500	0.3170	0.3134	0.2744
Coders	0.5000	0.4500	0.4596	0.4160

Tabla 5.1 Valores de correlación de similitud semántica utilizando la métrica *Path* con relaciones PAR/CHD para las versiones de SNOMED-CT con UMLS 2008AB y 2010AB. Correlaciones de Spearman basadas en valores mínimos del ranking (resultados replicados del trabajo de Pedersen y McInnes para la versión 2008AB) y en valores medios (empleado en este trabajo).

Como se observa, las correlaciones obtenidas (utilizando los coeficientes de Spearman), varían significativamente en estas dos versiones de SNOMED-CT, en torno a un 6-13%. Este hecho revela que en la versión 2010AB se han refinado las relaciones entre conceptos y, por tanto, los resultados son inferiores (relaciones de similitud que se daban anteriormente, ya no se encuentran). Por tanto, deberá tenerse en cuenta este hecho a la hora de comparar los resultados de los diferentes trabajos, ya que en muchos de ellos se pueden estar comparando el rendimiento de métricas que son ejecutadas sobre diferentes versiones del Metatesauro UMLS.

En estos resultados y en el resto de los obtenidos en este trabajo, se puede observar que las métricas se ajustan más a los criterios de similitud definidos por los expertos codificadores que a los indicados por los facultativos.

5.2. Estudio e impacto de las correlaciones utilizadas

Los resultados anteriores nos han conducido al análisis en profundidad de las variables empleadas en los trabajos de McInnes y Pedersen, observamos que no hay un criterio estándar en la utilización del coeficiente empleado. Por ejemplo, unos trabajos utilizan el coeficiente de correlación de Spearman (Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Garla, & Brandt, 2012), este muestra cómo de ajustado es el ordenamiento de los resultados obtenidos por las métricas, en relación al orden del rango de valores definidos por los expertos. Mientras que otros utilizan el índice de correlación lineal de Pearson (Batet, Sánchez, & Valls, 2011), el cual, muestra la proximidad o ajuste de los resultados de similitud obtenidos con respecto a los valores definidos por los expertos (Tabla 5.2). Esto resulta importante para la utilización de estas métricas en sistemas de recuperación de información. Por ello, en este trabajo se plantea la utilización del coeficiente de correlación lineal de Pearson como criterio de evaluación de la proximidad entre conceptos.

	SNOMED-CT 2008AB	**SNOMED-CT 2010AB**
Physicians	0.5400	0.5451
Coders	0.7100	0.7170

Tabla 5.2 Valores de correlación de similitud semántica utilizando la métrica *Path* con relaciones PAR/CHD para las versiones de SNOMED-CT con UMLS 2008AB y 2010AB. Coeficiente de correlación de Pearson (resultados replicados del trabajo de Pedersen y McInnes para la versión 2008AB y en este trabajo para la versión 2010AB).

Los cálculos realizados con el coeficiente de correlación de Pearson, demuestran variaciones poco significativas (1-1.5%) en los resultados obtenidos entre las versiones de UMLS 2008AB y 2010AB (Tabla 5.2). Esta circunstancia demuestra que, según Spearman, pequeñas diferencias en los valores entre pares de conceptos afectan en el rango de ordenamiento, reflejando mayores variaciones en el resultado final de correlación.

Además, hay que tener en cuenta que el propio Pedersen que utiliza Spearman en sus resultados (Pedersen, Pakhomov, Patwardhan, & Chute, 2007), apunta a una máxima correlación Pearson de 0.85, entre las valoraciones de los expertos evaluadores (codificadores y facultativos). Este hecho nos ha llevado al estudio e interpretación de los

dos tipos de correlación empleados, para el análisis de las diversas métricas de similitud semántica.

Por ello y para una mejor interpretación, realizamos el cálculo de similitud para los 29 pares de conceptos (McInnes, Pedersen, & Pakhomov, 2009) con las principales métricas basadas en "*Path finding*", donde se puede observar que existen variaciones importantes en los resultados según el coeficiente de correlación empleado (Tabla 5.3).

Al igual que en trabajos anteriores (McInnes, Pedersen, & Pakhomov, 2009; Nguyen & Al-Mubaid, 2006), la métrica *nam*, aplicada sobre los recursos SNOMED-CT, refleja mejores valores de correlación para el coeficiente de Spearman. Sin embargo, la correlación Pearson para los valores estimados por las diferentes métricas, ofrece un resultado más aproximado para la métrica *Path* (Tabla 5.3).

Lejos de entrar en una discusión sobre el tipo de correlación que debe ser utilizado (Pearson correlaciona los valores de similitud, mientras que Spearman correlaciona su ordenación), este estudio revela que los resultados en muchos trabajos nuevamente no son comparables entre sí, como ya apuntaba Garla (Garla, & Brandt, 2012). Por este motivo y para clarificar la idea, se muestran los resultados con ambos coeficientes de correlación.

		Path	**lch**	**wup**	**nam**
Spearman	**Physicians**	0.2744	0.2744	0.3377	**0.4063**
	Coders	0.4160	0.4156	0.4190	**0.5578**
Pearson	**Physicians**	**0.5451**	0.3348	0.3372	0.4301
	Coders	**0.7170**	0.4566	0.3840	0.4456

Tabla 5.3 Valores de correlación de similitud semántica según Spearman (basado en medias) y Pearson para un conjunto de métricas basadas en "*Path finding*" con relaciones PAR/CHD para SNOMED-CT con UMLS 2010AB.

Al igual que en los trabajos anteriores (McInnes, Pedersen, & Pakhomov, 2009; (Pedersen, Pakhomov, Patwardhan, & Chute, 2007), la métrica *nam* es la que obtiene un valor de correlación más alto con el coeficiente de Spearman basado en medias. Además, todos los resultados obtenidos se ajustan más a los criterios de similitud semántica definidos por los codificadores (*coders*) que los definidos por los facultativos (*physicians*). Por este motivo, en las figuras siguientes (Figura 5.1 y Figura 5.2), se comparan los resultados obtenidos con los valores establecidos por dichos codificadores.

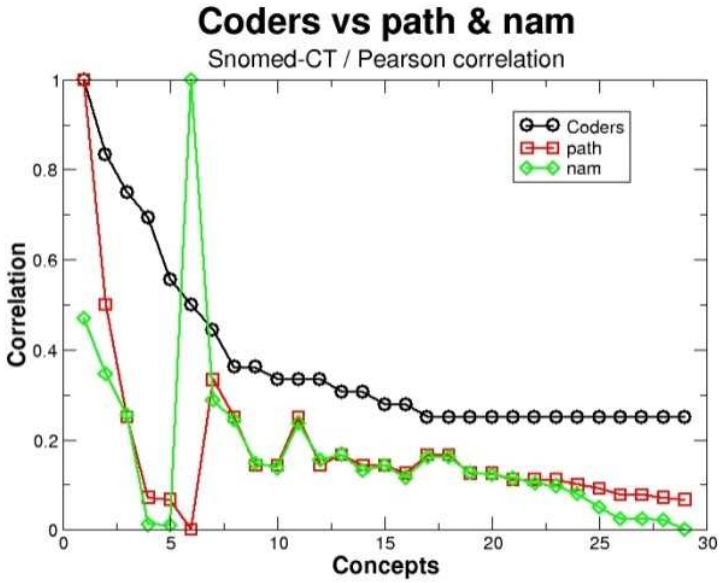

Figura 5.1 Valores de correlación de similitud semántica según Pearson para métricas *Path* y *nam* frente a los codificadores médicos (*coders*), basada en SNOMED-CT con UMLS 2010AB.

Con la utilización de la correlación de Pearson, la relevancia de la localización de los conceptos en la taxonomía y la profundidad de ésta para el cálculo de la similitud semántica entre términos (Figura 5.1), no se reflejan de una manera tan clara, como otros trabajos sugieren (Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Wu, & Palmer, 1994; Nguyen & Al-Mubaid, 2006).

5.3. Estudio e impacto de relaciones y recursos UMLS, en las métricas de similitud semántica

Los cálculos de similitud en los trabajos previos definidos por Pedersen (Pedersen, Pakhomov, Patwardhan, & Chute, 2007) y McInees (McInnes, Pedersen, & Pakhomov, 2009), fueron realiza-dos mediante relaciones jerárquicas directas de tipo (PAR-CHD), también definidas como relaciones semánticas de tipo "*is-a*", sobre un único recurso. Posteriores trabajos, Garla (Garla, & Brandt, 2012) y Batet (Batet, Sánchez, & Valls, 2011), no determinan el tipo de relaciones utilizadas en el cálculo de la similitud semántica, por lo que no se puede especificar que implicación pueden tener.

Por este motivo, en esta primera parte de este trabajo, también se ha evaluado el impacto que puede tener en el cálculo de la similitud semántica, el uso de diferentes tipos de relaciones entre conceptos. Los diferentes tipos de relaciones evaluadas son: las relaciones jerárquicas directas (PAR/CHD), las jerárquicas indirectas (RB/RN) y el resto de relaciones no jerárquicas existentes en el Metatesauro UMLS (SIB, RO, RL, SY, RQ, AQ y QB).

En primer lugar, se evalúa el impacto de añadir a las relaciones jerárquicas directas (PAR/CHD), otro tipo de relaciones como las jerárquicas indirectas (RB/RN), en el cálculo de la similitud para el *benchmark* de Pedersen con la métrica *Path*. En la Tabla 5.4 se muestran los nuevos valores de correlación obtenidos para la métrica *Path* (se reducen las métricas calculadas por el alto coste de computación requerido).

		Path
Spearman	**Physicians**	0.3393
	Coders	**0.4907**
Pearson	**Physicians**	0.5730
	Coders	**0.7438**

Tabla 5.4 Valores de correlación de similitud semántica para métrica *Path* con relaciones PAR/CHD y RB/RN para SNOMED-CT 2010AB.

Como se observa, los resultados obtenidos para la métrica *Path* ofrecen mejores valores de correlación con ambos tipos de relaciones jerárquicas. Esto es debido a la reducción en los caminos más cortos entre ciertos conceptos al introducir relaciones jerárquicas indirectas, como se explica en el siguiente ejemplo: Para los términos "*Carpal tunnel síndrome*" *C0007286* y "*Osteoarthritis*" *C0029408*, donde los expertos codificadores asignaron un valor 1.111 de similitud semántica entre ellos (0.2778 normalizado de 0 a 1), se mejoraron los resultados de la métrica *Path*, pasando de 0.1250 para (PAR/CHD) a 0.2000 para (PAR/CHD y RB/RN).

A continuación, se realizan los cálculos de similitud para el *benchmark* de Pedersen con la métrica *Path* aplicada al conjunto de recursos y a la totalidad de relaciones contenidas en UMLS 2010AB. (Más de 100 vocabularios controlados, entre ellos: MeSH, Snomed-CT).

Como se muestra en la Tabla 5.5, se observa una mejora significativa en los coeficientes de correlación para las relaciones jerárquicas. Sin embargo, el uso combinado de todas las relaciones (jerárquicas y no jerárquicas) degrada notablemente los resultados obtenidos. Esto se debe a que este tipo de relaciones no jerarquías, generan ciclos que no representan relaciones parentales o hermandad entre conceptos (sinonimia) (Bodenreider, 2001; Erdogan, Erdem, & Bodenreider, 2010), por lo que no recomendamos hacer uso de ellas, ya que añaden ruido en las relaciones. Además, se observa como la aplicación de la totalidad del conocimiento aportado por los recursos contenidos en el Metatesauro UMLS, muestra mejores resultados.

		PAR/CHD	PAR/CHD + RB/RN	TODAS
Spearman	**Phys.**	0.6382	0.5761	**0.4788**
	Coders	**0.6422**	**0.6495**	0.4338
Pearson	**Phys.**	0.7059	0.6740	0.6168
	Coders	**0.7982**	**0.8012**	**0.7046**

Tabla 5.5 Valores de correlación de similitud semántica para métrica *Path* con las relaciones (PAR/CHD), (PAR/CHD + RB/RN) y TODAS. Con todos los recursos existentes en UMLS 2010AB.

Las pruebas anteriores se realizaron también para la versión del Metatesauro UMLS 2011AB, obteniendo resultados semejantes.

Para las relaciones (PAR/CHD) y (PAR/CHD+RB/RN), los resultados muestran una mejora significativa en la correlación de los valores recogidos en la definición de la similitud semántica entre los 29 conceptos iniciales (Fig. 5.2). Se llega a obtener una

correlación muy alta (en ambos casos alrededor de 0.80) con respecto a los criterios de similitud semántica definidos por los codificadores. Es conveniente recordar que la correlación entre los valores establecidos por codificadores y facultativos es de 0.85.

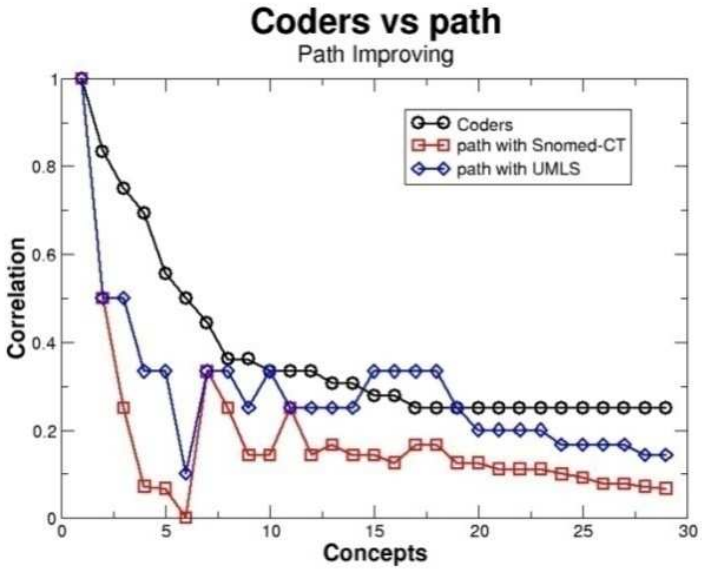

Figura 5.2 Mejora del valor de correlación de la métrica *Path* para los recursos SNOMED-CT y UMLS 2010AB en su totalidad, con relaciones PAR/CHD. Se toma como referencia el valor de correlación proporcionado por los codificadores.

5.4. Comparativa de las métricas *Path* vs *Intrinsic IC-Path*

Aunque son muchos los trabajos que han evaluado el rendimiento de las métricas con diferentes conjuntos de pares de conceptos, Garla aporta una visión definitiva sobre varios entornos teóricos de trabajo existentes y las diferentes métricas definidas (Garla, & Brandt, 2012). Como se observa en la Tabla 5.6 (resumen de los resultados concluidos por Garla), la métrica que ofrece mejores resultados en conjunto es *Intrinsic IC-Path*.

Benchmarks	Knowledge based		
	"Path Finding"		**"Intrinsic IC"**
	wup	***Path / lch***	***Path***
Pedersen Combined N=29	**0.70**	0.61	**0.70**
Mayo N=101	0.38	0.30	**0.41**
UMN relatedness N=430	0.33	**0.36**	**0.36**
UMN similarity N=566	0.39	0.40	**0.43**
UMN relatedness N=587	0.32	0.34	**0.35**

Tabla 5.6 Tabla resumida de resultados, extraída del trabajo de Garla para *concept graph* UMLS 2011AB (Garla, & Brandt, 2012).

Por este motivo, en este trabajo se evaluará el rendimiento de la métrica con mejores resultados (*Intrinsic IC-Path)* y la métrica computacionalmente más sencilla (*Path)*, en un contexto de recuperación de información real, es decir, trabajando con grandes volúmenes de información.

5.5. Conclusiones

En este capítulo se ha comprobado que los cambios en versiones del Metatesauro UMLS pueden influir en los resultados finales de las métricas de similitud semántica, debido a la inclusión de nueva información o correcciones de errores existentes en los cambios en versiones. (Cada año se distribuyen dos revisiones de la herramienta UMLS)

Se han replicado los diferentes trabajos del estado del arte bajo un marco común, realizando un estudio de las distintas métricas de similitud semántica aplicadas a un banco de pruebas teórico (Pedersen, Pakhomov, Patwardhan, & Chute, 2007), con pares de conceptos independientes y no interconectados. Los resultados obtenidos por las diferentes métricas, han sido comparados con los criterios de similitud aportados por los expertos. En dicho estudio y al igual que en los trabajos originales, se han utilizado las correlaciones basadas en el rango de ordenamiento de los resultados (Spearman). Pero además, se han estudiado las correlaciones centradas en la aproximación de valores (Pearson). Si bien las métricas correlacionan según el rango de ordenamiento de una forma muy similar a los estudios previos (en mismas condiciones de vocabularios) mostrando a la métrica "*nam"* como la más eficaz. Sin embargo, esos mismos resultados muestran a la métrica "*Path",* bajo una correlación de aproximación de valores, como la mejor métrica.

De esto se deduce que la métrica "*Path"*, tiene un mejor comportamiento de cara a ser aplicado en un sistema de recuperación basado en distancias semánticas. Ya que las aproximaciones de los valores de similitud, determinarán el mejor comportamiento de una métrica, en un contexto de recuperación de información.

Además, se ha puesto de manifiesto en este trabajo que el rendimiento de las métricas de similitud semántica será mayor, cuanto más amplia sea la cobertura de conceptos aportados por los recursos contenidos en el Metatesauro UMLS. Es decir, cuando el conocimiento que abarca el Metatesauro sea más extenso. Esto será válido, siempre y cuando se apliquen las correctas relaciones entre atributos que optimizan el rendimiento de las métricas de similitud semántica. Es decir, cuando se apliquen las relaciones jerárquicas entre conceptos en el cálculo de las distancias.

Por último, aunque en los trabajos finales aportados en otras investigaciones (Garla, & Brandt, 2012) sobre diferentes conjuntos de prueba teóricos y acotados, determinan que la métrica "*Intrinsic IC-Path"* ofrece mejores resultados frente a la métrica "*Path"*. Se observa que a medida que los conjuntos de prueba, contienen un mayor volumen de pares de conceptos de testeo, las correlaciones entre ambas métricas se aproximan más entre sí.

Capítulo 6

6. Evaluación de las métricas de similitud semántica en contexto real

Como se ha visto en el capítulo anterior, existen multitud de trabajos que han evaluado el comportamiento de las diferentes métricas de similitud semántica apoyadas en diferentes recursos del dominio biomédico, pero siempre con conjuntos de prueba cerrados y reducidos. Estos contextos teóricos, están formados por pares de conceptos preestablecidos y no interrelacionados entre sí.

Anteriormente, se ha mostrado la validez del marco de trabajo utilizado y la importancia de la correcta parametrización de este al trabajar con:

- Las últimas versiones del Metatesauro UMLS, ya que depuran errores e incorporan actualizaciones al conocimiento biomédico almacenado (cada año se distribuyen 2 actualizaciones de UMLS).
- Las relaciones jerárquicas directas e indirectas, que permiten obtener los resultados más ajustados a los juicios de proximidad semántica ofrecidos por los expertos. Reduciendo los ciclos entre pares de conceptos y por tanto se elimina parte del ruido de las relaciones entre ellos (Bodenreider, 2001; Erdogan, Erdem & Bodenreider, 2010; Caviedes & Cimino, 2004).

Por tanto, este capítulo de la tesis, se centrara en la aplicación de estas conclusiones en un estudio experimental sobre un entorno real de información biomédica. Para ello, y en contraste con los trabajos previos, se realizará una evaluación del impacto de las métricas *Path* e *Intrinsic IC-Path* en un contexto real de recuperación de información basado en registros médicos electrónicos contenidos en el TREC Medical Records Track 2011 (Voorhees, & Tong, 2011).

Para la consecución de esta evaluación, será necesario desarrollar un sistema de recuperación de información (R.I.) adecuado que permita la representación de cada uno de los documentos médicos (*reports*) que conforman el episodio médico de un paciente (*visit*), así como los criterios de búsqueda (*topics*). Estos se representarán mediante conceptos contenidos en UMLS. Esta representación, permitirá al sistema de recuperación,

relacionar semánticamente los conceptos de la consulta (*topic),* con los contenidos en cada documento médico (*report*). Para finalmente, determinar la relevancia del episodio médico de un paciente en función de la similitud semántica existente entre ellos.

6.1. Procesado previo de los documentos médicos y consultas del TREC - Medical Records Track

Como necesidad de representar, tratar y evaluar las métricas de similitud comentadas previamente, se deberán extraer los conceptos UMLS de los criterios de búsqueda (*topics*), así como de los documentos médicos (*reports*). Para el cálculo de las métricas de similitud semántica entre conceptos, se ha incorporado en el sistema de R.I. aquí propuesto, la herramienta Ytex desarrollada por Garla y Brandt (Garla, & Brandt, 2012). Estos resultados de similitud semántica entre *topic* y *report*, se agregarán en un único valor, el cual determinará la relevancia o no del documento (*report*) para dicha consulta (*topic*).

Pero como paso previo, al cálculo de la similitud semántica entre los conceptos del *report* y *topic*, será necesario un proceso de transformación del lenguaje natural que los representa, en unas estructuras basadas en conceptos. A continuación, se describe en detalle este proceso.

Pre-procesado de los documentos médicos y consultas

Cada documento médico (*report*) es almacenado en el repositorio del Text Retrieval Conference (TREC 2011) mediante un fichero con formato XML. Cada uno de estos ficheros conforma una estructura basada en etiquetas que aglutina su información de la siguiente manera:

- Versión xml y codificación
- <report>: Report
    - <checksum>: Identificador único del documento médico (*report*) y su asociación con la visita o episodio médico del paciente. La visita o episodio médico de un paciente, puede estar compuesta de entre 1 hasta 415 *reports.*
    - <subtype>: Subtipo de documento médico.
    - <type>: Tipo de documento médico.
    - <chief_complaint>: Motivo de la consulta.
    - <admit_diagnosis>: Diagnóstico de Admisión. (codificaciones ICD-9CM).
    - <discharge_diagnosis>: Diagnóstico de Alta. (codificaciones ICD-9CM).
    - <year>; <download_time>; <update_time>; <deid>: Etiquetas de fechas, control de cambios y versiones.
    - <report_text>: Informe médico en texto plano.

En este punto, se realiza un proceso de limpieza que consiste en eliminar todas las etiquetas, junto con la información de aquellas que contienen información no relevante para este estudio. Como es la versión xml; el identificador del documento, los códigos de diagnóstico para admisión y alta, así como las etiquetas de fechas y control de cambios. De esta manera se obtiene un nuevo fichero de texto plano, donde su principal información esta descrita en lenguaje natural. Evitando así, etiquetas, códigos y referencias que afectarían negativamente en su posterior procesado.

Conviene resaltar, que los códigos de diagnóstico (ICD9-CM) no han sido utilizados, a pesar de la su posible interés informacional. Ya que el objetivo principal de este trabajo, se centra en estudiar la viabilidad del sistema propuesto, en la estimación de la relevancia de un documento medico frente a una consulta, a partir de sus contenidos basados en texto. En posibles trabajos futuros se contempla la posibilidad de incluir esta codificación, la cual deberá ser tratada de una forma adicional.

Las consultas o *topics*, por otro lado, no requieren ningún tipo de pre-procesado ya que son básicamente cadenas de texto, con diferentes complejidades.

Procesado de las consultas

Cada consulta (*topic*), como se comenta en el punto anterior, Es una cadena de texto que refleja los criterios de evaluación de un usuario. Cada consulta es procesada mediante la correcta parametrización de la herramienta MetaMap. Esta herramienta analiza y descompone la cadena o frase original en cadenas sencillas denominadas *frases* que caracterizan síntomas, partes del cuerpo, enfermedades, etc. Para cada una de las *frases* creadas se obtienen los diferentes conceptos o CUIs que los representan. Como una frase o cadena se puede componer de varios términos y cada uno de ellos puede ser representado por distintos CUIs (como se describe en la sección 2.4), se combinan los CUIs de las *frases* entre sí dando lugar a varias *subfrases* por cada frase, realizando de esta manera una expansión de la consulta.

La herramienta MetaMap, aplicada en este proceso puede generar como resultado tanto un fichero interpretable por el usuario (*file.out*), como un fichero para su tratamiento informatizado (*file.xml*). La interpretación del fichero (*xml*), deberá ser procesada mediante un desarrollo Perl y diversos scripts (*bash-awk*) que darán lugar a una representación de la consulta basada en conceptos como se muestra en Tabla 6.1 y Tabla 6.2.

A modo de ejemplo, se describirá el procesamiento del *Topic 104* que define el criterio de búsqueda "*Patients diagnosed with localized prostate cáncer and treated with robotic surgery*". Las cadenas o frases en las que se descompone son:

1. *"Patients".*
2. *"diagnosed with localized prostate cancer".*
3. *"treated with robotic surgery".*

A continuación, para cada frase del *topic* se extraen los conceptos UMLS (CUIs) asociados, dando como resultado un conjunto de subfrases (Tabla 6.1). La generación de las subfrases, se obtiene a partir de la combinación de los distintos conceptos (CUIs) derivados de cada término de la frase inicial.

Por ejemplo, de la "frase 2"-> "*diagnosed with localized prostate cancer*", se extraen los términos independientes que conforman **dos subfrases contextuales**:

- "*diagnosed", "localized",* "*prostate cancer".*

- "*diagnosed", "localized cancer*", "*prostate".*

De cada una de estas subfrases contextuales, se realiza la expansión basada en conceptos UMLS (CUIs). Así, para la **primera subfrase contextual** se produce la siguiente expansión de sus términos:

- "***diagnosed***" se obtiene el CUI:
    - "Diagnosis" (C0011900)
- ***"localized***" se obtiene el CUI:
    - "Localized" (C0392752)
- "***prostate cáncer***" se obtienen tres diferentes CUIs:
    - "Malignant neoplasm of prostate" (C0376358)
    - "Prostate carcinoma" (C0600139)
    - "Prostate Cancer Pathway" (C2984325).

Para la **segunda subfrase contextual**, la expansión es la siguiente:

- "***diagnosed***" se obtiene el CUI:
    - "Diagnosis" (C0011900)
- ***"localized cancer***" se obtienen dos diferentes CUIs:
    - "Localized Malignant Neoplasm" (C0796563)
    - "Localized Carcinoma" (C1334407)
- ***"prostate***" se obtienen dos diferentes CUIs:
    - "Prostate" (C0033572)
    - "Entire prostate" (C1278980)

La combinación de todos los CUIs extraídos de los términos de las subfrases contextuales, dan como resultado las 11 subfrases que se muestran en la Tabla 6.1. Por ejemplo, la frase 2 da lugar a las siete subfrases definitivas (numeradas de la 1002 hasta la 1008), para la frase 1 ("*Patients")* sólo genera la subfrase 1001 y la frase 3 (*"treated with robotic surgery"*) genera las subfrases 1009, 1010 y 1011.

SUBFRASE	FRASE	**Topic 104:** ***"Patients diagnosed with localized prostate cancer and treated with robotic surgery"***
1001	1	CUI1= (C0030705) : podg : "Patients"
1002	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C0796563) : neop : "Localized Malignant Neoplasm" CUI3= (C0033572) : bpoc : "Prostate"
1003	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C0796563) :neop : "Localized Malignant Neoplasm" CUI3= (C1278980) :bpoc : "Entire prostate"
1004	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C1334407) : neop : "Localized Carcinoma" CUI3= (C0033572) : bpoc : "Prostate"
1005	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C1334407) : neop : "Localized Carcinoma" CUI3= (C1278980) : bpoc : "Entire prostate"
1006	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C0392752) : spco : "Localized" CUI3= (C0376358) : neop : "Malignant neoplasm of prostate"
1007	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C0392752) : spco : "Localized" CUI3= (C0600139) : neop : "Prostate carcinoma"
1008	2	CUI1= (C0011900) : fndg : "Diagnosis" CUI2= (C0392752) : spco : "Localized" CUI3= (C2984325) : ftcn : "Prostate Cancer Pathway"
1009	3	CUI1= (C0332293) : topp : " Treated with" CUI2= (C0035785) : ocdi : "Robotics" CUI3= (C0038894) : bmod : "Surgery specialty"
1010	3	CUI1= (C0332293) : topp : " Treated with" CUI2= (C0035785) : ocdi : "Robotics" CUI3= (C0038895) : ftcn : "Surgical aspects"
1011	3	CUI1= (C0332293) : topp : " Treated with" CUI2= (C0035785) : ocdi : "Robotics" CUI3= (C0543467) : diap : "Operative Surgical Procedures"

Tabla 6.1 Tabla de frases (Topic 104).

En el caso de procesar un criterio de búsqueda sencillo, por ejemplo el *Topic 101* "*Patients with hearing loss*", se generará una sola frase con las subfrases de conceptos 1001, 1002 y 1003, tal y como se muestra en la Tabla 6.2.

SUBFRASE	FRASE	**Topic 101:** ***"Patients with hearing loss"***
1001	1	CUI1= (C0030705) : podg : "Patients" CUI2= (C0011053) : dsyn : "Deafness"
1002	1	CUI1= (C0030705) : podg : "Patients" CUI2= (C0018772) : fndg: "Hearing Loss, Partial"
1003	1	CUI1= (C0030705) : podg : "Patients" CUI2= (C1384666) : fndg: "hearing impairment"

Tabla 6.2 Tabla de frases (Topic 101).

Procesado de los documentos médicos

Los documentos médicos (*reports*) son procesados de una manera idéntica a los criterios de búsqueda (*topics*), identificando los conceptos UMLS para cada frase del documento y generando cada una de las subfrases posibles a partir de las combinaciones de CUIs de sus diferentes frases contextuales.

Como ejemplo, se muestra un breve fragmento de un documento médico (*report90230*) el cual ha sido pre-procesado previamente para la generación de "frases" de esta fase (Figura 6.1). Estas "frases" son expandidas en diferentes "subfrases" a partir de las variaciones de conceptos (CUIs) que las representan (Tabla 6.3). De esta manera, se podrán combinar posteriormente con cada una de las subfrases que definen a la consulta (*topic*) enfrentándolas entre sí, para obtener la máxima proximidad semántica existente entre *topic* y *report* (esto se explica más detalladamente en la sección 6.3).

Conviene observar que aquellas frases que muestran la negación del correspondiente concepto, (asignación código 1), serán eliminadas del proceso del cálculo de similitud semántica. En ambos casos, consulta (*topic*) y documento (*report*), han sido expandidos conceptualmente a partir de las subfrases generadas en ambos procesos.

CONGESTIVE HEART FAILURE. CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE OF THE URINARY BLADDER WHICH CAUSES MASS EFFECT ON THE URINARY BLADDER AND ADJACENT TO UTERUS, DETECTED ON CT OF THE ABDOMEN. NO CHANGE IN 7 X 8 CM FOCAL CYSTIC STRUCTURE. HARD OF HEARING. IRON DEFICIENCY ANEMIA.

Figura 6.1 . Ejemplo sección o fragmento documento médico pre-procesado (*report90230*).

SUBFRASE	FRASE	Negación	Extracto *report90230*
190	254	0	C0018802 dsyn CONGESTIVE HEART FAILURE.
191	255	0	C0010709 dsyn CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
191	255	0	C0678594 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
191	255	0	C0456856 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
191	255	0	C0441987 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
192	255	0	C0010709 dsyn CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
192	255	0	C0678594 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
192	255	0	C0205095 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
192	255	0	C0205091 spco CYSTIC STRUCTURE AT THE POSTERIOR LEFT SIDE
193	256	0	C0577559 fndg MASS EFFECT ON THE URINARY BLADDER
193	256	0	C1280500 qlco MASS EFFECT ON THE URINARY BLADDER
193	256	0	C0005682 bpoc MASS EFFECT ON THE URINARY BLADDER
194	256	0	C0577559 fndg MASS EFFECT ON THE URINARY BLADDER
194	256	0	C2348382 qlco MASS EFFECT ON THE URINARY BLADDER
194	256	0	C0005682 bpoc MASS EFFECT ON THE URINARY BLADDER
195	256	0	C1280500 qlco MASS EFFECT ON THE URINARY BLADDER
195	256	0	C0042027 bpoc MASS EFFECT ON THE URINARY BLADDER
195	256	0	C0238775 fndg MASS EFFECT ON THE URINARY BLADDER
196	256	0	C1280500 qlco MASS EFFECT ON THE URINARY BLADDER
196	256	0	C1524119 qlco MASS EFFECT ON THE URINARY BLADDER
196	256	0	C0238775 fndg MASS EFFECT ON THE URINARY BLADDER
197	256	0	C2348382 qlco MASS EFFECT ON THE URINARY BLADDER
197	256	0	C0042027 bpoc MASS EFFECT ON THE URINARY BLADDER
197	256	0	C0238775 fndg MASS EFFECT ON THE URINARY BLADDER
198	256	0	C2348382 qlco MASS EFFECT ON THE URINARY BLADDER
198	256	0	C1524119 qlco MASS EFFECT ON THE URINARY BLADDER
198	256	0	C0238775 fndg MASS EFFECT ON THE URINARY BLADDER
199	257	0	C0442726 fndg DETECTED ON CT
200	258	**1**	C0205234 spco NO CHANGE IN 7 X 8 CM FOCAL CYSTIC STRUCTURE
200	258	**1**	C1511605 fndg NO CHANGE IN 7 X 8 CM FOCAL CYSTIC STRUCTURE
200	258	**1**	C0678594 spco NO CHANGE IN 7 X 8 CM FOCAL CYSTIC STRUCTURE
201	259	0	C0018772 fndg HARD OF HEARING
202	259	0	C1384666 fndg HARD OF HEARING
203	260	0	C0162316 dsyn IRON DEFICIENCY ANEMIA.

Tabla 6.3 Ejemplo sección o fragmento documento médico procesado (*report90230*).

6.2. Filtrado por tipos semánticos del topic

La expansión de la consulta realizada en el punto anterior, potencia el proceso de recuperación de información al reflejar nuevas relaciones entre las distintas variaciones de conceptos. Sin embargo, esta expansión puede dar lugar a relaciones entre conceptos pertenecientes a tipos semánticos con poca "especialización" o "particularidad" semántica (Bodenreider, & McCray, 2003; Bodenreider, 2001; Erdogan, Erdem, & Bodenreider, 2010; Plaza, & Díaz, 2010). Estas relaciones pueden desvirtuar los resultados finales de similitud para aquellos conceptos de mayor relevancia semántica en el contexto que aquí se trata.

Por este motivo, se ha procedido a realizar una clasificación de los tipos semánticos según su importancia en, "tipos genéricos" y "tipos específicos". Los tipos semánticos específicos agrupan conceptos que revelan una mayor importancia en este dominio, como por ejemplo: enfermedades, síntomas, procedimientos y medicamentos. A continuación, se muestran los tipos semánticos que aparecen en los criterios de búsqueda tomados como ejemplo (Topics 104 y 101):

- Genéricos:
spco - Spatial Concept (CONC)
podg - Patient or Disabled Group (LIVB)
ftcn - Functional Concept (CONC)

- Específicos:
dsyn - Disease or Syndrome (DISO)
diap - Diagnostic Procedure(PROC)
neop - Neoplastic Process (DISO)
fndg - Finding (DISO)
bpoc - Body Part, Organ, or Org.Component (ANAT)
topp - Thrapeutic or Preventive Procedure (PROC)
bmod - Biomedical Occupation or Discipline (OCCU)
ocdi - Occupation or Discipline (OCCU)

Aquellos conceptos asociados a tipos semánticos genéricos serán eliminados de la tabla de frases generada en el punto anterior (6.1 Procesado de la consulta). Por ejemplo, en la Tabla 6.1 y Tabla 6.2, se muestran en gris aquellos conceptos genéricos que han sido eliminados, para el Topic 104 y 101. Se identifican de esta manera, todos aquellos conceptos que no son de relevancia en el contexto de la frase.

Como ejemplo, nótese para el Topic 104, como:

- La frase 1, será completamente eliminada. Esta frase está conformada por una única variación de frase, donde su único concepto es genérico.
- Para la frase 2, se eliminarán los conceptos correspondientes al concepto "Localized" para las subfrases 1006, 1007 y 1008, por pertenecer a tipo genérico semántico "*spco*". Además también, en la subfrase 1008 será eliminado el concepto asociado a "Prostate Cancer Pathway", por pertenecer al tipo "*ftcn*".
- En frase 3, se elimina el concepto "Surgical aspect" para la subfrase 1010, por pertenecer al tipo "*ftcn*".

A continuación se muestran los Grupos semánticos del UMLS, donde para cada grupo, se han clasificado sus tipos semánticos en función de sus características, en **Genéricos** o **Específicos**.

LIVB Living Beings:

- **Tipos semánticos Genéricos.**
  - T001 orgm Organism LIVB Living Beings
  - T002 Plant LIVB Living Beings
  - T004 fngs Fungus LIVB Living Beings
  - T008 anim Animal LIVB Living Beings
  - T010 vtbt Vertebrate LIVB Living Beings
  - T011 amph Amphibian LIVB Living Beings
  - T012 bird Bird LIVB Living Beings
  - T013 fish Fish LIVB Living Beings
  - T014 rept Reptile LIVB Living Beings
  - T015 mamm Mammal LIVB Living Beings
  - T016 humn Human LIVB Living Beings
  - T096 grup Group LIVB Living Beings
  - T097 prog Professional or Occupational Group LIVB Living Beings
  - T099 famg Family Group LIVB Living Beings
  - T101 podg Patient or Disabled Group LIVB Living Beings
  - T194 arch Archaeon LIVB Living Beings

- **Tipos semánticos Específicos.**
  - T005 virs Virus LIVB Living Beings
  - T007 bact Bacterium LIVB Living Beings
  - T098 popg Population Group LIVB Living Beings
  - T100 aggp Age Group LIVB Living Beings
  - T204 euka Eukaryote LIVB Living Beings

ANAT Anatomy:

- **Tipos semánticos Genéricos.**
  - T018 emst Embryonic Structure ANAT Anatomy
  - T021 ffas Fully Formed Anatomical Structure ANAT Anatomy
  - T022 bdsy Body System ANAT Anatomy
  - T025 cell ANAT Anatomy
  - T026 celc Cell Component ANAT Anatomy

- **Tipos semánticos Específicos.**
  - T017 anst Anatomical Structure ANAT Anatomy
  - T023 bpoc Body Part, Organ, or Organ Component ANAT Anatomy
  - T029 blor Body Location or Region ANAT Anatomy
  - T030 bsoj Body Space or Junction ANAT Anatomy
  - T031 bdsu Body Substance ANAT Anatomy
  - T024 tisu Tissue ANAT Anatomy

DISO Disorders

- **Tipos semánticos Genéricos.**
  - -
- **Tipos semánticos Específicos.**
  - T019 cgab Congenital Abnormality DISO Disorders
  - T020 acab Acquired Abnormality DISO Disorders
  - T033 fndg Finding DISO Disorders
  - T037 inpo Injury or Poisoning DISO Disorders
  - T046 patf Pathologic Function DISO Disorders
  - T047 dsyn Disease or Syndrome DISO Disorders
  - T048 mobd Mental or Behavioral Dysfunction DISO Disorders
  - T049 comd Cell or Molecular Dysfunction DISO Disorders
  - T050 emod Experimental Model of Disease DISO Disorders
  - T184 sosy Sign or Symptom DISO Disorders
  - T190 anab Anatomical Abnormality DISO Disorders
  - T191 neop Neoplastic Process DISO Disorders

GENE Genes & Molecular Sequences

- **Tipos semánticos Genéricos.**
  T085 mosq Molecular Sequence GENE Genes & Molecular Sequences
  T086 nusq Nucleotide Sequence GENE Genes & Molecular Sequences
  T087 amas Amino Acid Sequence GENE Genes & Molecular Sequences
  T088 crbs Carbohydrate Sequence GENE Genes & Molecular Sequences

- **Tipos semánticos Específicos.**
  T028 gngm Gene or Genome GENE Genes & Molecular Sequences

PHYS Physiology

- **Tipos semánticos Genéricos.**
  T032 orga Organism Attribute PHYS Physiology
  T043 celf Cell Function PHYS Physiology
  T044 moft Molecular Function PHYS Physiology
  T045 genf Genetic Function PHYS Physiology
  T201 clna Clinical Attribute PHYS Physiology

- **Tipos semánticos Específicos.**
  T039 phsf Physiologic Function PHYS Physiology
  T040 orgf Organism Function PHYS Physiology
  T041 menp Mental Process PHYS Physiology
  T042 ortf Organ or Tissue Function PHYS Physiology

PHEN Phenomena

- **Tipos semánticos Genéricos.**
  T067 phpr Phenomenon or Process PHEN Phenomena
  T068 hcpp Human-caused Phenomenon or Process PHEN Phenomena
  T069 eehu Environmental Effect of Humans PHEN Phenomena
  T070 npop Natural Phenomenon or Process PHEN Phenomena

- **Tipos semánticos Específicos.**
  T034 lbtr Laboratory or Test Result PHEN Phenomena
  T038 biof Biologic Function PHEN Phenomena

ACTI Activities & Behaviors

- **Tipos semánticos Genéricos.**
  T051 evnt Event ACTI Activities & Behaviors
  T052 acty Activity ACTI Activities & Behaviors
  T053 bhvr Behavior ACTI Activities & Behaviors
  T054 socb Social Behavior ACTI Activities & Behaviors
  T055 inbe Individual Behavior ACTI Activities & Behaviors
  T056 dora Daily or Recreational Activity ACTI Activities & Behaviors
  T057 ocac Occupational Activity ACTI Activities & Behaviors
  T064 gora Governmental or Regulatory Activity ACTI Activities & Behaviors
  T066 mcha Machine Activity ACTI Activities & Behaviors

- **Tipos semánticos Específicos.**
  -

PROC Procedures

- **Tipos semánticos Genéricos.**
  T062 resa Research Activity PROC Procedures
  T063 mbrt Molecular Biology Research Technique PROC Procedures
  T065 edac Educational Activity PROC Procedures

- **Tipos semánticos Específicos.**
  T058 hlca Health Care Activity PROC Procedures
  T059 lbpr Laboratory Procedure PROC Procedures
  T060 diap Diagnostic Procedure PROC Procedures
  T061 topp Therapeutic or Preventive Procedure PROC Procedures

OBJC Objects

- **Tipos semánticos Genéricos.**
  T071 enty Entity OBJC Objects
  T072 phob Physical Object OBJC Objects
  T167 sbst Substance OBJC Objects
  T168 food Food OBJC Objects

- **Tipos semánticos Específicos.**
  T073 mnob Manufactured Object OBJC Objects

DEVI Devices

- **Tipos semánticos Genéricos.**
  T075 resd Research Device DEVI Devices

- **Tipos semánticos Específicos.**
  T074 medd Medical Device DEVI Devices
  T203 drdd Drug Delivery Device DEVI Devices

CONC Concepts & Ideas

- **Tipos semánticos Genéricos.**
  T077 cnce Conceptual Entity CONC Concepts & Ideas
  T078 idcn Idea or Concept CONC Concepts & Ideas
  T080 qlco Qualitative Concept CONC Concepts & Ideas
  T081 qnco Quantitative Concept CONC Concepts & Ideas
  T082 spco Spatial Concept CONC Concepts & Ideas
  T089 rnlw Regulation or Law CONC Concepts & Ideas
  T102 grpa Group Attribute CONC Concepts & Ideas
  T169 ftcn Functional Concept CONC Concepts & Ideas
  T170 inpr Intellectual Product CONC Concepts & Ideas
  T171 lang Language CONC Concepts & Ideas
  T185 clas Classification CONC Concepts & Ideas
  T079 tmco Temporal Concept CONC Concepts & Ideas

- **Tipos semánticos Específicos.**

GEOG Geographic Areas

- **Tipos semánticos Genéricos.**
  T083 geoa Geographic Area GEOG Geographic Areas

- **Tipos semánticos Específicos.**
  -

OCCU Occupations

- **Tipos semánticos Genéricos.**
  -

- **Tipos semánticos Específicos.**
  T090 ocdi Occupation or Discipline OCCU Occupations
  T091 bmod Biomedical Occupation or Discipline OCCU Occupations

ORGA Organizations

- **Tipos semánticos Genéricos.**
  - T094 pros Professional Society ORGA Organizations
  - T095 shro Self-help or Relief Organization ORGA Organizations

- **Tipos semánticos Específicos.**
  - T093 hcro Health Care Related Organization ORGA Organizations
  - T092 orgt Organization ORGA Organizations

CHEM Chemicals & Drugs

- **Tipos semánticos Genéricos.**
  - T104 chvs Chemical Viewed Structurally CHEM Chemicals & Drugs
  - T110 strd Steroid CHEM Chemicals & Drugs
  - T111 eico Eicosanoid CHEM Chemicals & Drugs
  - T114 nnon Nucleic Acid, Nucleoside, or Nucleotide CHEM Chemicals & Drugs
  - T115 opco Organophosphorus Compound CHEM Chemicals & Drugs
  - T116 aapp Amino Acid, Peptide, or Protein CHEM Chemicals & Drugs
  - T118 carb Carbohydrate CHEM Chemicals & Drugs
  - T119 lipd Lipid CHEM Chemicals & Drugs
  - T120 chvf Chemical Viewed Functionally CHEM Chemicals & Drugs
  - T122 bodm Biomedical or Dental Material CHEM Chemicals & Drugs
  - T123 bacs Biologically Active Substance CHEM Chemicals & Drugs
  - T124 nsba Neuroreactive Substance or Biogenic Amine CHEM Chemicals & Drugs
  - T130 irda Indicator, Reagent, or Diagnostic Aid CHEM Chemicals & Drugs
  - T131 hops Hazardous or Poisonous Substance CHEM Chemicals & Drugs
  - T192 rcpt Receptor CHEM Chemicals & Drugs
  - T196 elii Element, Ion, or Isotope CHEM Chemicals & Drugs
  - T197 inch Inorganic Chemical CHEM Chemicals & Drugs
  - T200 clnd Clinical Drug CHEM Chemicals & Drugs

- **Tipos semánticos Específicos.**
  - T103 chem Chemical CHEM Chemicals & Drugs
  - T109 orch Organic Chemical CHEM Chemicals & Drugs
  - T125 horm Hormone CHEM Chemicals & Drugs
  - T126 enzy Enzyme CHEM Chemicals & Drugs
  - T127 vita Vitamin CHEM Chemicals & Drugs
  - T129 imft Immunologic Factor CHEM Chemicals & Drugs
  - T195 antb Antibiotic CHEM Chemicals & Drugs
  - T196 elii Element, Ion, or Isotope CHEM Chemicals & Drugs
  - T197 inch Inorganic Chemical CHEM Chemicals & Drugs
  - T200 clnd Clinical Drug CHEM Chemicals & Drugs
  - T121 phsu Pharmacologic Substance CHEM Chemicals & Drugs

6.3. Matriz de Máxima Similitud y Cálculo de Relevancia

Con el fin de valorar la similitud semántica entre el la consulta (*topic*) y cada documento médico (*report*), se realiza una evaluación de su similitud a diferentes niveles de agregación: CUIs, subfrases y frases. Los cálculos de similitud (Sim) se consiguen gracias a una matriz que empareja los CUIs de la consulta (*topic*) frente a los CUIs del documento médico (*report*), para las dos métricas que han sido elegidas: Path e Intrinsic IC-Path. Seguidamente, se seleccionan para cada CUI de la subfrase de la consulta (*topic*), aquellos pares de conceptos (*topic-report*) con el mayor valor de similitud. Este proceso se repite en la consulta, para todas las subfrases de una frase.

$$Sim_{subphrase\ _{subphrase_i\ cui_j}} = max\left(Sim\left(CUI_{subphrase_{ij}} CUI_{report_k}\right)\right) \tag{6.1}$$

Siendo "i" cada una de las subfrases del topic, "j" cada uno de los CUIs de la subfrase y "k" cada uno de los CUIs del report.

A continuación, para cada frase individual de la consulta (*topic*), se elige el valor de máxima similitud de cada CUI que existe en sus sub-frases. En el caso particular del Topic 104 y la frase 2 (Tabla 6.1) se obtendrá el máximo valor de similitud del CUI1, CUI2 y CUI3.

$$Sim\_max\_phrase_{cui_j} = max\left(Sim\_subphrase_{subphrase_i\ cui_j}\right) \tag{6.2}$$

Hecho esto, seguidamente se calculará el valor medio de todos ellos, obteniendo un único valor de similitud por frase. En el ejemplo para el Topic 104 y frase 2, Sim_avg_phrase=CUI1+CUI2+CUI3/3.

$$Sim\_avg\_phrase_i = \sum_{i=0}^{num\_cuis\_phrase} (Sim\_max\_phrase_i / num\_cuis\_phrase) \tag{6.3}$$

Por último, se realiza una media aritmética de los valores de máxima similitud de cada una de las frases que conforman la búsqueda, lo que otorgará la relevancia final del documento médico (*report*) con respecto a la consulta (*topic*). En el ejemplo del Topic 104, *$Sim\_topic_{vs}report$ = (Sim_avg_phrase1 + Sim_avg_phrase2 + Sim_avg_phrase3)/3.*

Es interesante observar que en el caso concreto del Topic 104, el valor final de relevancia vendrá determinado por el valor medio de similitud de las dos últimas frases. La frase 1 *(Phrase 1 – "Patient")* es completamente eliminada del resultado, ya que todos los conceptos (CUIs) que la conforman, están asociados a tipos semánticos genéricos (*podg*).

$$Sim\_topic_{vs}report = \sum_{i=0}^{num\_phrases} (Sim\_avg\_phrase_i / num\_phrases) \quad (6.4)$$

Por tanto, podemos decir que el valor final ($Sim\_topic_{vs}report$), de la matriz de máxima similitud de un documento médico (*report*) con respecto a una consulta (*topic*), determinará si este es relevante o no, para las características de dicha consulta. Este valor se encontrará en el rango [0,1], reflejando la relevancia del documento médico para una cierta consulta. El extremo inferior (valor 0), indica la máxima "no relevancia" y el extremo superior (valor 1), la máxima "relevancia".

Para poder comparar el valor final obtenido por la matriz de similitud semántica, con los criterios de relevancia otorgados por los expertos evaluadores en cada caso. Será necesario establecer un valor de corte indicativo (dentro del rango [0,1]), el cual permitirá determinar cuándo un documento médico es relevante o no, para una consulta. Esto será estudiado y definido en la siguiente sección (6.4.3).

Como un episodio médico de un paciente (*visit*) puede estar conformado por más de un de un documento médico (*report*), la relevancia del episodio vendrá determinado por el valor máximo de relevancia de sus documentos (*reports*).

Este método pretende preservar la originalidad y completitud informacional de la consulta (*topic*), para su tratamiento automatizado sin intervención por parte del usuario. Por ello es necesario, como se describe en el proceso anterior, la inclusión de cada uno de los componentes de la consulta (*topic*), mediante un proceso de agregación de la media de máximos valores de similitud de las distintas frases. De esta forma cada subfrase, que es expandida a partir de las frases que componen la consulta, son medidas con la misma precisión al realizar la agregación de sus medias. Sin embargo, lo que determinará la relevancia de cada componente, será la máxima similitud semántica de los conceptos de la consulta (*topic*) frente al documento médico (*report*), así como el tipo semántico al que pertenezcan.

Topic 101	Report90230	Path Max. Sim	IC-Path Max. Sim
1001 1 C0030705 podg Patients	168 52 C0030705 podg the patient on consultation	1.0000	1.0000
1001 1 C0011053 dsyn Deafness	201 73 C0018772 fndg HARD OF HEARING.	0.5000	0.8042
1002 1 C0030705 podg Patients	113 11 C0030705 podg the patient in consultation	1.0000	1.0000
1002 1 C0018772 fndg Hearing Loss, Partial	201 73 C0018772 fndg HARD OF HEARING.	1.0000	1.0000
1003 1 C0030705 podg Patients	111 9 C0030705 podg The patient apparently	1.0000	1.0000
1003 1 C1384666 fndg hearing impairment	202 73 C1384666 fndg HARD OF HEARING.	1.0000	1.0000

Tabla 6.4 . Ejemplo matriz de máximos valores de similitud de cada subfrase: "Sim_subphrase" del "topic101 vs Report90230".

Se puede observar en este sencillo ejemplo (Tabla 6.4), la importancia de la expansión basada en conceptos tanto de la consulta (*topic*), como del documento (*report*). Entre los que se obtienen, las relaciones de máxima similitud de sus conceptos, aunque los términos

o cadenas sean diferentes entre sí. Así por ejemplo, los términos asociados a los CUIs de la consulta ("*Deafness*"; "*Hearing Loss, Partial*"; "*hearing impairment*"), son diferentes a los términos asociados a los CUIs del documento médico ("*HARD OF HEARING*"), pero se obtiene la máxima similitud posible.

Las tablas 6.4 y 6.5, se componen de los siguientes elementos:

- Las dos primeras columnas (*topic* y *report)*, las forman sus "subphrases", "phrases", "CUIs" y "String phrases".
- Las dos últimas columnas corresponden a las máximas similitudes para cada métrica entre pares de CUIs *topic-report*.

Topic 104	Report51139	Path Max. Sim	IC-Path Max. Sim
1001 1 C0030705 podg Patients	43 15 C0030705 podg he patient	1.0000	1.0000
1002 2 C0011900 fndg Diagnosis	75 28 C0543467 diap DESCRIPTION OF OPERATION	0.3333	0.7172
1002 2 C0796563 neop Localized Malignant Neoplasm	28 12 C0796563 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1002 2 C0033572 bpoc Prostate	61 18 C0033572 bpoc now for removal of his prostate	1.0000	1.0000
1003 2 C0011900 fndg Diagnosis	40 14 C0376358 neop LOCALIZED PROSTATE CANCER.	0.3333	0.7172
1003 2 C0796563 neop Localized Malignant Neoplasm	28 12 C0796563 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1003 2 C1278980 bpoc Entire prostate	380 19 C1278980 bpoc The prostate	1.0000	1.0000
1004 2 C0011900 fndg Diagnosis	40 14 C0376358 neop LOCALIZED PROSTATE CANCER.	0.3333	0.7172
1004 2 C1334407 neop Localized Carcinoma	30 12 C1334407 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1004 2 C0033572 bpoc Prostate	171 75 C0033572 bpoc at the prostate.	1.0000	1.0000
1005 2 C0011900 fndg Diagnosis	63 19 C0184661 diap benefits of the procedure	0.3333	0.7172
1005 2 C1334407 neop Localized Carcinoma	30 12 C1334407 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1005 2 C1278980 bpoc Entire prostate	236 113 C1278980 bpoc sharp dissection until the prostate	1.0000	1.0000
1006 2 C0011900 fndg Diagnosis	15 7 C0184661 diap PROCEDURE	0.3333	0.7172
1006 2 C0392752 spco Localized	53 16 C0392752 spco 50s-year-old male with localized adenocarcinoma of	1.0000	1.0000
1006 2 C0376358 neop Malignant neoplasm of prostate	32 12 C0376358 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1007 2 C0011900 fndg Diagnosis	14 6 C0543467 diap SURGERY DATE	0.3333	0.7172
1007 2 C0392752 spco Localized	44 16 C0392752 spco 50s-year-old male with localized adenocarcinoma of	1.0000	1.0000
1007 2 C0600139 neop Prostate carcinoma	33 12 C0600139 neop LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1008 2 C0011900 fndg Diagnosis	32 12 C0376358 neop LOCALIZED PROSTATE CANCER.	0.3333	0.7172
1008 2 C0392752 spco Localized	44 16 C0392752 spco 50s-year-old male with localized adenocarcinoma of	1.0000	1.0000
1008 2 C2984325 ftcn Prostate Cancer Pathway	42 14 C2984325 ftcn LOCALIZED PROSTATE CANCER.	1.0000	1.0000
1009 3 C0332293 topp Treated with	523 24 C0444667 qnco present for the entire procedure.	0.0000	0.0000
1009 3 C0035785 ocdi Robotics	17 8 C0035785 ocdi ROBOTIC-ASSISTED LAPAROSCOPIC RADICAL PROSTATECTOMY	1.0000	1.0000
1009 3 C0038894 bmod Surgery specialty	9 6 C0038894 bmod SURGERY DATE	1.0000	1.0000
1010 3 C0332293 topp Treated with	522 24 C0450011 topp present for the entire procedure.	0.0000	0.0000
1010 3 C0035785 ocdi Robotics	19 8 C0035785 ocdi ROBOTIC-ASSISTED LAPAROSCOPIC RADICAL PROSTATECTOMY	1.0000	1.0000
1010 3 C0038895 ftcn Surgical aspects	11 6 C0038895 ftcn SURGERY DATE	1.0000	1.0000
1011 3 C0332293 topp Treated with	522 24 C0450011 topp present for the entire procedure.	0.0000	0.0000
1011 3 C0035785 ocdi Robotics	19 8 C0035785 ocdi ROBOTIC-ASSISTED LAPAROSCOPIC RADICAL PROSTATECTOMY	1.0000	1.0000
1011 3 C0543467 diap Operative Surgical Procedures	75 28 C0543467 diap DESCRIPTION OF OPERATION	1.0000	1.0000

Tabla 6.5 Ejemplo matriz de máximos valores de similitud de cada subfrase: "Sim_subphrase" del "topic104 vs report51139".

6.4. Análisis previo de los Resultados

En esta sección, se hace un análisis previo de los resultados obtenidos después de la evaluación de las consultas (*topics*) frente a los documentos médicos (*reports*) según el procedimiento descrito en la sección anterior (6.3).

Para contrastar la opinión de relevancia otorgada por los expertos y los resultados del sistema de recuperación propuesto en este trabajo, se generan unos histogramas (Figura 6.2) que reflejen la similitud de cada episodio médico o *visita* (correspondiendo con el *report* de mayor similitud de todos los que conforman la *visita*) frente a una consulta. Estos documentos médicos (*reports*) son distribuidos en el eje X según su grado de relevancia (0 siendo "No Relevante", y 1 si es "Relevante"). Por último, para facilitar la comprensión de las figuras de esta sección, se destacan en negro aquellos documento médicos (*reports*) que fueron valorados por los expertos como "Relevantes" y en ocre los "No Relevantes".

6.4.1. Justificación del filtrado por tipos semánticos

En primer lugar, se han realizado un conjunto de experimentos que validen el filtrado por los conceptos asociados a los tipos semánticos "específicos" de una consulta (*topic*). Así en la Figura 6.2, se muestra los resultados de la evaluación de los documentos médicos (*reports*) enfrentados al Topic 107 ("*Patients with ductal carcinoma in situ (DCIS)*") filtrando por tipos semánticos (Figura 6.2b) y sin filtrar (Figura 6.2a). Se observa fácilmente, cómo tras el filtrado, aquellos documentos médicos (*reports*) "No Relevantes" (ocre) y "Relevantes" (negro) más significativos, sufren un desplazamiento hacia zonas de menor y mayor relevancia respectivamente.

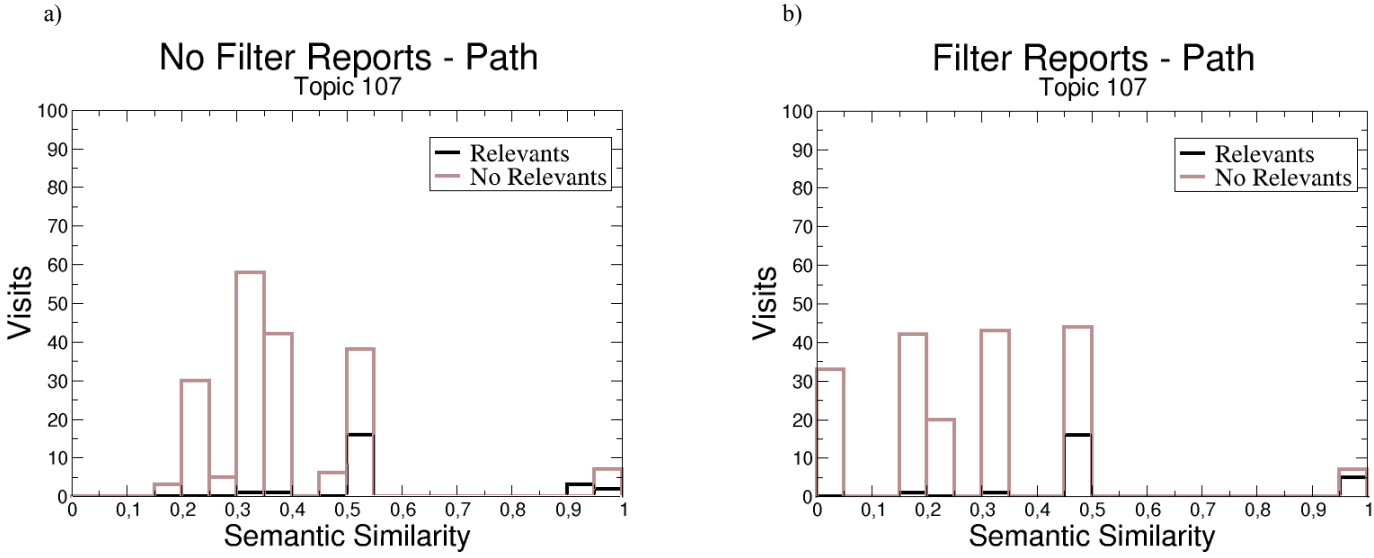

Figura 6.2 **a) Histograma para Topic 107 sin filtrar por tipos semánticos.**
b) Histograma para Topic 107 filtrado por tipos semánticos.

Estos resultados demuestran la necesidad de aplicar la expansión de la consulta únicamente por tipos semánticos "específicos", con el fin de obtener resultados más precisos y de menor coste computacional (eliminando cálculos de similitud para los tipos semánticos "genéricos").

6.4.2. Comportamiento de las métricas Path e Intrinsic IC-Path

Para evaluar comparativamente el rendimiento (en términos de similitud semántica) de las métricas *Path* e *Intrinsic IC-Path* en un contexto real, se muestra un experimento preliminar sobre dos criterios de búsqueda. Uno es un consulta sencilla, *Topic 101* ("*Patients with hearing loss*"), aplicado sobre 4073 *reports* agrupados en 249 *visitas* y el otro una consulta compleja, *Topic 104* ("*Patients diagnosed with localized prostate cáncer and treated with robotic surgery*"), aplicado sobre 3439 *reports* agrupados en 196 *visitas*.

Los resultados obtenidos al aplicar la métrica *Path* sobre una consulta sencilla (*Topic 101*), muestran una distribución discreta de los resultados, derivada de su definición basada en el inverso de las distancias (Fig. 6.3a). Este hecho, provoca que se generen zonas de incertidumbre, ya que se localizan documentos médicos (*reports*) con valores de similitud entre 0.45 y 0.50 (27 reports no relevantes y 9 relevantes).

En el caso de la métrica *Intrinsic IC-Path*, la naturaleza interna de su cálculo elimina dicho carácter discreto (Fig. 6.3b). Los resultados globales respecto a la métrica *Path* son similares pero distribuidos de una manera más suavizada y continua hacia los extremos.

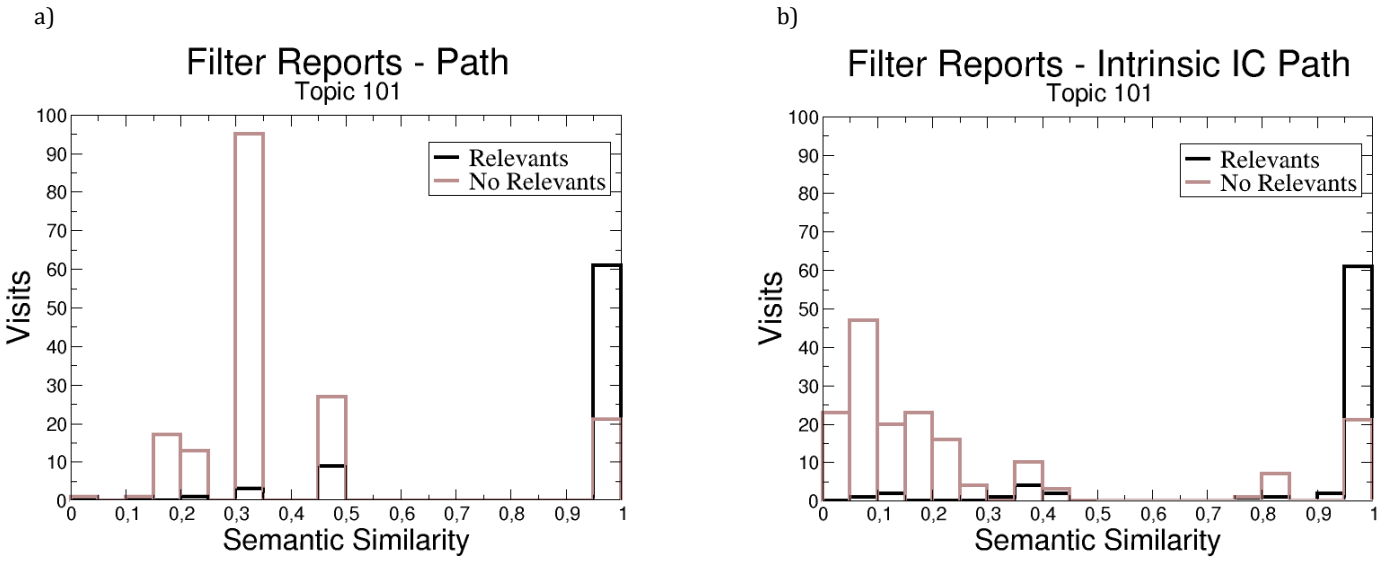

Figura 6.3 *Path* vs *Intrinsic IC-Path* para un tema de búsqueda sencillo (Topic 101).

Por el contrario, al procesar consultas complejas (múltiples frases) como es el *Topic 104*, los cálculos basados en las medias agregadas sobre los valores de máxima similitud obtenidos (sección 6.3), hace que se pierda el carácter discreto de la métrica *Path*. Además, para ambas métricas, los valores de similitud de los documentos médicos (*reports*) tienden a repartirse según una función de distribución normal (Figuras 6.4, 6.4b), desapareciendo las diferencias comentadas anteriormente.

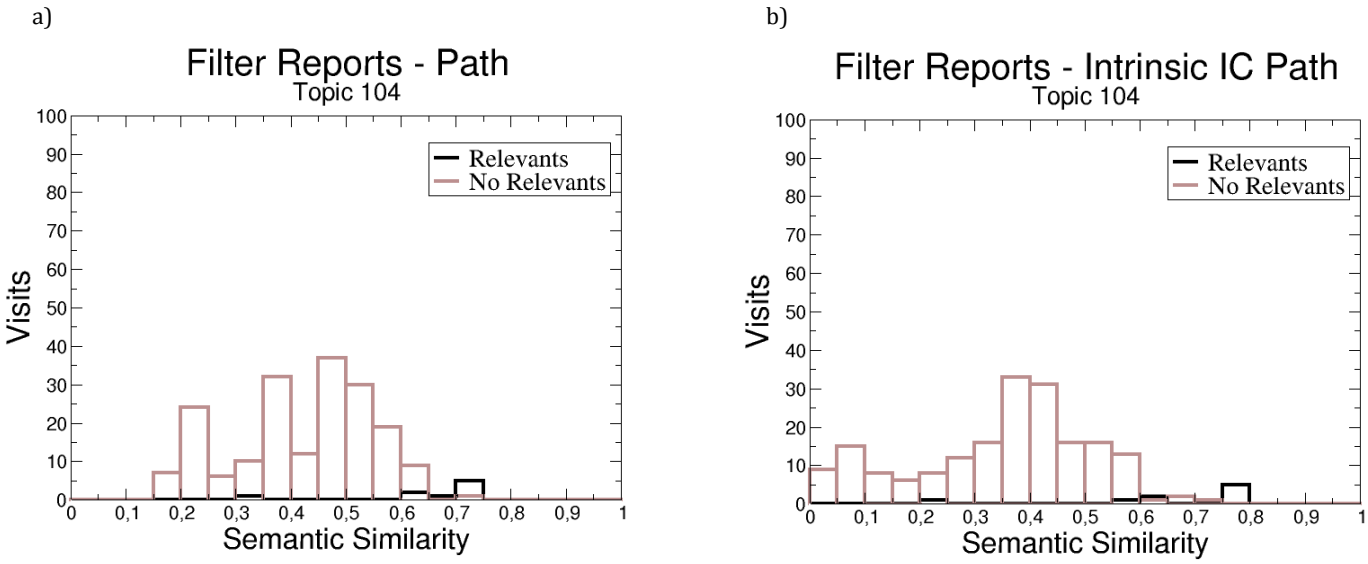

Figura 6.4 *Path* vs *Intrinsic IC-Path* para un tema de búsqueda complejo (Topic 104).

6.4.3. Valor de corte

A partir de las distribuciones de similitud de los documentos médicos (*reports*) generadas para cada criterio de búsqueda, como se muestra en el apartado anterior (Figuras 6.3, 6.4), se debe establecer un valor de corte que determine si un documento (*report*) es relevante o no. En función de dicho valor, los documentos (*reports*) con una estimación de similitud mayor o igual, serán clasificados por el sistema como "Relevantes" y el resto como "No Relevantes". Al trabajar con documentos, evaluados y categorizados previamente por expertos como "relevantes" y "no relevantes" para cada criterio de búsqueda (*topic*), se podrá estimar y valorar la precisión del sistema de recuperación propuesto en este trabajo.

Como se desprende del apartado anterior, en caso de trabajar con consultas sencillas (*topics*) resulta fácil determinar este valor de corte por la distribución de los valores hacia los extremos. Sin embargo, en el caso de evaluar búsquedas complejas (*topics*), la decisión resultará una tarea más compleja y crítica para el buen rendimiento del sistema. Por todo ello y para definir dicho valor de corte, se tendrá en cuenta las siguientes premisas:

- El valor debe ser común para ambas métricas y estar entre "0" y "1".
- Debe ser superior a "0.5", ya que este valor representa una relación de sinonimia entre dos conceptos, para la métrica *Path*. Sin embargo, no es suficiente para determinar si un documento (*report*) es relevante frente a consultas complejas (*topics*).
- Debe mostrar un equilibrio en la clasificación de los documentos por su relevancia, es decir, cuanto mayor sea el valor de corte beneficiará más a la clasificación de los documentos "No Relevantes", perjudicando a los "Relevantes".

Con las premisas anteriores y para una matriz simple de máxima similitud de un solo par de conceptos, sólo se determinaría la relevancia del documento médico (*report*) con

respecto a una consulta (*topic*), si dicho par de conceptos tuviese una similitud de "1.0" (distancia igual a 1 y representa el mismo concepto) o "0.5" (distancia igual a 2 y representa una relación de sinonimia). Al trabajar en un contexto real, con frases complejas de múltiples pares de conceptos, la aplicación de un valor medio entre todas las similitudes acarrea errores de varianza que distorsionan los resultados finales de similitud. Por este motivo, es necesario aplicar un requisito añadido que se debe cumplir para considerar correcta la aplicación del valor medio de similitud: Al menos uno de los pares de conceptos debe tener el valor de similitud "1.0" y sólo puede haber un valor inferior a "0.5" (los valores inferiores a 0.5 representan sinonimia lejana entre conceptos). En el caso de que no se cumpla esta restricción, se considerará el documento (*report*) como "No Relevante". Aplicando esta restricción a un conjunto de entrenamiento de 1000 documentos, todos los documentos "Relevantes" poseían valores mayores o iguales a "0.6".

Por esta razón, se establece un valor de corte en "0.6", ya que es el valor mínimo que cumple con las tres premisas anteriores.

Por lo tanto, el resultado final de la matriz de máxima similitud ($Sim\_topic_{vs}report$) reflejará la relevancia de un documento médico (*report*) confrontado con un criterio de búsqueda (*topic*) de la siguiente manera:

- Si valor de ($Sim\_topic_{vs}report$) está en el rango [0.0; 0.6):
  - El *report*, es "No Relevante", para dicho *topic.*
- Si valor de ($Sim\_topic_{vs}report$) está en el rango [0.6; 1.0]:
  - El *report*, es "Relevante", para dicho *topic.*

De esta forma, los ejemplos mostrados en la sección 6.3 (Tabla 6.4; Tabla 6.5), corresponderán a dos documentos médicos (*reports*) "Relevantes" con respecto a los criterios de búsqueda (*topics*) para los que fueron evaluados con ambas métricas:

	Path	**Intrinsic IC-Path**
***Sim_topic101*$_{vs}$*report90230* =**	1.000	1.000
***Sim_topic104*$_{vs}$*report51139* =**	0.7149	0.7783

Tabla 6.6 Valores finales de relevancia, para los ejemplos de la sección 6.3.

Así, en el caso de ejemplo de evaluación del *Topic 104*, para la métrica *Path* con el valor de corte propuesto (Figura 6.5a) se aprecian cómo 9 documentos (*reports*) etiquetados por los expertos como "no relevantes", son clasificados por el sistema como "relevantes" y 1 documento (*report*) etiquetado como "relevante", es clasificado como "no relevante". En el caso de la métrica *Intrinsic IC-Path*, 4 documentos (*reports*) etiquetados por expertos como "no relevantes", son clasificados como "relevantes" y 2 "relevantes", como "no relevantes" (Figura 6.5b).

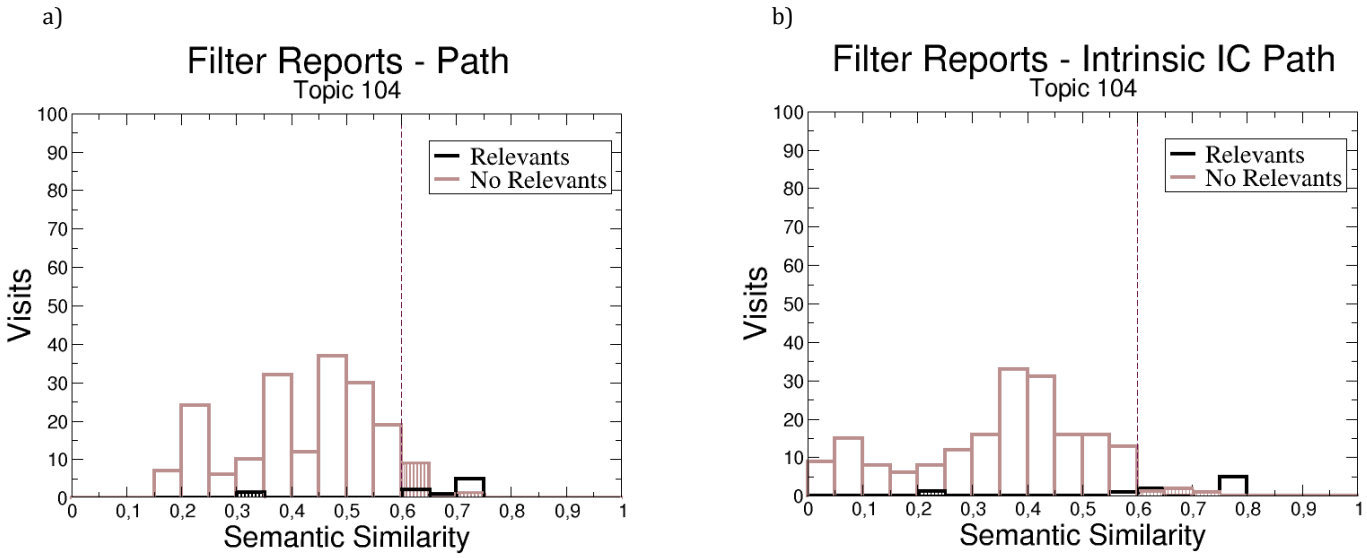

Figura 6.5 Documentos evaluados por el sistema propuesto para el Topic 104, con métricas *Path* e *Intrinsic IC-Path*.

En resumen, en el caso particular del *Topic 104*, los resultados obtenidos para la métrica *Path* son: (Precisión = 44.4%, Recall = 88.9% y F-Measure = 59.3%). Mientras que para la métrica *Intrinsic IC-Path* son: (Precisión = 63.6%, Recall = 77.8% y F- Measure = 70.0%). En este caso, la métrica *Intrinsic IC-Path* muestra un mejor comportamiento que la métrica *Path*.

6.5. Evaluación final de los Resultados

En esta sección se evalúa el rendimiento de las dos métricas analizadas en este trabajo (*Path* e *Intrinsic IC-Path*), en un contexto real de recuperación de información. Para ello, se toman los 35 *Topics* (o criterios de búsqueda) que se proponen en el marco del TREC 2011, tomando como fuente de información 101.712 *reports* o documentos médicos (agrupados en 17.265 visitas).

En la sección 6.5.1., se mostraran los resultados obtenidos por el sistema de recuperación propuesto para cada una de las consultas definidas en el TREC. Seguidamente en la sección 6.5.2, se representan los resultados finales agregados, mediante las principales medidas utilizadas en la evaluación del rendimiento de los sistemas de recuperación de información. Siendo estos la ***Precision***, ***Recall*** (o exhaustividad) *y* ***F-Mesaure***.

6.5.1. Resultados de la métricas Path e Intrinsic IC-Path, sobre el conjunto real de prueba del TREC

A continuación se muestran los resultados obtenidos para cada uno de los criterios de búsqueda propuestos por el TREC 2011.

Para cada criterio de búsqueda (*topic*), se referencia la cadena original completa. Seguidamente se muestran mediante una tabla, el número de *visitas* evaluadas por los expertos como *Relevantes* y *No Relevantes* para dicha búsqueda o (*topic*), junto con el porcentaje de aciertos en los resultados obtenidos para ambas métricas, así como el número de las visitas que conforman dichos aciertos y fallos.

(Conviene recordar que la *visita* resultante, viene definida por el documento médico o *report* cuya matriz de máxima similitud es superior a todos los *reports* que componen dicho *episodio médico* o *visita*, pudiendo estar compuesta desde 1 hasta 415 documentos o *reports*). Se mostrarán en color rojo aquellas visitas que fueron calificadas por los expertos como "*No Relevantes*" y en verde como "*Relevantes*"

Por último se muestran los histogramas resultantes para cada una de las métricas evaluadas, *Path* e *Intrinsic IC-Path*.

En el **Apéndice A**, se representa la descomposición y expansión de cada consulta o *topic*, en:

- Identificador de subfrase.
- Identificador de frase.
- Digito de negación.
- CUI o concepto.
- Tipo semántico al que pertenece dicho concepto
- Subcadena de referencia y término preferido para dicho concepto en UMLS.
- La clasificación de los tipos semánticos asociados a la consulta en tipos "Específicos" y "Genéricos".

TOPIC 101: Patients with hearing loss

PATH 101				Corte 0.6 Visitas
RELEV.	74	Aciertos (>=0,6):	82.4%	61
		Fallos (<0,6):		13
NO RELEV.	175	Aciertos(<0,6):	88,0%	154
		Fallos (>=0,6):		21

INTRINSIC IC-PATH 101				Corte 0.6 Visitas
RELEV.	74	Aciertos (>=0,6):	82.4%	61
		Fallos (<0,6):		13
NO RELEV.	175	Aciertos(<0,6):	88,0%	154
		Fallos (>=0,6):		21

Tabla 6.7 Resultados Máxima Similitud *Topic 101 - Path* e *Intrinsic IC-Path*

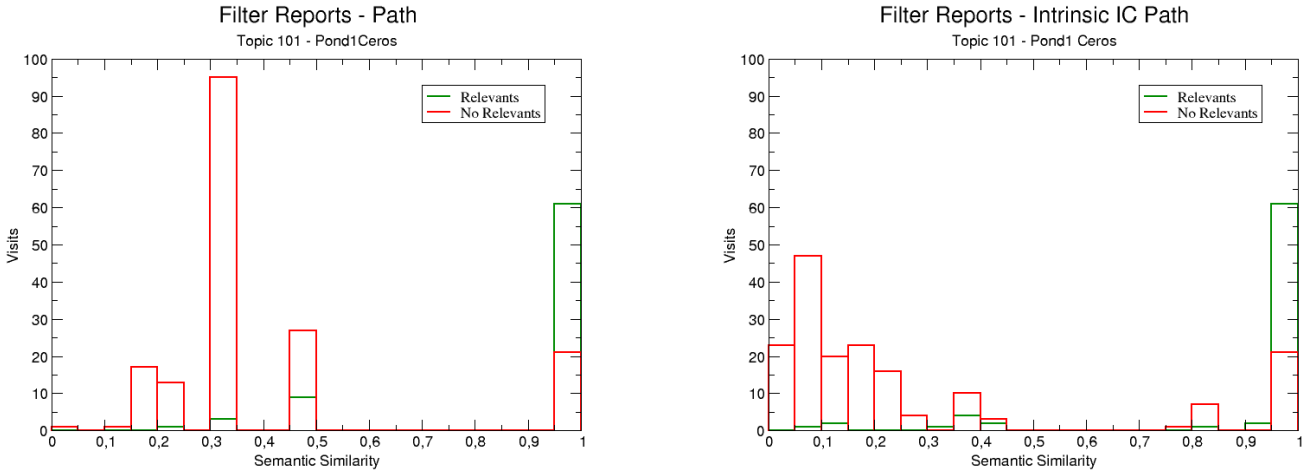

Figura 6.6 Histograma Máxima Similitud *Topic 101 - Path* e *Intrinsic IC-Path*

TOPIC 102: Patients with complicated GERD who receive endoscopy

PATH 102				Corte 0.6 Visitas
RELEV.	89	Aciertos (>=0,6):	100%	89
		Fallos (<0,6):		0
NO RELEV.	202	Aciertos(<0,6):	29.70%	60
		Fallos (>=0,6):		142

INTRINSIC IC-PATH 102				Corte 0.6 Visitas
RELEV.	89	Aciertos (>=0,6):	78.7%	70
		Fallos (<0,6):		19
NO RELEV.	202	Aciertos(<0,6):	63.9%	129
		Fallos (>=0,6):		73

Tabla 6.8 Resultados Máxima Similitud *Topic 102 - Path* e *Intrinsic IC-Path*

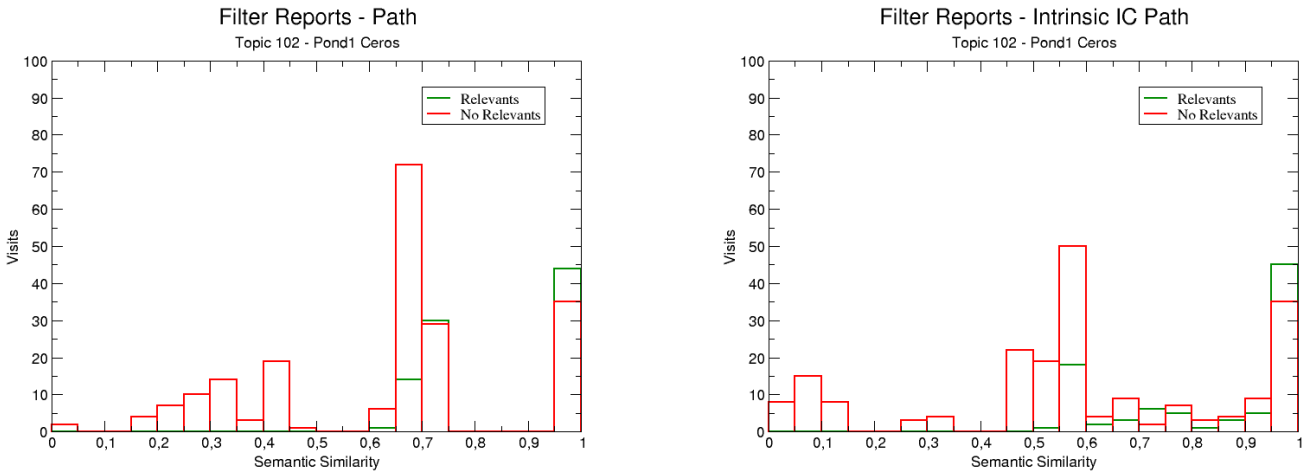

Figura 6.7 Histograma Máxima Similitud *Topic 102 - Path* e *Intrinsic IC-Path*

TOPIC 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis

PATH 103				Corte 0.6 Visitas
RELEV.	12	Aciertos (>=0,6):	50%	6
		Fallos (<0,6):		6
NO RELEV.	188	Aciertos(<0,6):	93.1%	175
		Fallos (>=0,6):		13

INTRINSIC IC-PATH 103				Corte 0.6 Visitas
RELEV.	12	Aciertos (>=0,6):	75%	9
		Fallos (<0,6):		3
NO RELEV.	188	Aciertos(<0,6):	90.2%	173
		Fallos (>=0,6):		15

Tabla 6.9 Resultados Máxima Similitud *Topic 103 - Path* e *Intrinsic IC-Path*

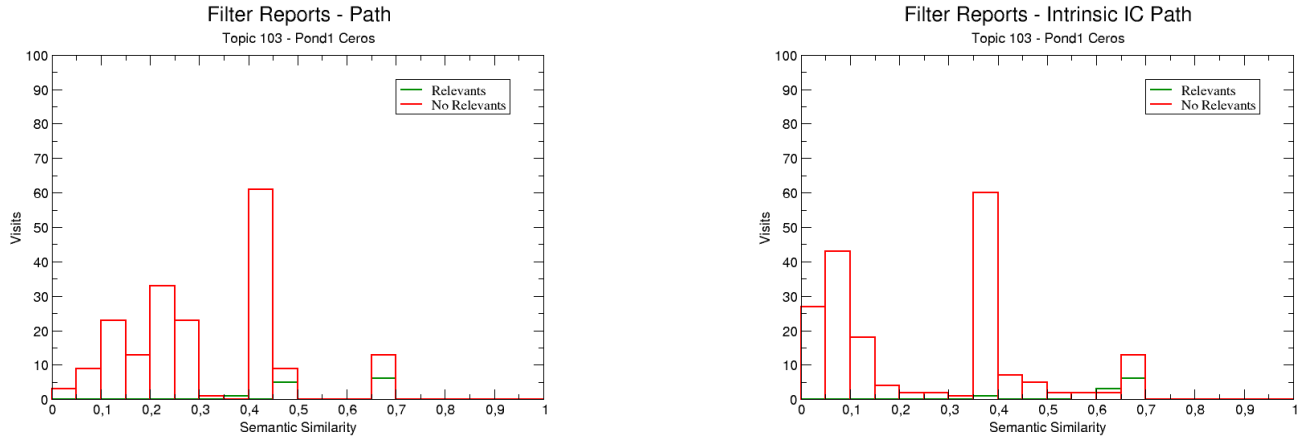

Figura 6.8 Histograma Máxima Similitud *Topic 103 - Path* e *Intrinsic IC-Path*

TOPIC 104: Patients diagnosed with localized prostate cancer and treated with robotic surgery

PATH 104				Corte 0.6 Visitas
RELEV.	9	Aciertos (>=0,6):	88.9%	8
		Fallos (<0,6):		1
NO RELEV.	187	Aciertos(<0,6):	94.7%	177
		Fallos (>=0,6):		10

INTRINSIC IC-PATH 104				Corte 0.6 Visitas
RELEV.	9	Aciertos (>=0,6):	77.8%	7
		Fallos (<0,6):		2
NO RELEV.	187	Aciertos(<0,6):	97.9%	183
		Fallos (>=0,6):		4

Tabla 6.10 Resultados Máxima Similitud *Topic 104 - Path* e *Intrinsic IC-Path*

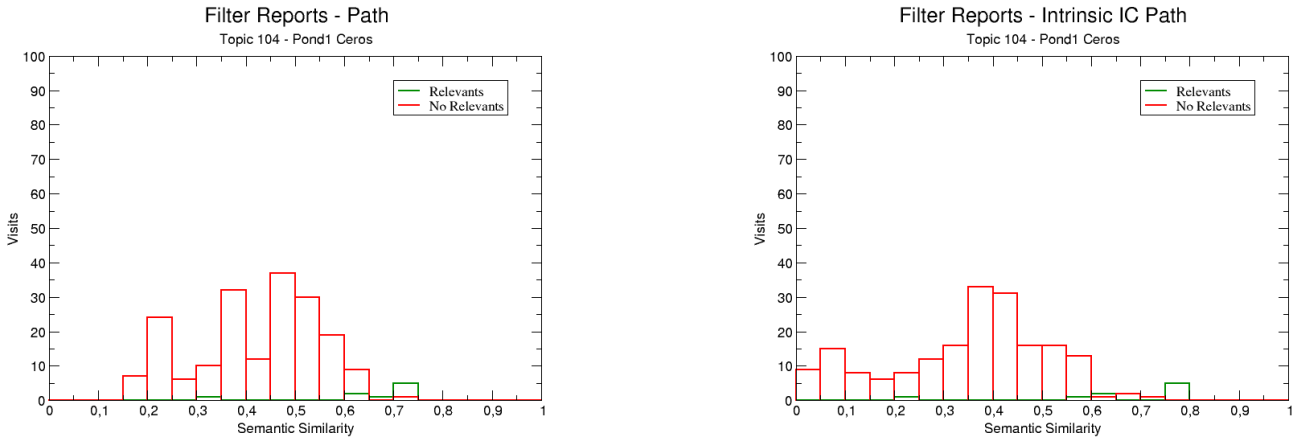

Figura 6.9 Histograma Máxima Similitud *Topic 104 - Path* e *Intrinsic IC-Path*

TOPIC 105: Patients with dementia

PATH 105				Corte 0.6 Visitas
RELEV.	145	**Aciertos (>=0,6):**	**92.4%**	**134**
		Fallos (<0,6):		11
NO RELEV.	84	**Aciertos(<0,6):**	**88.1%**	**74**
		Fallos (>=0,6):		10

INTRINSIC IC-PATH 105				Corte 0.6 Visitas
RELEV.	145	**Aciertos (>=0,6):**	**93.1%**	**135**
		Fallos (<0,6):		10
NO RELEV.	84	**Aciertos(<0,6):**	**88.1%**	**74**
		Fallos (>=0,6):		10

Tabla 6.11 Resultados Máxima Similitud *Topic 105 - Path* e *Intrinsic IC-Path*

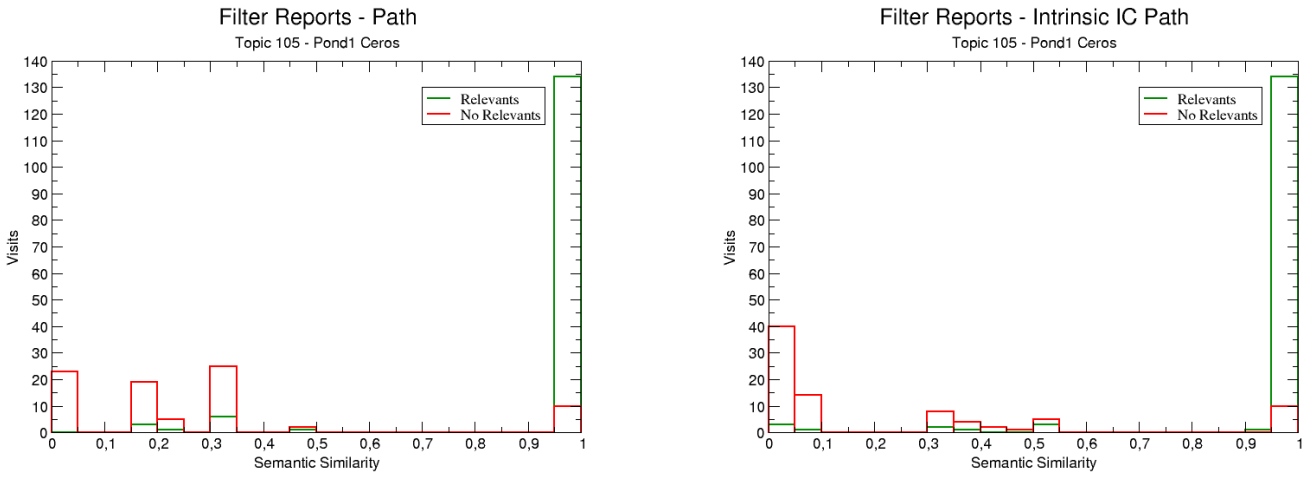

Figura 6.10 Histograma Máxima Similitud *Topic 105 - Path* e *Intrinsic IC-Path*

TOPIC 106: Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer

PATH 106				Corte 0.6 Visitas
RELEV.	85	**Aciertos (>=0,6):**	**95.3%**	**81**
		Fallos (<0,6):		4
NO RELEV.	213	**Aciertos(<0,6):**	**35.2 %**	**75**
		Fallos (>=0,6):		138

INTRINSIC IC-PATH 106				Corte 0.6 Visitas
RELEV.	85	**Aciertos (>=0,6):**	**84.7%**	**72**
		Fallos (<0,6):		13
NO RELEV.	213	**Aciertos(<0,6):**	**55.9%**	**119**
		Fallos (>=0,6):		94

Tabla 6.12 Resultados Máxima Similitud *Topic 106 - Path* e *Intrinsic IC-Path*

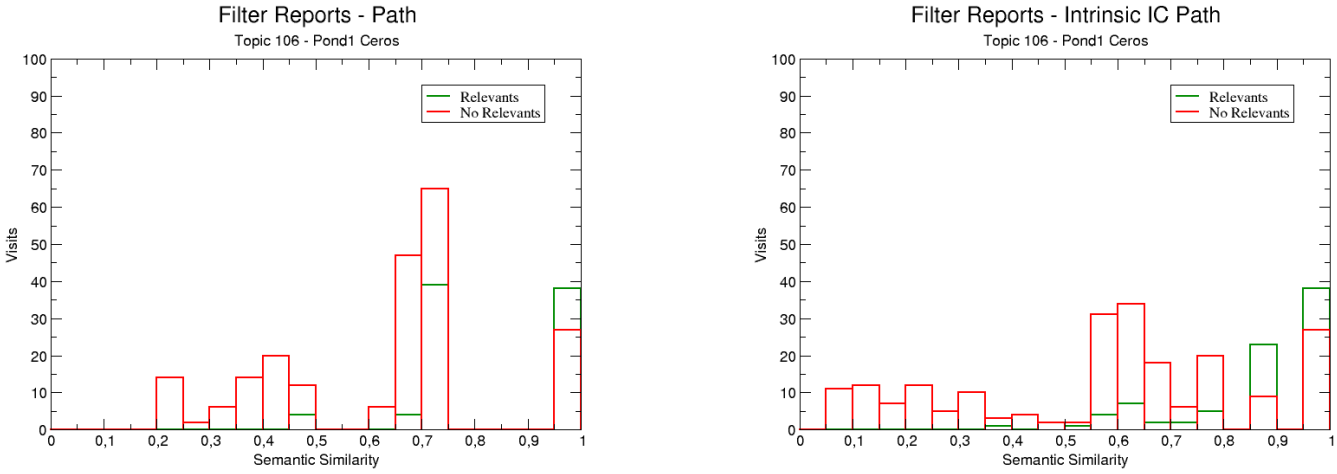

Figura 6.11 Histograma Máxima Similitud *Topic 106 - Path* e *Intrinsic IC-Path*

TOPIC 107: Patients with ductal carcinoma in situ (DCIS)

PATH 107				Corte 0.6 Visitas
RELEV.	23	Aciertos (>=0,6):	21.8%	5
		Fallos (<0,6):		18
NO RELEV.	189	Aciertos(<0,6):	96.3 %	182
		Fallos (>=0,6):		7

INTRINSIC IC-PATH 107				Corte 0.6 Visitas
RELEV.	23	Aciertos (>=0,6):	91.3%	21
		Fallos (<0,6):		2
NO RELEV.	189	Aciertos(<0,6):	83.1%	157
		Fallos (>=0,6):		32

Tabla 6.13 Resultados Máxima Similitud *Topic 107 - Path* e *Intrinsic IC-Path*

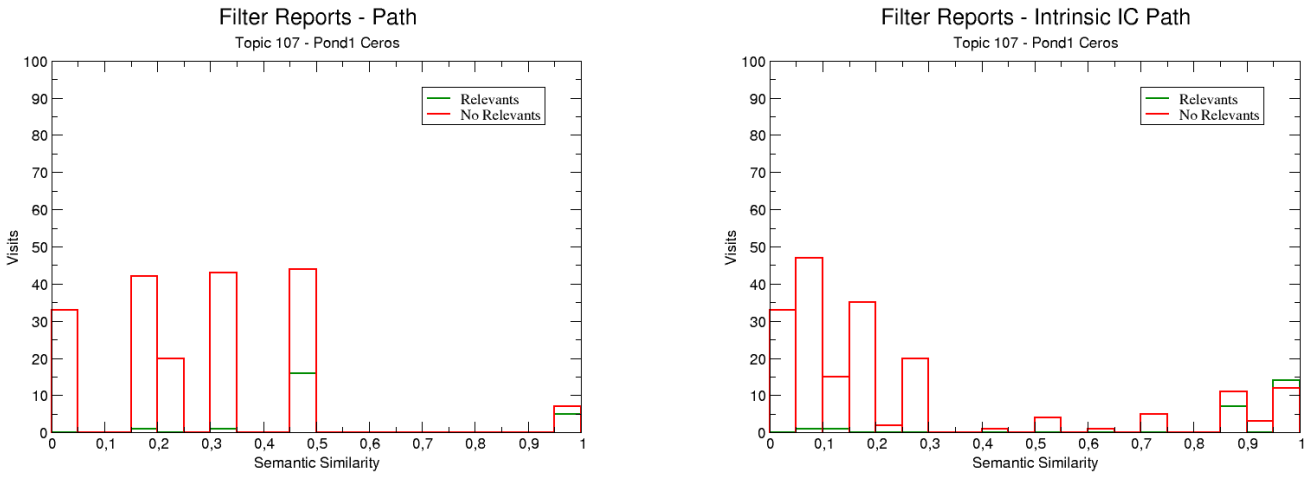

Figura 6.12 Histograma Máxima Similitud *Topic 107 - Path* e *Intrinsic IC-Path*

TOPIC 108: Patients treated for vascular claudication surgically

PATH 108				Corte 0.6 Visitas
RELEV.	13	Aciertos (>=0,6):	69.2%	9
		Fallos (<0,6):		4
NO RELEV.	278	Aciertos(<0,6):	82.4 %	229
		Fallos (>=0,6):		46

INTRINSIC IC-PATH 108				Corte 0.6 Visitas
RELEV.	13	Aciertos (>=0,6):	15.4%	2
		Fallos (<0,6):		11
NO RELEV.	278	Aciertos(<0,6):	95.7%	266
		Fallos (>=0,6):		12

Tabla 6.14 Resultados Máxima Similitud *Topic 108 - Path* e *Intrinsic IC-Path*

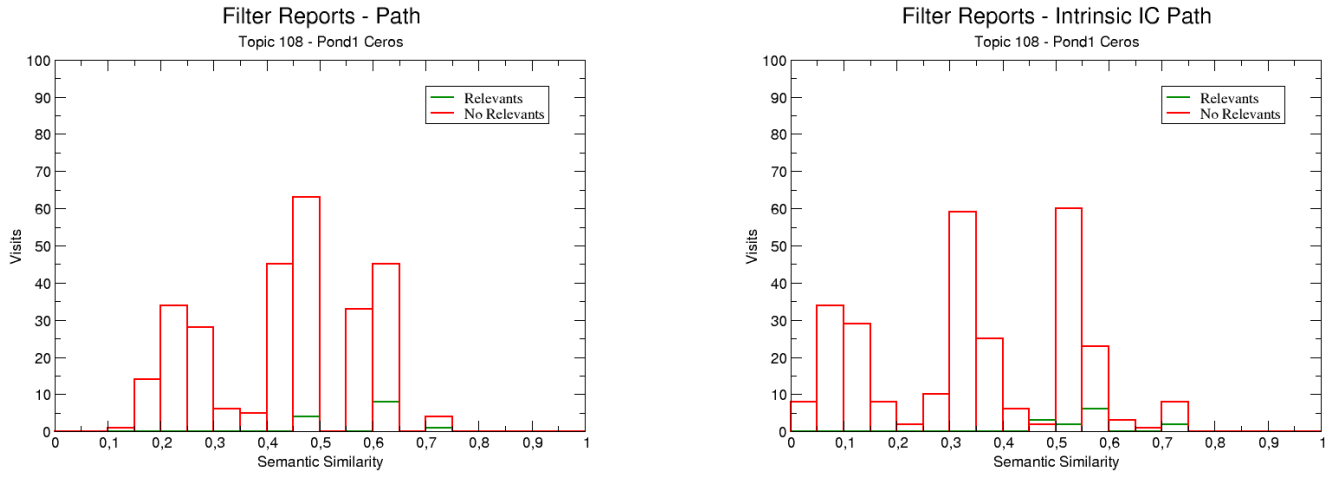

Figura 6.13 Histograma Máxima Similitud *Topic 108 - Path* e *Intrinsic IC-Path*

TOPIC 109: Women with osteopenia

PATH 109				Corte 0.6 Visitas
RELEV.	123	Aciertos (>=0,6):	95.1%	117
		Fallos (<0,6):		6
NO RELEV.	146	Aciertos(<0,6):	26.7 %	39
		Fallos (>=0,6):		107

INTRINSIC IC-PATH 109				Corte 0.6 Visitas
RELEV.	123	Aciertos (>=0,6):	35%	43
		Fallos (<0,6):		80
NO RELEV.	146	Aciertos(<0,6):	89.04%	130
		Fallos (>=0,6):		16

Tabla 6. 15 Resultados Máxima Similitud *Topic 109 - Path* e *Intrinsic IC-Path*

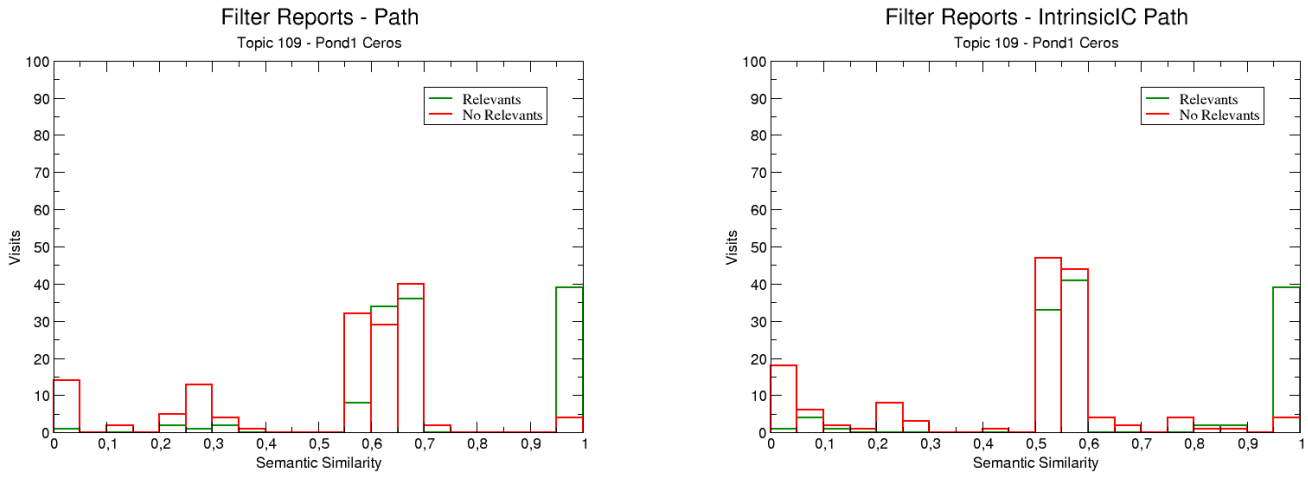

Figura 6.14 Histograma Máxima Similitud *Topic 109 - Path* e *Intrinsic IC-Path*

TOPIC 110: Patients being discharged from the hospital on hemodialysis

PATH 110				Corte 0.6 Visitas
RELEV.	95	Aciertos (>=0,6):	91.6%	87
		Fallos (<0,6):		8
NO RELEV.	181	Aciertos(<0,6):	79.6 %	144
		Fallos (>=0,6):		37

INTRINSIC IC-PATH 110				Corte 0.6 Visitas
RELEV.	95	Aciertos (>=0,6):	8.4%	8
		Fallos (<0,6):		87
NO RELEV.	181	Aciertos(<0,6):	97.2%	176
		Fallos (>=0,6):		5

Tabla 6.16 Resultados Máxima Similitud *Topic 110 - Path* e *Intrinsic IC-Path*

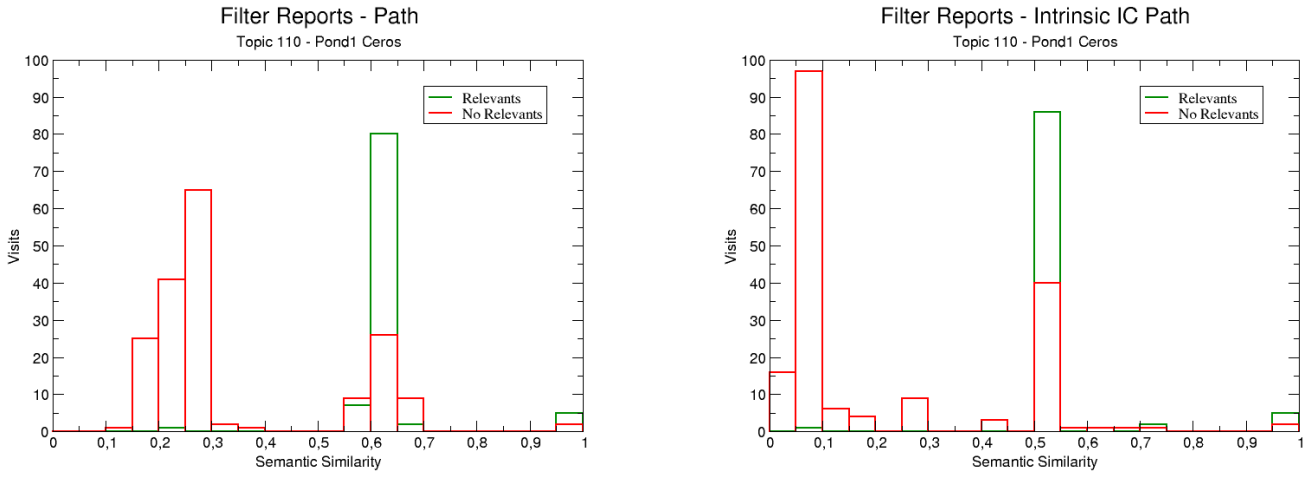

Figura 6.15 Histograma Máxima Similitud *Topic 110 - Path* e *Intrinsic IC-Path*

TOPIC 111: Patients with chronic back pain who receive an intraspinal pain-medicine pump

PATH 111				Corte 0.6 Visitas
RELEV.	21	Aciertos (>=0,6):	23.8%	5
		Fallos (<0,6):		16
NO RELEV.	263	Aciertos(<0,6):	98.5%	259
		Fallos (>=0,6):		4

INTRINSIC IC-PATH 111				Corte 0.6 Visitas
RELEV.	21	Aciertos (>=0,6):	52.4%	11
		Fallos (<0,6):		10
NO RELEV.	263	Aciertos(<0,6):	93.9%	247
		Fallos (>=0,6):		16

Tabla 6 17 Resultados Máxima Similitud *Topic 111 - Path* e *Intrinsic IC-Path*

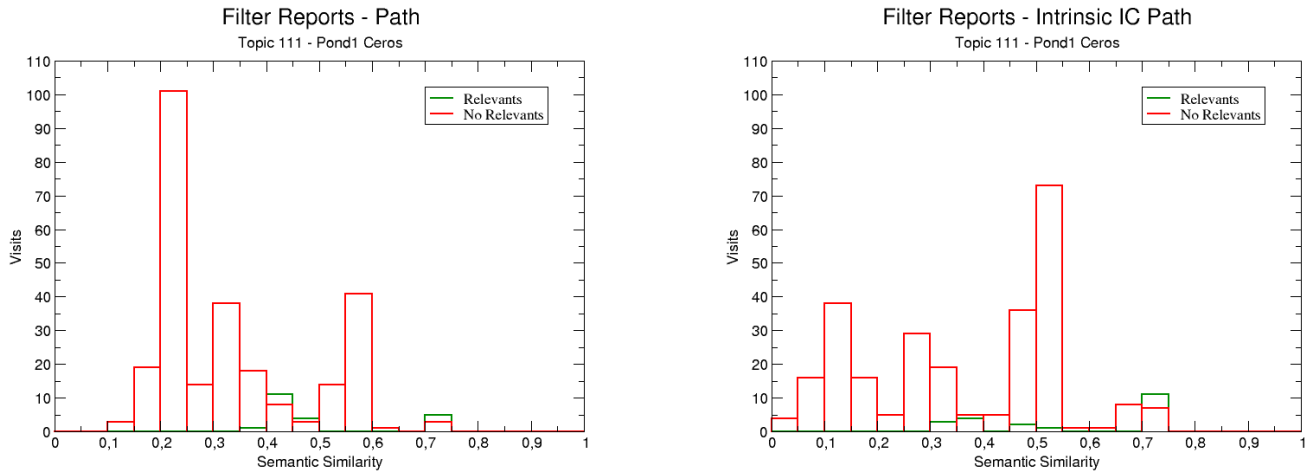

Figura 6.16 Histograma Máxima Similitud *Topic 111 - Path* e *Intrinsic IC-Path*

TOPIC 112: Female patients with breast cancer with mastectomies during admission

PATH 112				Corte 0.6 Visitas
RELEV.	73	Aciertos (>=0,6):	89.1%	65
		Fallos (<0,6):		8
NO RELEV.	125	Aciertos(<0,6):	62.4%	78
		Fallos (>=0,6):		47

INTRINSIC IC-PATH 112				Corte 0.6 Visitas
RELEV.	73	Aciertos (>=0,6):	89.1%	65
		Fallos (<0,6):		8
NO RELEV.	125	Aciertos(<0,6):	87.2%	109
		Fallos (>=0,6):		16

Tabla 6.18 Resultados Máxima Similitud *Topic 112 - Path* e *Intrinsic IC-Path*

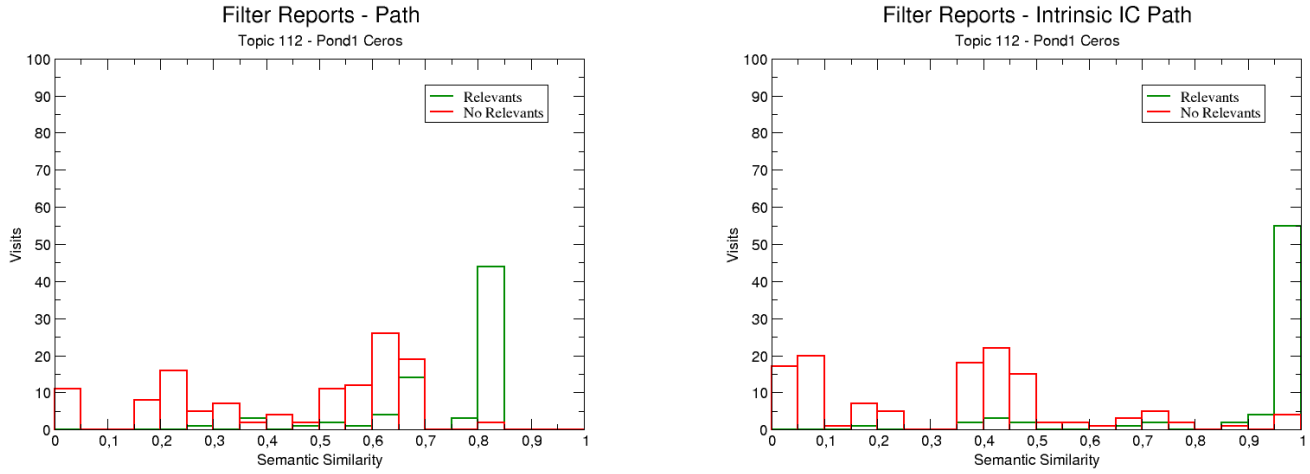

Figura 6.17 Histograma Máxima Similitud *Topic 112 - Path* e *Intrinsic IC-Path*

TOPIC 113: Adult patients who received colonoscopies during admission which revealed adenocarcinoma

PATH 113				Corte 0.6 Visitas
RELEV.	14	Aciertos (>=0,6):	92.9%	13
		Fallos (<0,6):		1
NO RELEV.	196	Aciertos(<0,6):	65.3%	128
		Fallos (>=0,6):		68

INTRINSIC IC-PATH 113				Corte 0.6 Visitas
RELEV.	14	Aciertos (>=0,6):	92.9%	13
		Fallos (<0,6):		1
NO RELEV.	196	Aciertos(<0,6):	88.8%	174
		Fallos (>=0,6):		22

Tabla 6.19 Resultados Máxima Similitud *Topic 113 - Path* e *Intrinsic IC-Path*

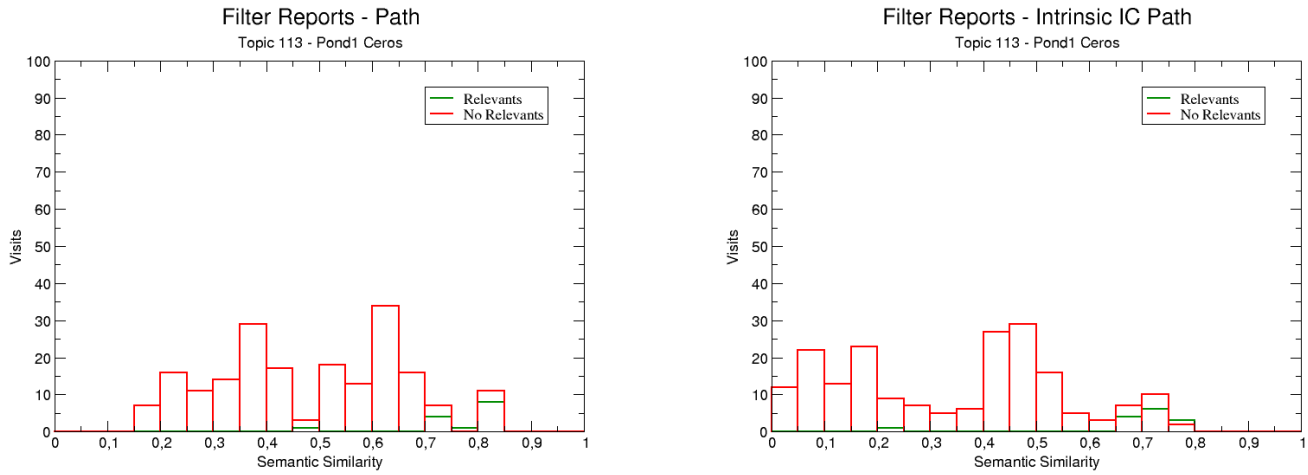

Figura 6.18 Histograma Máxima Similitud *Topic 113 - Path* e *Intrinsic IC-Path*

TOPIC 114: Adult patients discharged home with palliative care / home hospice

PATH 114				Corte 0.6 Visitas
RELEV.	55	Aciertos (>=0,6):	85.5%	47
		Fallos (<0,6):		8
NO RELEV.	146	Aciertos(<0,6):	64.4%	94
		Fallos (>=0,6):		52

INTRINSIC IC-PATH 114				Corte 0.6 Visitas
RELEV.	55	Aciertos (>=0,6):	78.2%	43
		Fallos (<0,6):		12
NO RELEV.	146	Aciertos(<0,6):	87.0%	127
		Fallos (>=0,6):		19

Tabla 6.20 Resultados Máxima Similitud *Topic 114 - Path* e *Intrinsic IC-Path*

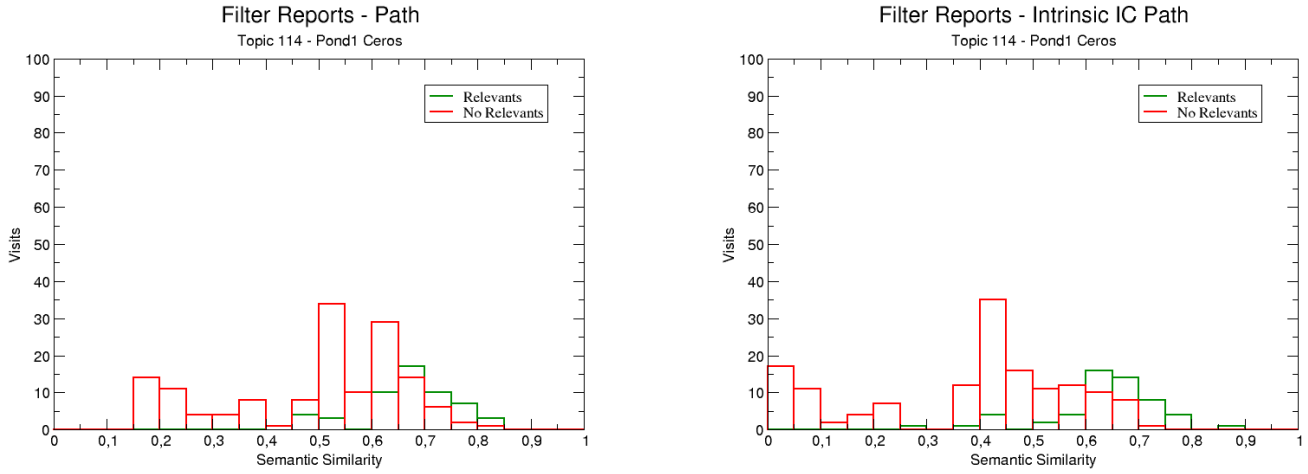

Figura 6.19 Histograma Máxima Similitud *Topic 114 - Path* e *Intrinsic IC-Path*

TOPIC 115: Adult patients who are admitted with an asthma exacerbation

PATH 115				Corte 0.6 Visitas
RELEV.	36	Aciertos (>=0,6):	73.2%	26
		Fallos (<0,6):		10
NO RELEV.	206	Aciertos(<0,6):	85.9%	177
		Fallos (>=0,6):		29

INTRINSIC IC-PATH 115				Corte 0.6 Visitas
RELEV.	35	Aciertos (>=0,6):	33.3%	12
		Fallos (<0,6):		24
NO RELEV.	206	Aciertos(<0,6):	85.9%	177
		Fallos (>=0,6):		14

Tabla 6.21 Resultados Máxima Similitud *Topic 115 - Path* e *Intrinsic IC-Path*

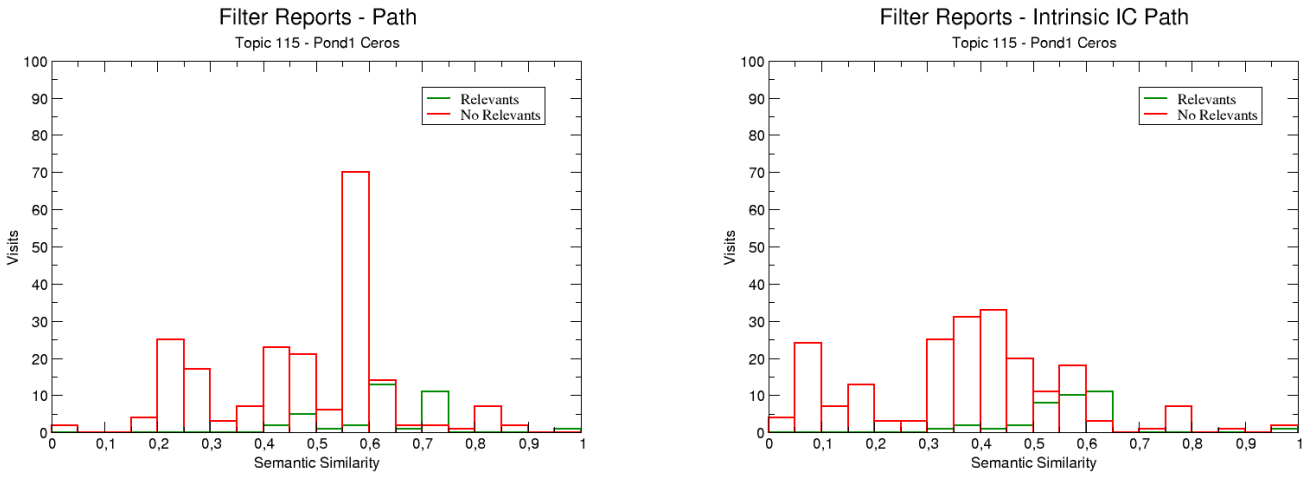

Figura 6.20 Histograma Máxima Similitud *Topic 115 - Path* e *Intrinsic IC-Path*

TOPIC 116: Patients received methotrexate for cancer treatment while in the hospital

PATH 116				Corte 0.6 Visitas
RELEV.	10	Aciertos (>=0,6):	100%	10
		Fallos (<0,6):		0
NO RELEV.	216	Aciertos(<0,6):	50.9%	110
		Fallos (>=0,6):		106

INTRINSIC IC-PATH 116				Corte 0.6 Visitas
RELEV.	10	Aciertos (>=0,6):	60%	6
		Fallos (<0,6):		4
NO RELEV.	216	Aciertos(<0,6):	66.2%	143
		Fallos (>=0,6):		73

Tabla 6.22 Resultados Máxima Similitud *Topic 116 - Path* e *Intrinsic IC-Path*

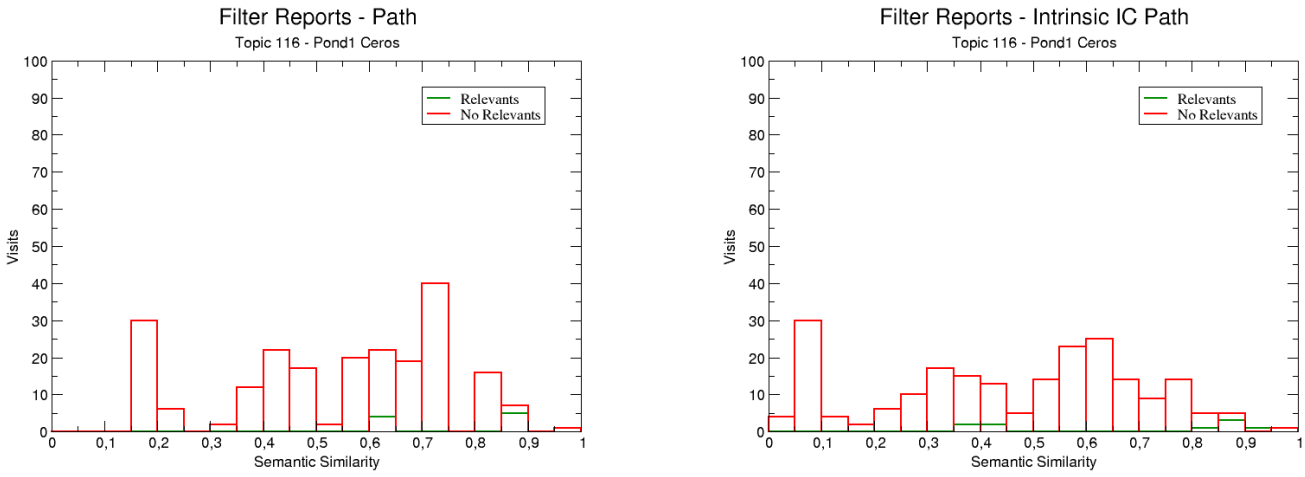

Figura 6.21 Histograma Máxima Similitud *Topic 116 - Path* e *Intrinsic IC-Path*

TOPIC 117: Patients with Post-traumatic Stress Disorder

PATH 117				Corte 0.6 Visitas
RELEV.	22	**Aciertos (>=0,6):**	**95.5%**	**21**
		Fallos (<0,6):		1
NO RELEV.	240	**Aciertos(<0,6):**	**96.7%**	**232**
		Fallos (>=0,6):		8

INTRINSIC IC-PATH 117				Corte 0.6 Visitas
RELEV.	22	**Aciertos (>=0,6):**	**95.5%**	**21**
		Fallos (<0,6):		1
NO RELEV.	240	**Aciertos(<0,6):**	**88.8%**	**213**
		Fallos (>=0,6):		27

Tabla 6.23 Resultados Máxima Similitud *Topic 117 - Path* e *Intrinsic IC-Path*

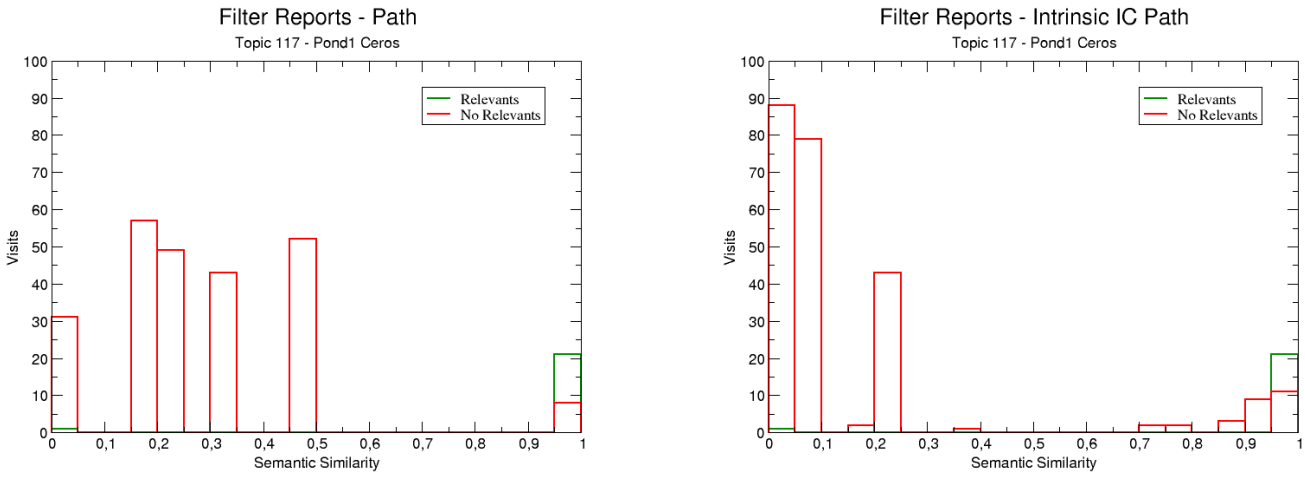

Figura 6.22 Histograma Máxima Similitud *Topic 117 - Path* e *Intrinsic IC-Path*

TOPIC 118: Adults who received a coronary stent during an admission

PATH 118				Corte 0.6 Visitas
RELEV.	52	**Aciertos (>=0,6):**	**98.1%**	**51**
		Fallos (<0,6):		1
NO RELEV.	273	**Aciertos(<0,6):**	**36.3%**	**99**
		Fallos (>=0,6):		174

INTRINSIC IC-PATH 118				Corte 0.6 Visitas
RELEV.	52	**Aciertos (>=0,6):**	**98.1%**	**51**
		Fallos (<0,6):		1
NO RELEV.	273	**Aciertos(<0,6):**	**46.9%**	**128**
		Fallos (>=0,6):		145

Tabla 6.24 Resultados Máxima Similitud *Topic 118 - Path* e *Intrinsic IC-Path*

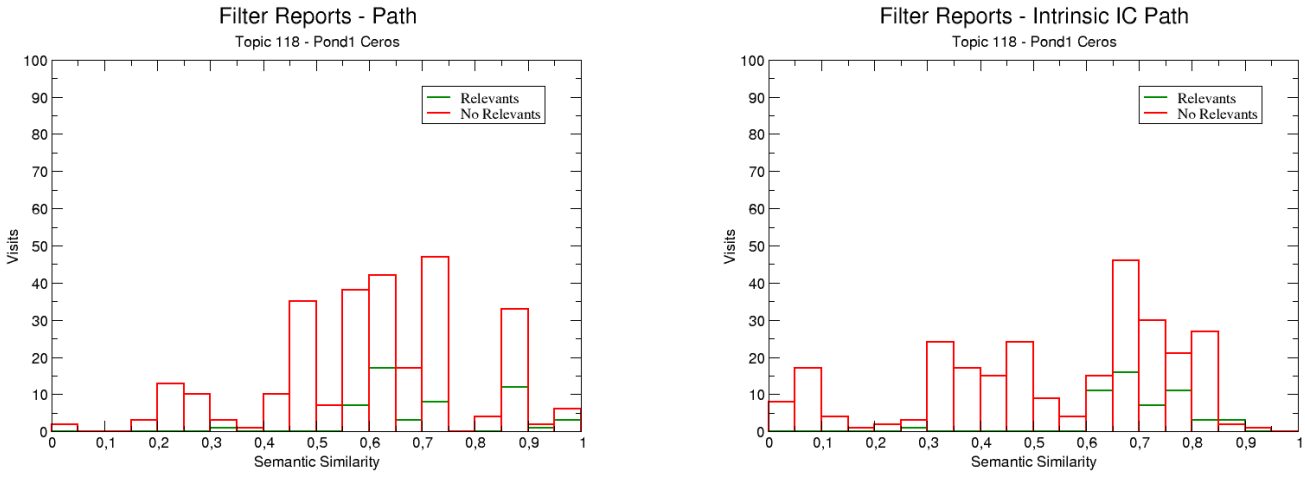

Figura 6.23 Histograma Máxima Similitud *Topic 118 - Path* e *Intrinsic IC-Path*

TOPIC 119: Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes

PATH 119				Corte 0.6 Visitas
RELEV.	46	Aciertos (>=0,6):	26.1%	12
		Fallos (<0,6):		34
NO RELEV.	190	Aciertos(<0,6):	95.3%	181
		Fallos (>=0,6):		9

INTRINSIC IC-PATH 119				Corte 0.6 Visitas
RELEV.	46	Aciertos (>=0,6):	0%	0
		Fallos (<0,6):		46
NO RELEV.	190	Aciertos(<0,6):	99.5%	189
		Fallos (>=0,6):		1

Tabla 6.25 Resultados Máxima Similitud *Topic 119 - Path* e *Intrinsic IC-Path*

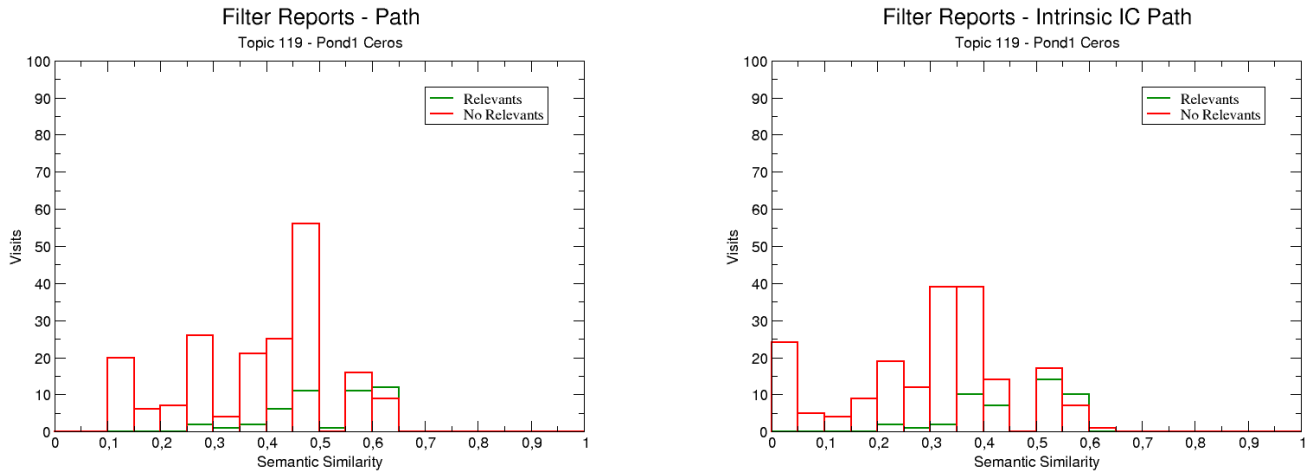

Figura 6.24 Histograma Máxima Similitud *Topic 119 - Path* e *Intrinsic IC-Path*

TOPIC 120: Patients admitted for treatment of CHF exacerbation

PATH 120				Corte 0.6 Visitas
RELEV.	117	Aciertos (>=0,6):	99.1%	116
		Fallos (<0,6):		1
NO RELEV.	205	Aciertos(<0,6):	28.8%	59
		Fallos (>=0,6):		146

INTRINSIC IC-PATH 120				Corte 0.6 Visitas
RELEV.	117	Aciertos (>=0,6):	99.1%	116
		Fallos (<0,6):		1
NO RELEV.	205	Aciertos(<0,6):	38.1%	78
		Fallos (>=0,6):		127

Tabla 6.26 Resultados Máxima Similitud *Topic 120 - Path* e *Intrinsic IC-Path*

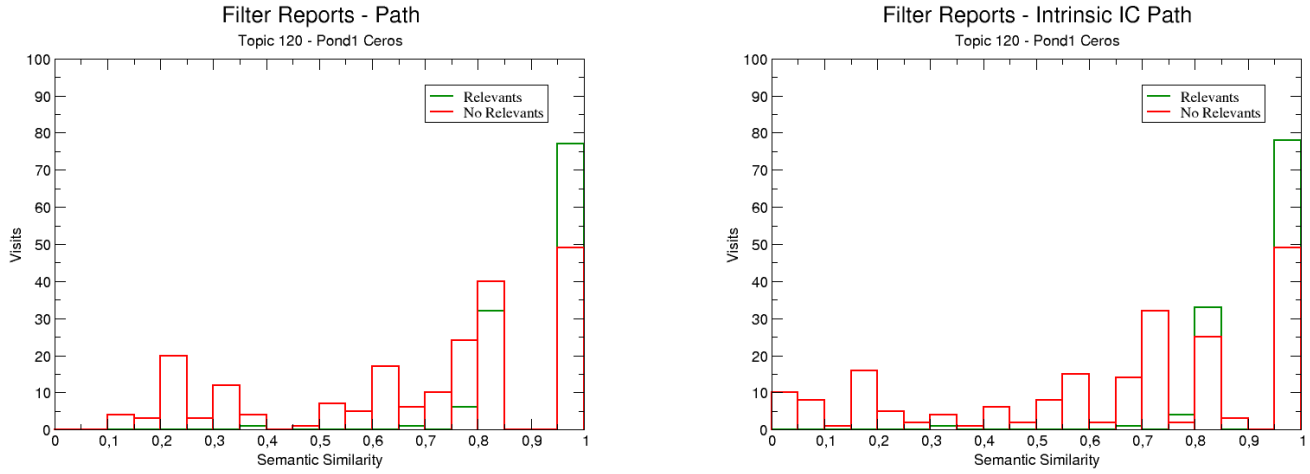

Figura 6.25 Histograma Máxima Similitud *Topic 120 - Path* e *Intrinsic IC-Path*

TOPIC 121: Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix

PATH 121				Corte 0.6 Visitas
RELEV.	40	Aciertos (>=0,6):	92.5%	37
		Fallos (<0,6):		3
NO RELEV.	258	Aciertos(<0,6):	58.1%	150
		Fallos (>=0,6):		108

INTRINSIC IC-PATH 121				Corte 0.6 Visitas
RELEV.	40	Aciertos (>=0,6):	97.5%	39
		Fallos (<0,6):		1
NO RELEV.	258	Aciertos(<0,6):	38.4%	99
		Fallos (>=0,6):		159

Tabla 6.27 Resultados Máxima Similitud *Topic 121 - Path* e *Intrinsic IC-Path*

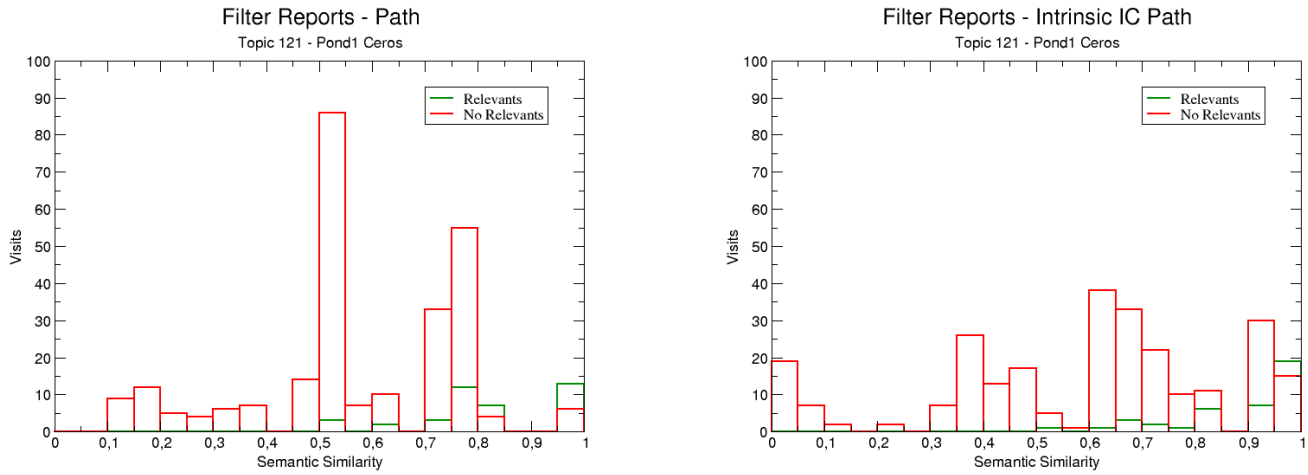

Figura 6.26 Histograma Máxima Similitud *Topic 121 - Path* e *Intrinsic IC-Path*

TOPIC 122: Patients who received total parenteral nutrition while in the hospital

PATH 122				Corte 0.6 Visitas
RELEV.	24	Aciertos (>=0,6):	91.7%	22
		Fallos (<0,6):		2
NO RELEV.	193	Aciertos(<0,6):	45.6%	88
		Fallos (>=0,6):		105

INTRINSIC IC-PATH 122				Corte 0.6 Visitas
RELEV.	24	Aciertos (>=0,6):	58.3%	14
		Fallos (<0,6):		10
NO RELEV.	193	Aciertos(<0,6):	69.9%	135
		Fallos (>=0,6):		58

Tabla 6.28 Resultados Máxima Similitud *Topic 122 - Path* e *Intrinsic IC-Path*

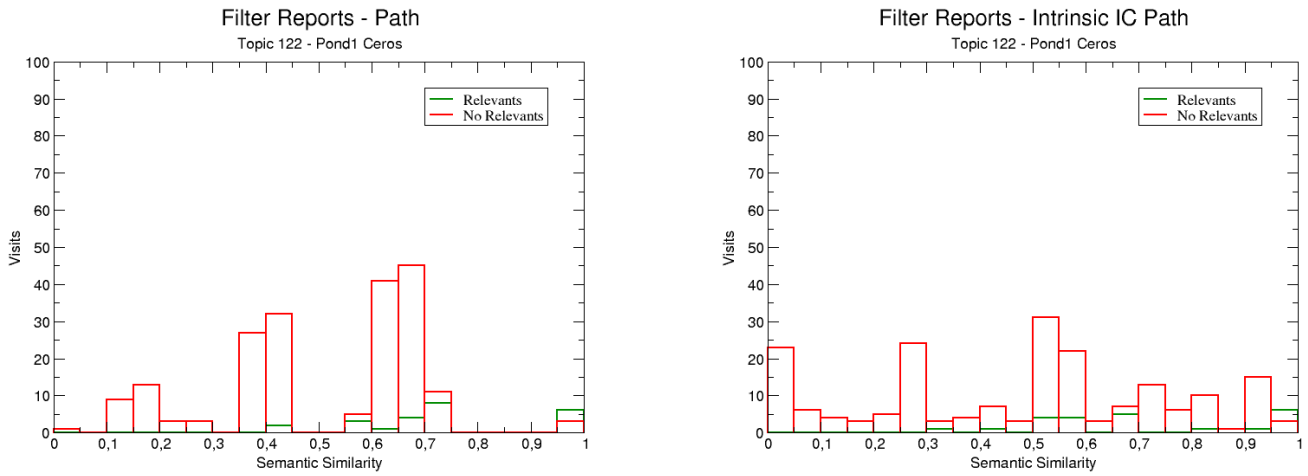

Figura 6.27 Histograma Máxima Similitud *Topic 122 - Path* e *Intrinsic IC-Path*

TOPIC 123: Diabetic patients who received diabetic education in the hospital

PATH 123				Corte 0.6 Visitas
RELEV.	33	Aciertos (>=0,6):	0%	0
		Fallos (<0,6):		33
NO RELEV.	214	Aciertos(<0,6):	100%	214
		Fallos (>=0,6):		0

INTRINSIC IC-PATH 123				Corte 0.6 Visitas
RELEV.	33	Aciertos (>=0,6):	0%	0
		Fallos (<0,6):		33
NO RELEV.	214	Aciertos(<0,6):	99.1%	212
		Fallos (>=0,6):		2

Tabla 6.29 Resultados Máxima Similitud *Topic 123 - Path* e *Intrinsic IC-Path*

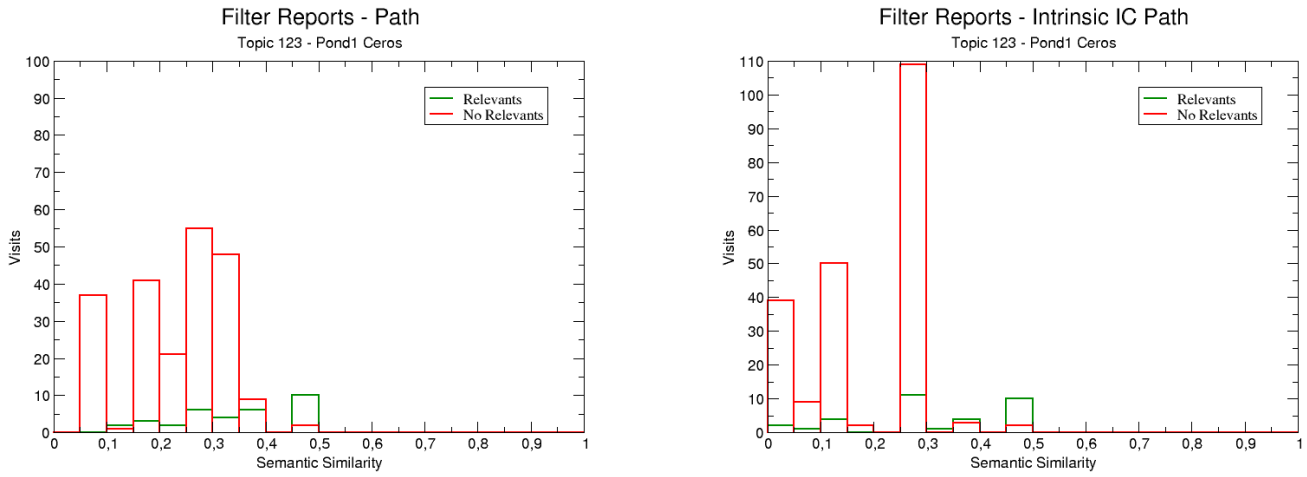

Figura 6.28 Histograma Máxima Similitud *Topic 123 - Path* e *Intrinsic IC-Path*

TOPIC 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma

PATH 124				Corte 0.6 Visitas
RELEV.	6	Aciertos (>=0,6):	100%	6
		Fallos (<0,6):		0
NO RELEV.	289	Aciertos(<0,6):	39.5%	114
		Fallos (>=0,6):		175

INTRINSIC IC-PATH 124				Corte 0.6 Visitas
RELEV.	6	Aciertos (>=0,6):	83.3%	5
		Fallos (<0,6):		1
NO RELEV.	289	Aciertos(<0,6):	58.1%	168
		Fallos (>=0,6):		121

Tabla 6.30 Resultados Máxima Similitud *Topic 124 - Path* e *Intrinsic IC-Path*

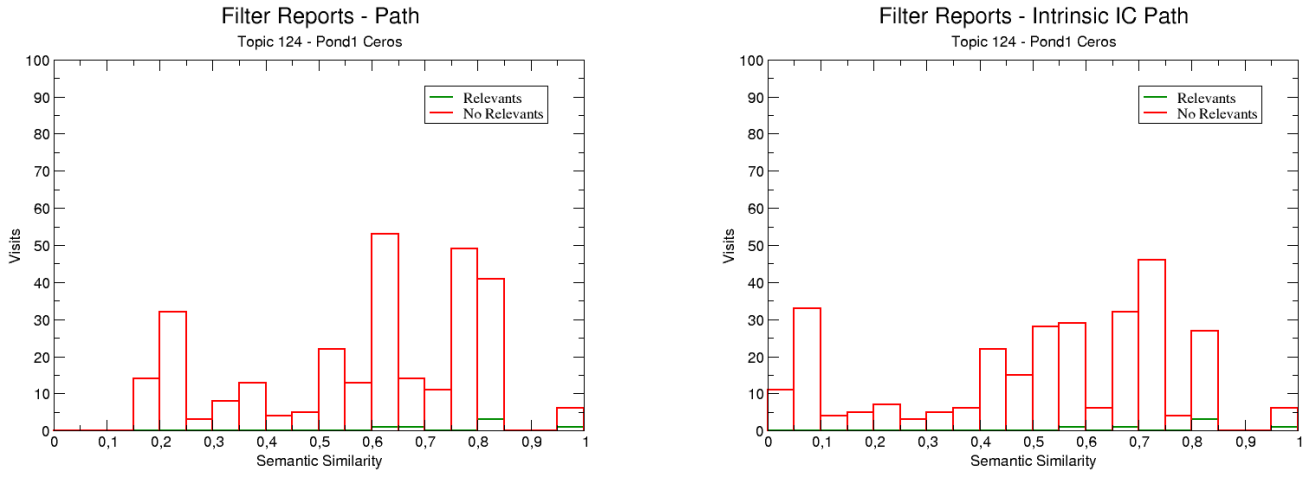

Figura 6.29 Histograma Máxima Similitud *Topic 124 - Path* e *Intrinsic IC-Path*

TOPIC 125: Patients co-infected with Hepatitis C and HIV

PATH 125				Corte 0.6 Visitas
RELEV.	14	Aciertos (>=0,6):	21.4%	3
		Fallos (<0,6):		11
NO RELEV.	225	Aciertos(<0,6):	57.8%	130
		Fallos (>=0,6):		95

INTRINSIC IC-PATH 125				Corte 0.6 Visitas
RELEV.	14	Aciertos (>=0,6):	21.4%	3
		Fallos (<0,6):		11
NO RELEV.	225	Aciertos(<0,6):	62.2%	140
		Fallos (>=0,6):		85

Tabla 6.31 Resultados Máxima Similitud *Topic 125 - Path* e *Intrinsic IC-Path*

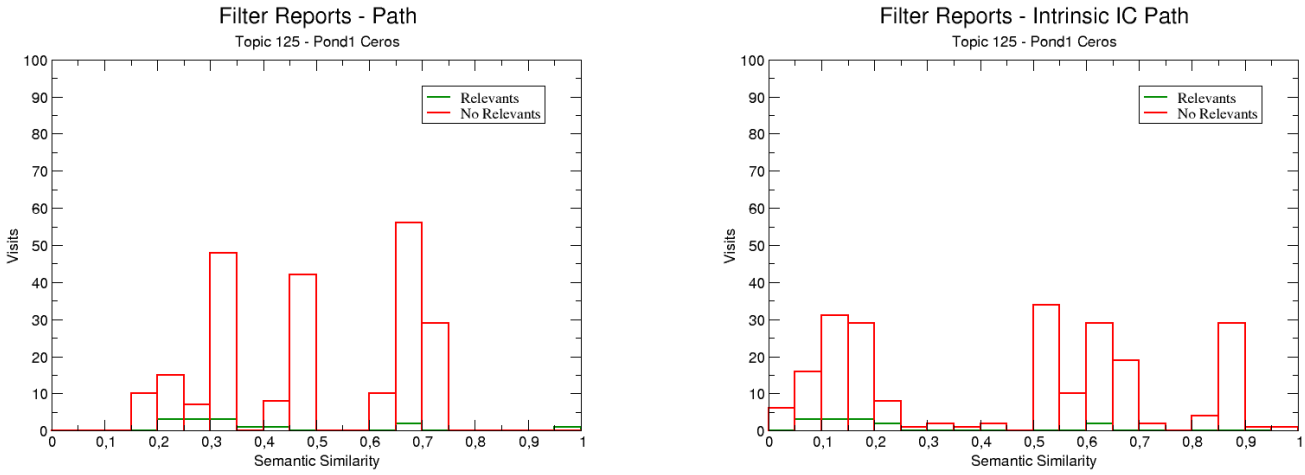

Figura 6.30 Histograma Máxima Similitud *Topic 125 - Path* e *Intrinsic IC-Path*

TOPIC 126: Patients admitted with a diagnosis of multiple sclerosis

PATH 126				Corte 0.6 Visitas
RELEV.	5	Aciertos (>=0,6):	100%	5
		Fallos (<0,6):		0
NO RELEV.	180	Aciertos(<0,6):	66.7%	120
		Fallos (>=0,6):		60

INTRINSIC IC-PATH 126				Corte 0.6 Visitas
RELEV.	5	Aciertos (>=0,6):	20%	1
		Fallos (<0,6):		4
NO RELEV.	180	Aciertos(<0,6):	92.8%	167
		Fallos (>=0,6):		13

Tabla 6.32 Resultados Máxima Similitud *Topic 126 - Path* e *Intrinsic IC-Path*

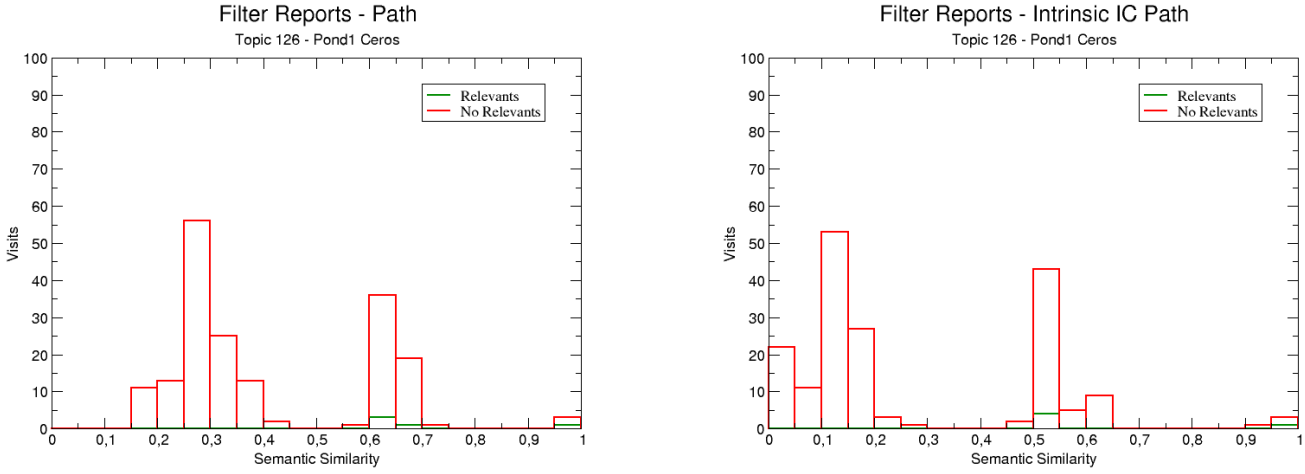

Figura 6.31 Histograma Máxima Similitud *Topic 126 - Path* e *Intrinsic IC-Path*

TOPIC 127: Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension

PATH 127				Corte 0.6 Visitas
RELEV.	85	Aciertos (>=0,6):	94.1%	80
		Fallos (<0,6):		5
NO RELEV.	165	Aciertos(<0,6):	42.4%	70
		Fallos (>=0,6):		95

INTRINSIC IC-PATH 127				Corte 0.6 Visitas
RELEV.	85	Aciertos (>=0,6):	85.9%	73
		Fallos (<0,6):		12
NO RELEV.	165	Aciertos(<0,6):	57.0%	94
		Fallos (>=0,6):		7

Tabla 6.33 Resultados Máxima Similitud *Topic 127 - Path* e *Intrinsic IC-Path*

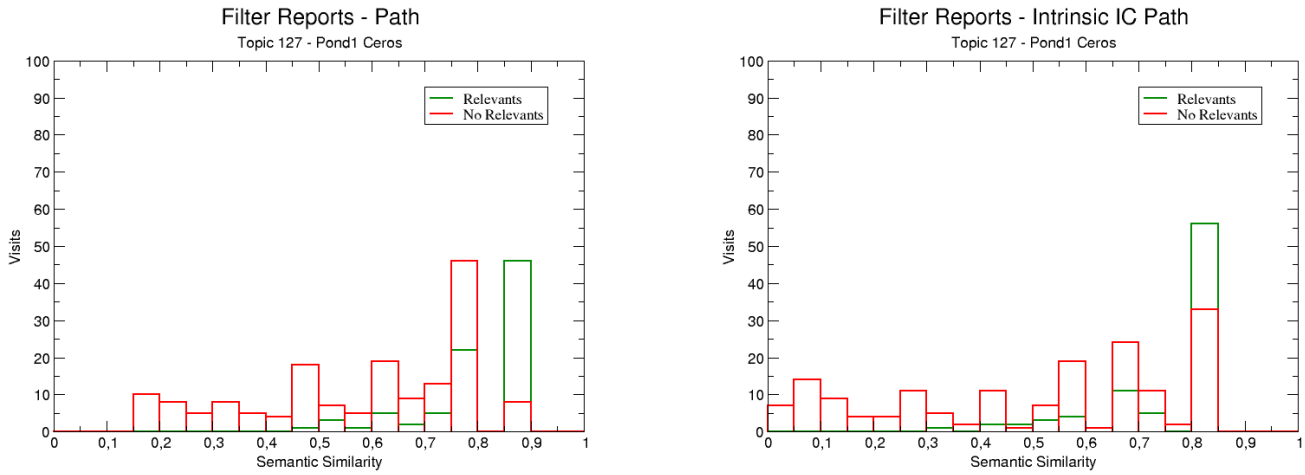

Figura 6.32 Histograma Máxima Similitud *Topic 127 - Path* e *Intrinsic IC-Path*

TOPIC 128: Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op

PATH 128				Corte 0.6 Visitas
RELEV.	85	Aciertos (>=0,6):	29.4%	25
		Fallos (<0,6):		60
NO RELEV.	239	Aciertos(<0,6):	87.5%	209
		Fallos (>=0,6):		30

INTRINSIC IC-PATH 128				Corte 0.6 Visitas
RELEV.	85	Aciertos (>=0,6):	16.5%	14
		Fallos (<0,6):		71
NO RELEV.	239	Aciertos(<0,6):	95.0%	227
		Fallos (>=0,6):		13

Tabla 6.34 Resultados Máxima Similitud *Topic 128 - Path* e *Intrinsic IC-Path*

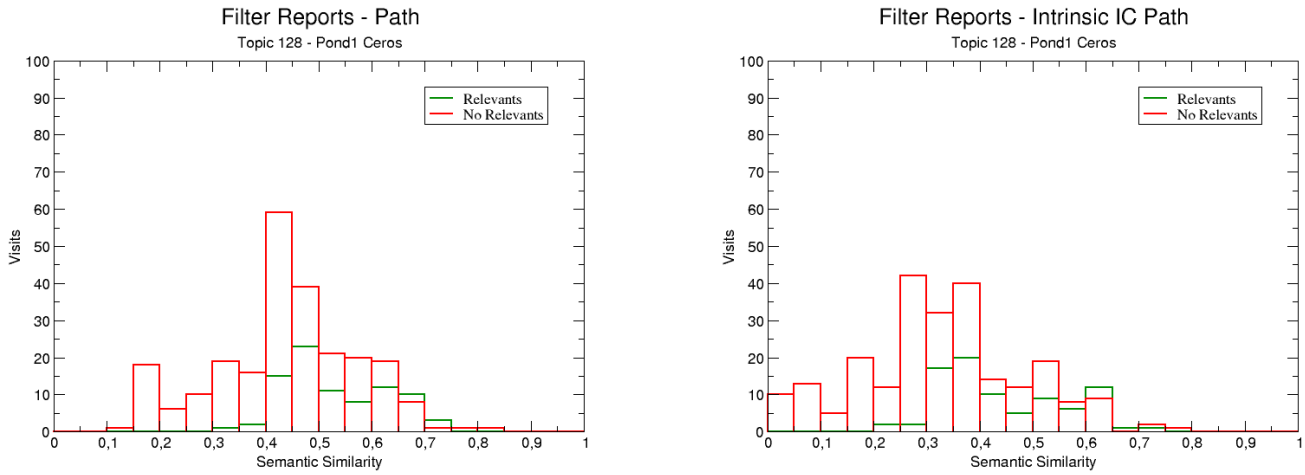

Figura 6.33 Histograma Máxima Similitud *Topic 128 - Path* e *Intrinsic IC-Path*

TOPIC 129: Patients admitted with chest pain and assessed with CT angiography

PATH 129				Corte 0.6 Visitas
RELEV.	53	Aciertos (>=0,6):	94.3%	50
		Fallos (<0,6):		3
NO RELEV.	271	Aciertos(<0,6):	55.4%	150
		Fallos (>=0,6):		121

INTRINSIC IC-PATH 129				Corte 0.6 Visitas
RELEV.	53	Aciertos (>=0,6):	58.5%	31
		Fallos (<0,6):		22
NO RELEV.	271	Aciertos(<0,6):	81.5%	221
		Fallos (>=0,6):		50

Tabla 6.35 Resultados Máxima Similitud *Topic 129 - Path* e *Intrinsic IC-Path*

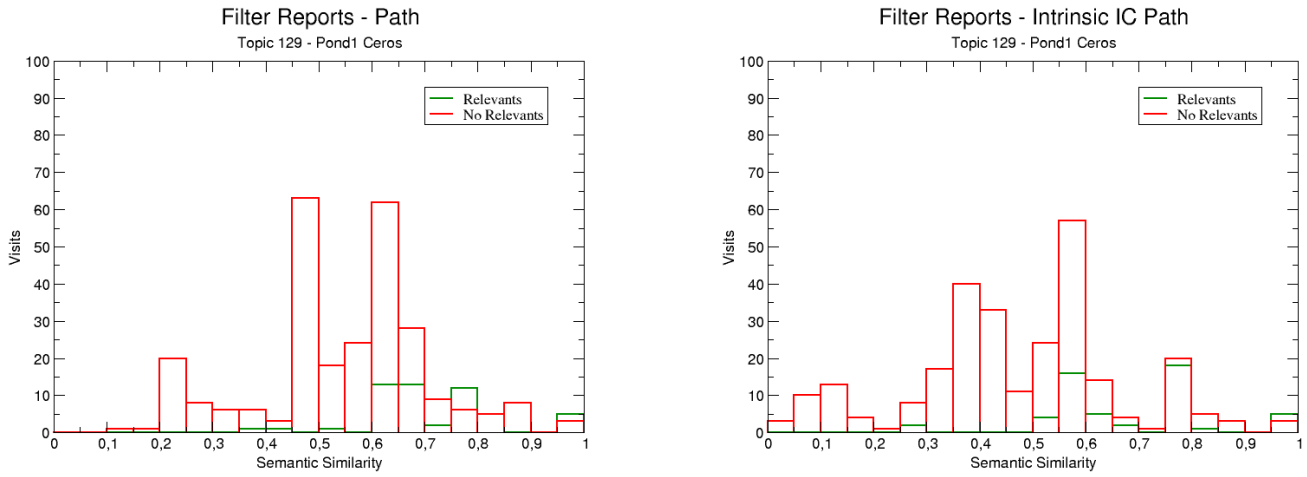

Figura 6.34 Histograma Máxima Similitud *Topic 129 - Path* e *Intrinsic IC-Path*

TOPIC 130: Children admitted with cerebral palsy who received physical therapy

PATH 130				Corte 0.6 Visitas
RELEV.	1	Aciertos (>=0,6):	100%	1
		Fallos (<0,6):		0
NO RELEV.	278	Aciertos(<0,6):	68.0%	189
		Fallos (>=0,6):		89

INTRINSIC IC-PATH 130				Corte 0.6 Visitas
RELEV.	1	Aciertos (>=0,6):	100%	1
		Fallos (<0,6):		0
NO RELEV.	278	Aciertos(<0,6):	94.6%	263
		Fallos (>=0,6):		15

Tabla 6.36 Resultados Máxima Similitud *Topic 130 - Path* e *Intrinsic IC-Path*

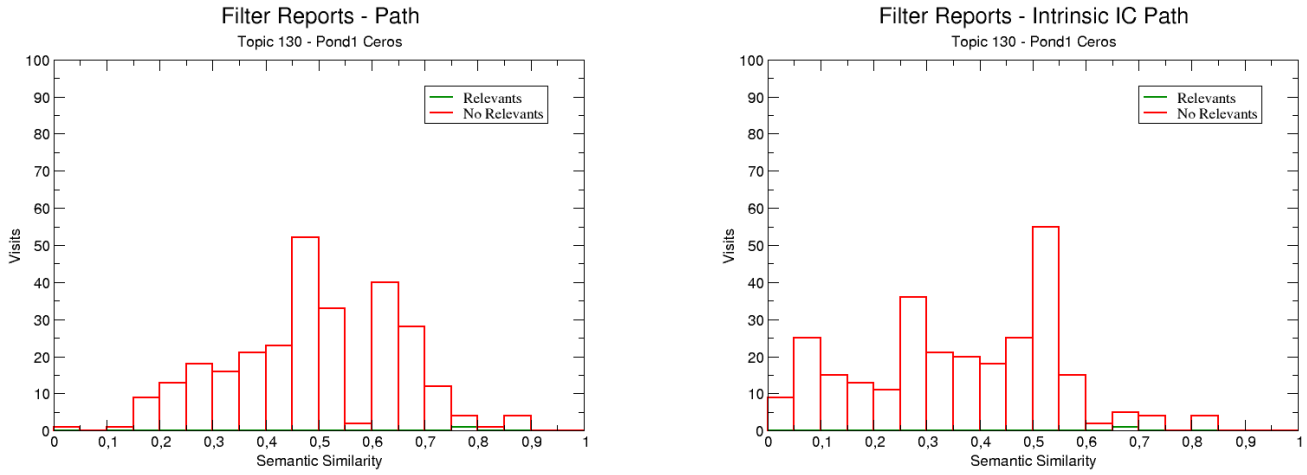

Figura 6.35 Histograma Máxima Similitud *Topic 130 - Path* e *Intrinsic IC-Path*

TOPIC 131: Patients who underwent minimally invasive abdominal surgery

PATH 131				Corte 0.6 Visitas
RELEV.	99	Aciertos (>=0,6):	100%	99
		Fallos (<0,6):		0
NO RELEV.	236	Aciertos(<0,6):	19.9%	47
		Fallos (>=0,6):		189

INTRINSIC IC-PATH 131				Corte 0.6 Visitas
RELEV.	99	Aciertos (>=0,6):	98.0%	97
		Fallos (<0,6):		2
NO RELEV.	236	Aciertos(<0,6):	35.6%	84
		Fallos (>=0,6):		152

Tabla 6.37 Resultados Máxima Similitud *Topic 131 - Path* e *Intrinsic IC-Path*

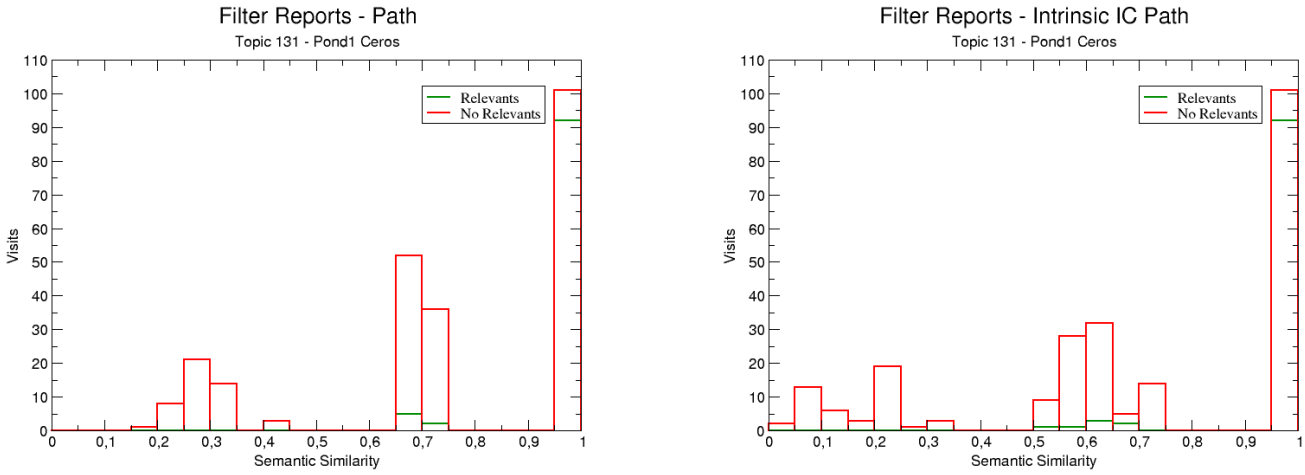

Figura 6.36 Histograma Máxima Similitud *Topic 131 - Path* e *Intrinsic IC-Path*

TOPIC 132: Patients admitted for surgery of the cervical spine for fusion or discectomy

PATH 132				Corte 0.6 Visitas
RELEV.	94	Aciertos (>=0,6):	93.6%	88
		Fallos (<0,6):		6
NO RELEV.	131	Aciertos(<0,6):	84.0%	110
		Fallos (>=0,6):		21

INTRINSIC IC-PATH 132				Corte 0.6 Visitas
RELEV.	94	Aciertos (>=0,6):	91.5%	86
		Fallos (<0,6):		8
NO RELEV.	131	Aciertos(<0,6):	87.8%	115
		Fallos (>=0,6):		16

Tabla 6.38 Resultados Máxima Similitud *Topic 132 - Path* e *Intrinsic IC-Path*

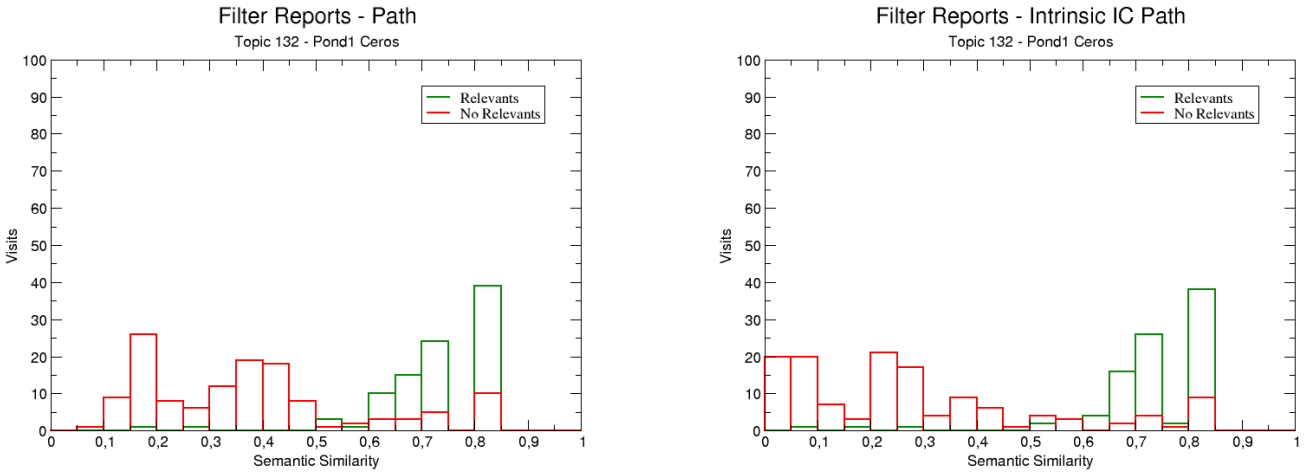

Figura 6.37 Histograma Máxima Similitud *Topic 132 - Path* e *Intrinsic IC-Path*

TOPIC 133: Patients admitted for care who take herbal products for osteoarthritis

PATH 133				Corte 0.6 Visitas
RELEV.	**20**	**Aciertos (>=0,6):**	**85.0%**	**17**
		Fallos (<0,6):		3
NO RELEV.	**321**	**Aciertos(<0,6):**	**34.6%**	**111**
		Fallos (>=0,6):		210

INTRINSIC IC-PATH 133				Corte 0.6 Visitas
RELEV.	**20**	**Aciertos (>=0,6):**	**85.0%**	**17**
		Fallos (<0,6):		3
NO RELEV.	**321**	**Aciertos(<0,6):**	**36.8%**	**118**
		Fallos (>=0,6):		203

Tabla 6.39 Resultados Máxima Similitud *Topic 133 - Path* e *Intrinsic IC-Path*

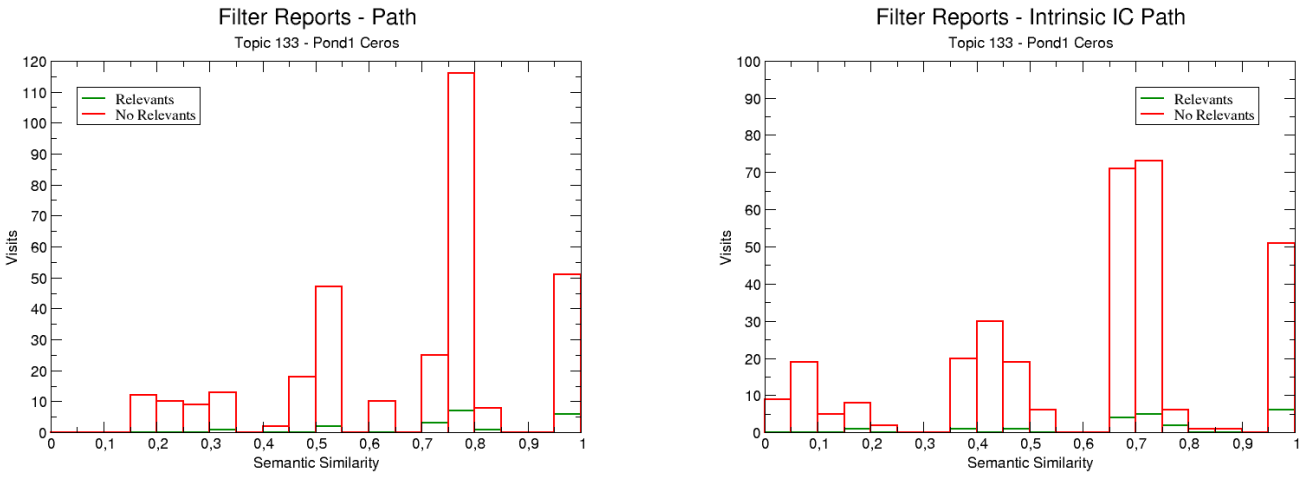

Figura 6.38 Histograma Máxima Similitud *Topic 133 - Path* e *Intrinsic IC-Path*

TOPIC 134: Patients admitted with chronic seizure disorder to control seizure activity

PATH 134				Corte 0.6 Visitas
RELEV.	**34**	**Aciertos (>=0,6):**	**0%**	**0**
		Fallos (<0,6):		34
NO RELEV.	**233**	**Aciertos(<0,6):**	**100%**	**233**
		Fallos (>=0,6):		0

INTRINSIC IC-PATH 134				Corte 0.6 Visitas
RELEV.	**34**	**Aciertos (>=0,6):**	**0%**	**0**
		Fallos (<0,6):		34
NO RELEV.	**233**	**Aciertos(<0,6):**	**100%**	**233**
		Fallos (>=0,6):		0

Tabla 6.40 Resultados Máxima Similitud *Topic 134 - Path* e *Intrinsic IC-Path*

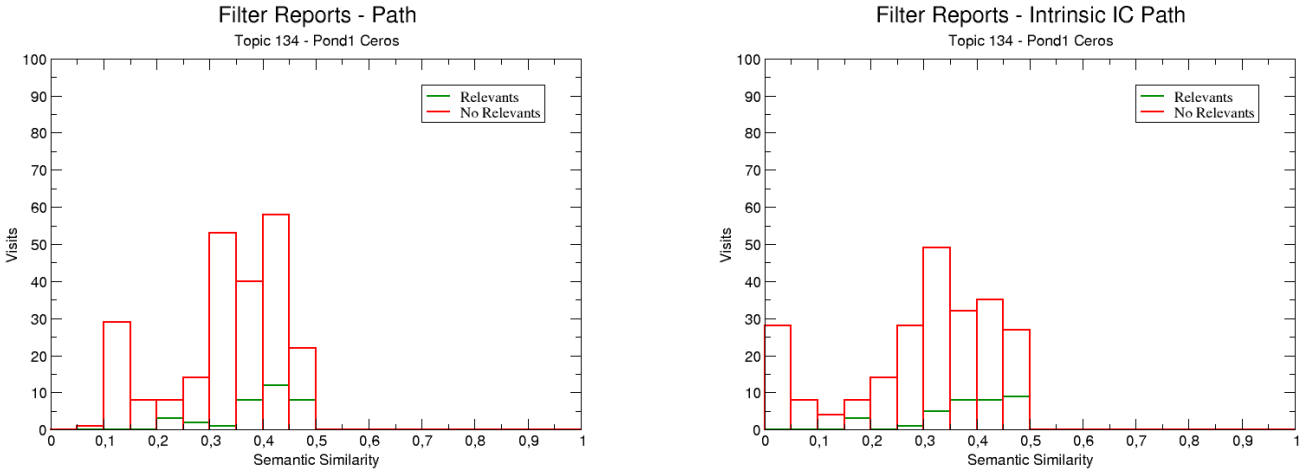

Figura 6.39 Histograma Máxima Similitud *Topic 134 - Path* e *Intrinsic IC-Path*

<u>TOPIC 135</u>: Cancer patients with liver metastasis treated in the hospital who underwent a procedure

PATH 135				Corte 0.6 Visitas
RELEV.	**59**	**Aciertos (>=0,6):**	**64.4%**	**38**
		Fallos (<0,6):		21
NO RELEV.	**218**	**Aciertos(<0,6):**	**78.4%**	**171**
		Fallos (>=0,6):		47

INTRINSIC IC-PATH 135				Corte 0.6 Visitas
RELEV.	**59**	**Aciertos (>=0,6):**	**78.0%**	**46**
		Fallos (<0,6):		13
NO RELEV.	**218**	**Aciertos(<0,6):**	**83.5%**	**182**
		Fallos (>=0,6):		36

Tabla 6.41 Resultados Máxima Similitud *Topic 135 - Path* e *Intrinsic IC-Path*

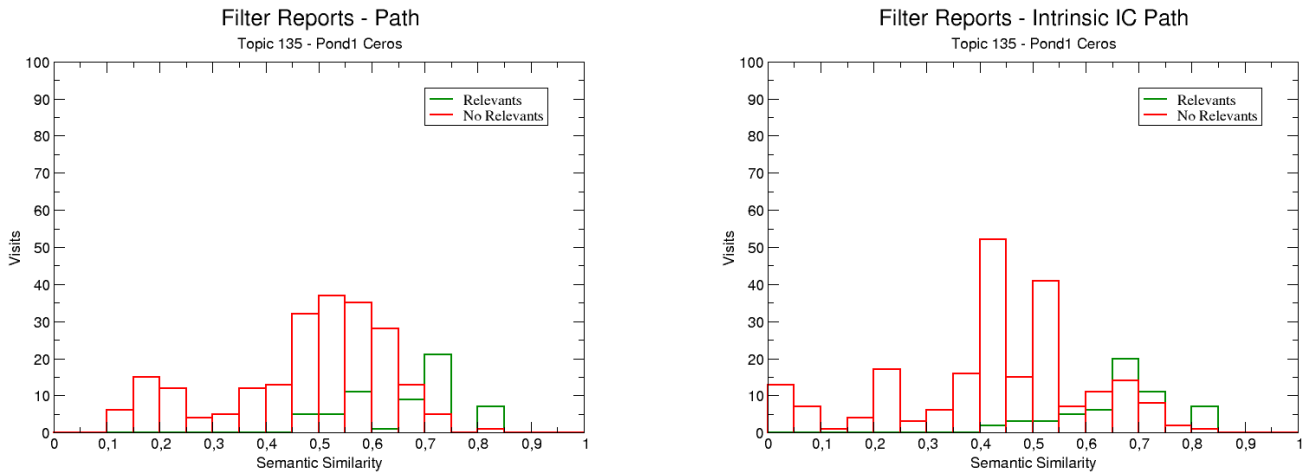

Figura 6.40 Histograma Máxima Similitud *Topic 135 - Path* e *Intrinsic IC-Path*

6.5.2. Resultados Finales Agregados

A continuación se muestra los resultados finales agregados. Las métricas utilizadas para la evaluación son las estándares en el campo de la recuperación de información: *Precision*, *Recall* y *F-Measure*. Siendo esta última la que recoge mejor el equilibrio entre las dos anteriores. Dichas métricas se definen como:

$$Precision = \left(\frac{Nº\ de\ doc.relevantes\ recuperados}{Nº\ total\ de\ doc.recuperados}\right) \tag{6.5}$$

$$Recall = \left(\frac{Nº\ de\ doc.relevantes\ recuperados}{Nº\ de\ doc.relevantes\ en\ la\ colección}\right) \tag{6.6}$$

$$F\_Mesaure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{6.7}$$

Primero, se muestran los valores de *Precision*, *Recall* y *F-Measure,* para ambas métricas (*Path* e *Intrinsic IC-Path*), en cada uno de los criterios de búsqueda evaluados (*topics*) sobre el conjunto de documentos médicos (*reports*) del TREC Medical Track (Tabla 6.42 y Tabla 6.43).

Resultados Métrica PATH

	101	102	103	104	105	106	107	108	109	110	111	112
Recall	82,4%	100,0%	50,0%	88,9%	92,4%	95,3%	21,7%	69,2%	95,1%	91,6%	23,8%	89,0%
Precision	74,4%	38,5%	31,6%	44,4%	93,1%	37,0%	41,7%	15,5%	52,2%	70,2%	55,6%	58,0%
F-Mesaure	78,2%	55,6%	38,7%	59,3%	92,7%	53,3%	28,6%	25,4%	67,4%	79,5%	33,3%	70,3%

	113	114	115	116	117	118	119	120	121	122	123
Recall	92,9%	85,5%	72,2%	100,0%	95,5%	98,1%	26,1%	99,1%	92,5%	91,7%	0,0%
Precision	16,0%	47,5%	47,3%	8,6%	72,4%	22,7%	57,1%	44,3%	25,5%	17,3%	0,0%
F-Mesaure	27,4%	61,0%	57,1%	15,9%	82,4%	36,8%	35,8%	61,2%	40,0%	29,1%	0,0%

	124	125	126	127	128	129	130	131	132	133	134	135
Recall	100,0%	21,4%	100,0%	94,1%	29,4%	94,3%	100,0%	100,0%	93,6%	85,0%	0,0%	64,4%
Precision	3,3%	3,1%	7,7%	45,7%	45,5%	29,2%	1,1%	34,4%	80,7%	7,5%	0,0%	44,7%
F-Mesaure	6,4%	5,4%	14,3%	61,5%	35,7%	44,6%	2,2%	51,2%	86,7%	13,8%	0,0%	52,8%

Tabla 6.42 Resultados finales métrica *Path* (Recall; Precision; F-Measure) para cada *topic*.

Resultados Métrica INTRINSIC IC-PATH

	101	102	103	104	105	106	107	108	109	110	111	112
Recall	86.5%	78.7%	75.0%	77.8%	93.1%	84.7%	91.3%	15.4%	35.0%	8.4%	52.4%	89.0%
Precision	68.8%	50.0%	37.5%	63.6%	93.1%	43.4%	39.6%	14.3%	72.9%	61.5%	40.7%	80.2%
F-Mesaure	76.6%	61.1%	50.0%	70.0%	93.1%	57.4%	55.3%	14.8%	47.3%	14.8%	45.8%	84.4%

	113	114	115	116	117	118	119	120	121	122	123
Recall	92.9%	78.2%	33.3%	60.0%	95.5%	98.1%	0.0%	99.1%	97.5%	58.3%	0.0%
Precision	37.1%	69.4%	46.2%	7.6%	43.8%	26.0%	0.0%	47.7%	19.7%	19.4%	0.0%
F-Mesaure	53.1%	73.5%	38.7%	13.5%	60.0%	41.1%	0.0%	64.4%	32.8%	29.2%	0.0%

	124	125	126	127	128	129	130	131	132	133	134	135
Recall	83.3%	21.4%	20.0%	85.9%	16.5%	58.5%	100.0%	98.0%	91.5%	85.0%	0.0%	78.0%
Precision	4.0%	3.4%	7.1%	91.3%	51.9%	38.3%	6.3%	39.0%	84.3%	7.7%	0.0%	56.1%
F-Mesaure	7.6%	5.9%	10.5%	88.5%	25.0%	46.3%	11.8%	55.7%	87.8%	14.2%	0.0%	65.2%

Tabla 6.43 Resultados finales métrica *Intrinsic* IC-*Path* (Recall; Precision; F-Measure) para cada *topic*.

En la Tabla 6.44, se muestra la media de todos los resultados en la recuperación de aquellos documentos médicos (*reports*) relevantes respecto a cada uno de los criterios de búsqueda propuestos (*topics*). Como se observa en la tabla, el valor de *F-Measure* de ambas métricas son muy similares (*Path* = 0.430 e *Intrinsic IC-Path* = 0.427), reflejando una pequeña mejora la métrica *Path*. Aunque como se muestra en los resultados, en términos de *Recall* se podría considerar a *Path* como mejor métrica y de Precisión a *Intrinsic IC-Path*, no es un resultado concluyente ya que ambos indicadores son complementarios.

	Path	***Intrinsic IC-Path***
Recall	0.753	0.639
Precision	0.364	0.392
F-Measure	0.430	0.427

Tabla 6.44 Resultados Agregados (Recall; Precision; F-Measure) para ambas métricas.

Profundizando en estos resultados y analizando su dispersión, Figura 6.41 y Figura 6.42. Se muestran los resultados detallados de los anteriores indicadores para todos los *topics*, sobre ambas métricas objeto de estudio. Esta figura revela la complejidad de algunos *Topics* (como el 116, 123, 124, 125, 126, 130, 133 o 134) en los cuales los resultados en términos de *F-Measure* se encuentran por debajo del 20%, para ambas métricas. Unos ejemplos son, el *Topic 123* ("*Diabetic patients who received diabetic education in the hospital*") o el *Topic 133* ("*Patients admitted for care who take herbal products for osteoarthritis*").

Los *Topics 123* y *134* producen un resultado completamente anómalo, debido a un error detectado en las relaciones UMLS para dos conceptos concretos. Estos conceptos son, *"C0241863 – Diabetic"* para el *Topic 123* y "*C1148454 – Seizure activity"* en el *Topic 134*, no ofrecen ningún distancia de similitud y son particularmente importantes para dichas consultas o *topics*.

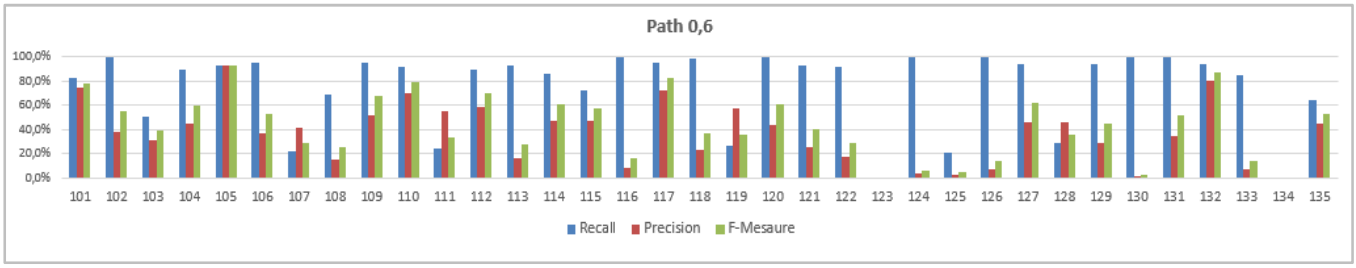

Figura 6.41 Resultados métrica *Path* para los 135 Topics. *Recall, Precision y F-Measure.*

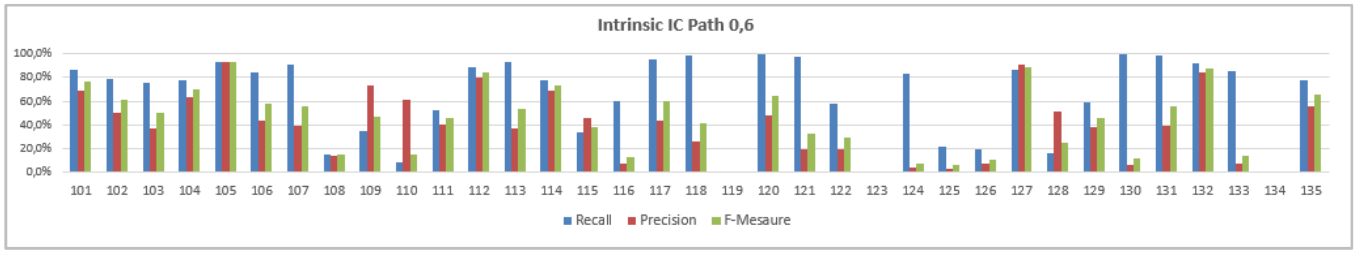

Figura 6.42 Resultados métrica *Intrinsic IC-Path* para los 135 Topics. *Recall Precision* y *F-Measure*.

6.5.3. Comparativa con propuestas del Medical Records Track 2011

A continuación, se mostrará una comparativa de los resultados obtenidos por el sistema de recuperación propuesto en este trabajo para ambas métricas de similitud semántica, en comparación con los mejores resultados enviados por los diversos trabajos recogidos en el Medical Records Track 2011 (TREC).

El TREC 2011, recibió un total de 127 ejecuciones resultantes de los 29 grupos participantes en dicha conferencia, de los cuales 109 eran ejecuciones automáticas y 18 ejecuciones con intervención manual. Las ejecuciones enviadas fueron evaluadas para 34 de las consultas o *topics*, (el conjunto de consultas inicial estaba formado por 35 "*querys*", pero el TREC tomó la decisión de eliminar el *Topic 130*, ya que únicamente contenía una única *visita* relevante de entre un total de 279 evaluadas). Las medidas oficiales aplicadas para la evaluación de los resultados propuestos en la conferencia del TREC, son las siguientes:

- ***Bpref***: Es una medida aplicada a los documentos juzgados que pretende contabilizar en el *ranking* resultante, cuantos documentos juzgados como relevantes (siendo realmente no relevantes) son recuperados antes de los documentos juzgados como relevantes de manera correcta.

$$bpref = \frac{1}{R}\sum_{r}\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}\right) \tag{6.8}$$

R= número de documentos juzgados como Relevantes.
N= número de documentos juzgados como no Relevantes.
r= documento relevante recuperado.
n= conjunto de documentos recuperados del conjunto R, siendo no relevantes.

- **R-Precision**: Para esta medida, es necesario conocer todos aquellos documentos relevantes para una consulta dada (*R)*. Se tomará *R,* como punto de corte de los *r* documentos relevantes recuperados en el *ranking* establecido. En resumen, son los *r* documentos relevantes recuperados entre el *top-R.*

$$R - Precision = \left(\frac{r}{R}\right) \tag{6.9}$$

- **Precision at 10 (P10):** Corresponde a la proporción de documentos relevantes recuperados *r*, para un punto de corte definido en el *top-n* (*n* corresponde a los 10 primeros resultados).

$$P10 = \left(\frac{r}{n}\right) \tag{6.10}$$

Las universidades y entidades que conformaron los grupos participantes del TREC 2011 fueron las siguientes:

- Australian e-Health Research Center
- Beijing Univ. of Posts and Telecommunications
- Bibliomics and Text Mining Group
- Cengage Learning
- CentrumWiskunde & Informatica
- CSIRO
- Dokuz Eylul University
- Erasmus Medical Center
- Fraunhofer SCAI
- Inst. for Research in Informatics of Toulouse
- Kobe University
- Mayo Clinic
- Merck KGaA
- NICTA - National ICT Australia
- Oregon Health & Science University
- RMIT University
- U. Amsterdam, U. Twente, Erasmus U.
- US National Library of Medicine
- Universities of Huelva, Europea de Madrid and Vigo
- University College Dublin (UCD CSI)
- University of Colorado School of Me
- University of Delaware (Fang)
- University of Delaware (Carterette)
- University of Glasgow (Terrier)
- University of Iowa (UI ICTS)
- University of Iowa (UIowaS)
- University of Michigan
- University of Texas at Dallas
- York University

Todas las ejecuciones enviadas por los anteriores participantes, fueron evaluadas con las tres medidas previamente citadas (*bpref*, *P10* y *R-Precision*), sobre los 34 topics (exceptuando el topic 130). La Tabla 6.45, recoge los valores medios de los ocho mejores resultados obtenidos entre todas las ejecuciones (manuales y automáticas) propuestas en el TREC 2011.

Ejecuciones Manuales				**Ejecuciones automáticas**			
Run	**bpref**	**P10**	**R-Prec**	**Run**	**bpref**	**P10**	**R-Prec**
NLMManual	0.658	0.727	0.500	CegageM11R3	0.552	0.656	0.440
buptpris01	0.474	0.547	0.342	SCAIMED7	0.552	0.603	0.425
IRITm1QE1	0.462	0.488	0.344	UTDHLTCIR	0.545	0.603	0.422
SCAIMED1	0.457	0.506	0.324	udelgn	0.522	0.544	0.407
UCDCSIrun3	0.456	0.459	0.324	WWOCorrect	0.494	0.415	0.306
mayolbrst	0.426	0.279	0.220	uogTrDeNIo	0.493	0.568	0.401
ohsuManAll	0.379	0.582	0.328	NICTA6	0.490	0.503	0.355
merckkgaamer	0.275	0.459	0.247	EssieAuto	0.482	0.497	0.337

Tabla 6.45 Evaluación de resultados para las ocho mejores ejecuciones manuales y automáticas ordenadas por *bpref*.

A continuación, en las Tablas 6.46 y 6.47, se muestran los resultados obtenidos para el sistema de recuperación propuesto en este trabajo de investigación sobre ambas métricas de similitud semántica (*Path* e *Intrinsic IC-Path*). Primeramente se reflejarán los valores obtenidos para las distintas medidas propuestas en el TREC 2001 (*bpref*, *P10* y *R-Precision*), aplicadas a cada consulta o *topic*. Seguidamente se mostraran los resultados finales medios de cada medida para las métricas *Path* e *Intrinsic IC-Path* y se incorporarán a la lista de ejecuciones automáticas propuestas en el TREC 2011 para su comparación.

Resultados Métrica PATH con medidas estándar del TREC 2001

	101	102	103	104	105	106	107	108	109	110	111	112
R-Precision	0.824	0.607	0.500	0.778	0.924	0.553	0.217	0.692	0.626	0.768	0.238	0.836
Precision at 10	1.000	1.000	0.600	0.700	1.000	0.900	0.500	0.600	1.000	0.700	0.500	1.000
bpref	0.851	0.610	0.778	0.828	0.972	0.524	0.800	0.605	0.478	0.854	0.421	0.972

	113	114	115	116	117	118	119	120	121	122	123
R-Precision	0.571	0.673	0.611	0.600	0.955	0.308	0.261	0.658	0.750	0.583	0.000
Precision at 10	0.800	0.700	0.100	0.600	1.000	0.400	1.000	1.000	1.000	0.700	0.000
bpref	0.675	0.699	0.549	0.000	0.777	0.000	0.646	0.675	0.580	0.093	0.000

	124	125	126	127	128	129	130	131	132	133	134	135
R-Precision	0.167	0.071	0.200	0.682	0.294	0.453	0.000	0.929	0.830	0.300	0.000	0.644
Precision at 10	0.400	0.100	0.200	1.000	0.400	0.500	0.100	1.000	1.000	0.600	0.000	0.900
bpref	0.000	0.000	0.000	0.729	0.574	0.367	0.000	0.809	0.913	0.000	0.000	0.927

Tabla 6.46 Resultados para *Path* en cada *topic*, evaluados con las medidas del TREC (bpref, P10; R-Precision)

Resultados Métrica INTRINSIC IC-PATH con medidas estándar del TREC 2001

	101	102	103	104	105	106	107	108	109	110	111	112
R-Precision	0.824	0.551	0.500	0.667	0.931	0.482	0.522	0.154	0.350	0.084	0.524	0.836
Precision at 10	1.000	1.000	0.600	0.700	1.000	0.900	0.500	0.200	1.000	0.700	0.700	1.000
bpref	0.984	0.725	0.469	0.878	0.972	0.463	0.528	0.000	0.990	0.083	0.469	0.980

	113	114	115	116	117	118	119	120	121	122	123
R-Precision	0.571	0.745	0.333	0.500	0.955	0.231	0.000	0.658	0.625	0.292	0.000
Precision at 10	0.800	0.900	0.100	0.500	1.000	0.300	0.000	1.000	1.000	0.700	0.000
bpref	0.515	0.866	0.083	0.389	0.727	0.000	0.000	0.714	0.269	0.000	0.000

	124	125	126	127	128	129	130	131	132	133	134	135
R-Precision	0.167	0.071	0.200	0.659	0.165	0.434	0.000	0.929	0.851	0.300	0.000	0.576
Precision at 10	0.100	0.100	0.100	0.900	0.400	0.500	0.100	1.000	1.000	0.600	0.000	0.800
bpref	0.000	0.000	0.200	0.801	0.505	0.258	0.000	0.926	0.922	0.000	0.000	0.695

Tabla 6.47 Resultados para *Intrinsic IC-Path* en cada *topic*, evaluados con las medidas del TREC (bpref, P10; R-Precision)

Seguidamente, se muestran los resultados anteriores de forma gráfica, Figura 6.43 y 6.44, donde se puede observar el comportamiento de ambas métricas (*Path* e *Intrinsic IC-Path*, evaluadas con las medidas estándar propuestas en el Medical Records Track (TREC 2011)

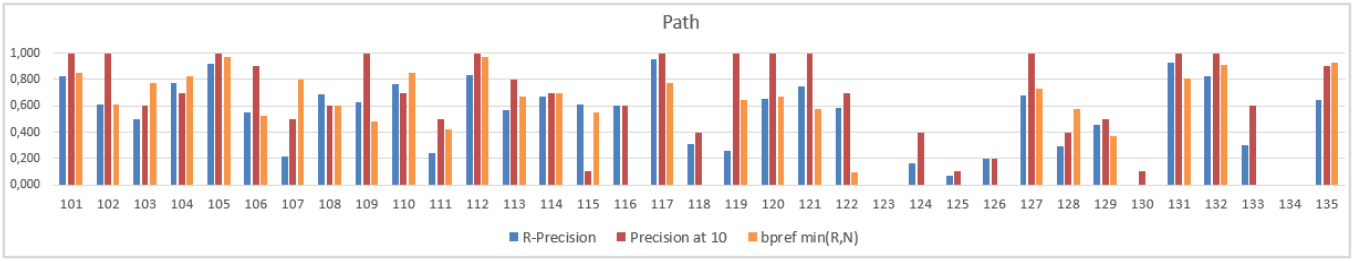

Figura 6.43 Resultados métrica *Path* para los 135 Topics. *P10, R-Precision y bpref*.

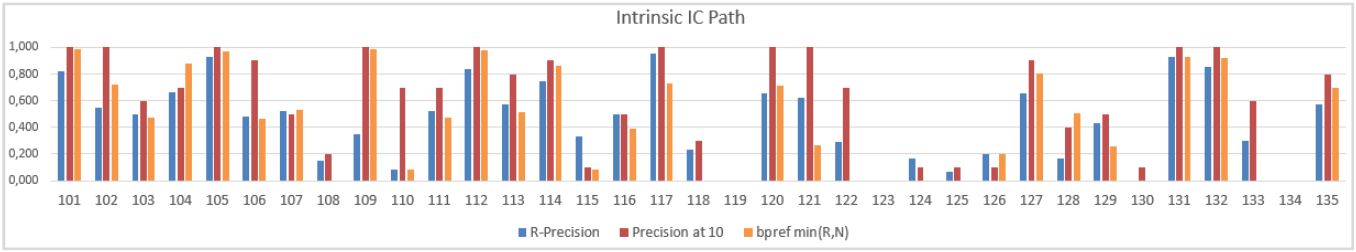

Figura 6.44 Resultados métrica *Intrinsic-IC Path* para los 135 Topics. *P10, R-Precision y bpref*.

A continuación, se muestran las valores medios para cada una de las tres medidas estándares del TREC 2011, sobre el *ranking* de resultados obtenidos por el sistema de recuperación aquí propuesto en la evaluación de cada uno de los *topics*. En la Tabla 6.48, se muestran los valores medios para todas las consultas o *topics* del TREC 2011. Mientras que en la Tabla 6.49, se ha excluido la consulta o *topic 130*, tal y como se realizó con las evaluaciones de los diferentes trabajos remitidos al TREC 2011.

	Path	***Intrinsic IC-Path***
R-Precision	0.517	0.448
Precision at 10	0.657	0.606
bpref	0.506	0.440

Tabla 6.48 Resultados agregados para los 35 *Topics*

	Path	***Intrinsic IC-Path***
R-Precision	0.532	0.461
Precision at 10	0.585	0.621
bpref	0.521	0.453

Tabla 6.49 Resultados agregados, excluido el *Topic 130* del conjunto.

Los resultados de la Tabla 6.49 son integrados en el conjunto de las 109 ejecuciones automáticas enviadas por los diferentes grupos de investigación al TREC 2011, para así comparar los resultados obtenidos en la propuesta de esta investigación, con respecto a las mejores ejecuciones o resultados del TREC 2011, Tabla 6.50.

Ejecuciones automáticas			
Run	**bpref**	**P10**	**R-Prec**
CegageM11R3	0.552	0.656	0.440
SCAIMED7	0.552	0.603	0.425
UTDHLTCIR	0.545	0.603	0.422
udelgn	0.522	0.544	0.407
Path	***0.521***	***0.585***	***0.532***
WWOCorrect	0.494	0.415	0.306
uogTrDeNIo	0.493	0.568	0.401
NICTA6	0.490	0.503	0.355
EssieAuto	0.482	0.497	0.337
Intrinsic IC-Path	***0.453***	***0.621***	***0.461***

Tabla 6.50 Resultados de las mejores ejecuciones del TREC 2011, ordenadas por *bpref* e incluidos los resultados del sistema de recuperación propuesto basado en las métricas *Path* e *Intrinsic IC-Path*.

Como se puede observar a partir de los valores medios de las tres medidas estándar aplicadas a las principales ejecuciones enviadas al Medical Records Track – TREC 2011, los resultados ofrecidos por el sistema de recuperación de documentos médicos propuesto para ambas métricas de similitud semántica, se encuentran entre las 10 mejores ejecuciones (de un total de 109). Corroborando de esta manera, el mejor comportamiento de la métrica *Path*, frente a *Intrinsic IC-Path*.

Esta tabla refleja una leve mejora en el *ranking* de ordenación de los resultados ofrecidos por la métrica *Path* para las medidas *bpref* y *R-Precisión*. Sin embargo, tomando como referencia los diez primeros resultados, la métrica *Intrinsic IC-Path* muestra unos resultados ligeramente mejores.

6.6. Conclusiones

Como resumen de este capítulo, podemos decir que se ha realizado un novedoso experimento que ha permitido la evaluación del comportamiento de las métricas *Path* e *Intrinsic IC-Path* en un contexto real de información biomédica, basada en documentos médicos. Para ello, ha sido necesario crear una serie de procedimientos que transformen la información expresada en lenguaje natural tanto de los documentos médicos, como los criterios de búsqueda, en representaciones basadas en conceptos. Estas representaciones basadas en conceptos, se definen en estructuras de frases y subfrases que permiten la expansión no solo de la consulta, sino también del documento, optimizando los resultados de las matrices de máxima similitud.

Aprovechando estas ventajas, se ha desarrollado un entorno que integre el uso del conocimiento del Metatesauro UMLS, en un sistema de recuperación de documentación médica basada en conceptos, a través de su matriz de máxima similitud. Para la eficiencia de esta sistema de recuperación, se ha mostrado necesario el proceso de filtrado por tipos semánticos ante la necesidad de eliminaran aquellos conceptos poco relevantes, que pueden distorsionar o perjudicar los resultados finales de similitud.

El comportamiento de las métricas en este ámbito, se puede definir de la siguiente manera:

La métrica *Path* muestra una distribución discreta y escalonada hacia los extremos como propia definición de su métrica. Sin embargo, la naturaleza de la métrica *Intrinsic IC-Path*, pierde ese carácter discreto, haciendo la distribución de sus resultados más suavizada hacia los extremos. Sin embargo, los resultados globales para ambas métricas suelen ser muy similares. Además se observa que cuanto más complejo es el criterio de búsqueda, los resultados en ambas métricas tienden a representarse como distribuciones normales.

Por último, se concluye que los resultados agregados de ambas métricas en contexto reales (grandes volúmenes de información), tienen un comportamiento muy similar.

- *Path* (F-Measure=0.430)
- *Intrinsic IC-Path* (F-Measure=0.427).

Lo que nos lleva a determinar que los experimentos propuestos en entornos teóricos y no interconectados, no tienen el mismo comportamiento que en entornos reales. Llegando de esta manera a la conclusión de que la complejidad computacional de la métrica *Intrincic IC-Path,* no justifica su aplicación en dichos contextos.

Además esto viene a corroborar en cierta medida, la incertidumbre en las diferentes interpretaciones que se realizaron en estudios previos sobre entornos teóricos, al utilizar una única correlación (Spearman) representativa del orden de los resultados propuestos por las métricas frente a los propuestos por expertos. Mientras que la correlación Pearson, representa como de aproximados son los valores aportados por las métricas con respecto a los aportados por los expertos. Siendo este criterio el más razonable, a la hora de utilizar en un sistema que trata de emular el comportamiento de un experto al identificar la similitud semántica más próxima, entre documentos médicos y criterios de búsqueda.

Capítulo 7

7. Otras aplicaciones de la representación basada en conceptos UMLS: Generación Automática de Resúmenes

Como se ha visto en el trabajo anterior, la representación basada en conceptos tanto de los documentos médicos, como de la consulta, han sido un elemento de importancia que ha permitido el desarrollo de un sistema de recuperación biomédica aplicado a la similitud semántica contenida. En esta sección, se pretende valorar la capacidad semántica del Metatesauro UMLS y dar otro uso y validez a la representación y expansión basada en conceptos UMLS propuesto en este trabajo, sobre las frases originales de un documento especializado en el ámbito de la biomedicina.

Esta representación basada en conceptos de un documento, es la base inicial para la generación automática de su resumen mediante la aplicación de técnicas o enfoques basados en la extracción de información.

7.1 Trabajos previos en la Generación Automática Resúmenes

Estudios recientes, demuestran como el exceso de información puede provocar en el desempeño de nuestras tareas un resultado contrario a nuestras expectativas, obteniendo un rendimiento menor del esperado, junto con un agravante añadido de estrés y ansiedad (Klingberg, 2008). Es por ello que surgen nuevas técnicas e investigaciones que pretenden eliminar los problemas asociados a la sobrecarga de información, mediante la generación automática de resúmenes que sean de utilidad para el usuario final. Estos resúmenes pueden ser de ayuda en tareas como la recuperación de información o agrupación de documentos, ofreciendo la posibilidad de ahorrar esfuerzos y mejorar los resultados obtenidos. (Mani et al., 2001); (Saggion, Lloret & Palomar, 2010).

Los resúmenes en definitiva son una transformación de un documento o documentos originales en otro nuevo, donde su contenido se ve reducido mediante un proceso de *selección* o *extensión* de la información más relevante (Spark, 1999). Para realizar este proceso de reducción o resumen de un texto, es necesario identificar tres factores: La ***entrada***, el ***propósito*** y la ***salida*** del resumen.

- Los ***factores de entrada*** corresponden a las características del texto a resumir. Como son su forma en base a su estructura, tamaño y estilo utilizados, o la especialización del texto incluido y el volumen de documentos que formarán parte del resumen a generar.
- Los ***factores de propósito*** definen el objetivo final de dicho resumen. Como es el entorno o público al que irá dirigido y lo más importante, si este resumen será de utilidad al usuario para comprender o identificar la información que puede ser de mayor interés o la de crear un documento completamente nuevo y resumido.
- Por último, los ***factores de salida*** determinan las características del texto a obtener en el resumen. Como podría ser la cobertura del texto original, es decir, sobre todo o partes de éste. Así como la representación del resumen, como un texto completamente redactado y continuo, o expresado mediante diferentes secciones. Estas salidas pueden ser resúmenes que aporten al usuario las ideas principales del contenido más relevante (*resúmenes indicativos o informativos*) o incluir nueva información no existente en el documento original (*resúmenes críticos o agregativos*).

Como consecuencia de los anteriores problemas y necesidades citadas, surge el desarrollo de sistemas de generación automática de resúmenes. Estos se pueden clasificar de una manera generalizada en función de sus técnicas aplicadas, en aquellos que hacen uso de **abstracción** de información y los que aplican la **extracción** de información. Los sistemas de generación de resúmenes basados en la abstracción, pueden incluir texto y contenidos no presentes en el documento original, de manera que requerirán de técnicas de reescritura y compresión del texto original (Banko & Vanderwende, 2004). Mientras que los sistemas basados en la extracción de información, se componen de texto o información contenida en el documento original. Ciertos experimentos basados en la generación de resúmenes realizados por personas, han mostrado que la mayor parte del texto incluido (alrededor del 81%), corresponden a frases exactas contenidas en los documentos originales (Jing, 2002).

Los resúmenes basados en estas técnicas pueden tener por tanto, dos funcionalidades principales (Mani, 2001):

- Abstraer un documento final resumido que puede ser tomado como sustituto del documento original.
- Extraer un conjunto de referencias fundamentales que permitan dar una visión de la temática o ideas principales contenidas en dicho documento.

Actualmente algunos de los mayores esfuerzos, se han asociado a la generación de resúmenes mediante las técnicas basadas en la **extracción** de información, donde se han aplicado diferentes enfoques en función del análisis de documentos, como son las técnicas heurísticas, posición de frases, repetición de términos, relación entre palabras o estructuras complejas. La **abstracción** conlleva una mayor profundidad del estudio realizado sobre el contenido original del documento para su transformación y reescritura mediante técnicas de generación de lenguaje natural (el cual no es el objetivo principal de este estudio).

Ambas técnicas en su proceso de generación de resúmenes, realizan una transformación del documento original, en su paso por dos fases que son el *análisis y síntesis* de su contenido (Hahn & Mani, 2000). De esta manera, las **técnicas basadas en la abstracción**, se componen de una primera fase de análisis **consistente en la representación semántica del texto mediante conceptos genéricos y sus relaciones**, para finalmente aplicar una fase de síntesis fundamentada en la generación de lenguaje natural que defina la reescritura del resumen final. Mientras que las **técnicas basadas en la extracción**, en su fase de análisis se centra en la recopilación de **segmentos principales en el texto** que serán tratados en la fase de síntesis **para la eliminación de redundancias, incoherencias y referencias entre dichos segmentos**.

Diferentes enfoques para la selección o extracción de segmentos de texto principales han sido estudiados previamente (Paice, 1980); (Paice, 1990), como son por ejemplo:

- **Frecuencia de palabras**
  Este en un sencillo enfoque que se fundamenta en la búsqueda de ocurrencias de palabras o expresiones que representan la información principal del documento. Previamente se suele utilizar una lista de palabras a eliminar (*stop words*) sin significado independiente pleno, pasando después a estudiar la frecuencia del resto de palabras. De estas palabras se descartarán aquellas que tengan una frecuencia menor a un umbral establecido (Sparck Jones, 1972); (Hovy & Lin, 1998); (Teufel & Moens, 1997), otros trabajos puntúan la frecuencia de grupos de palabras (Luhn, 1958); (Tombros & Sanderson, 1998) con la idea de identificar oraciones relevantes. La simplicidad y el amplio uso de esta heurística, no ha hecho decaer su aplicación en trabajos recientes en combinación con otras técnicas (Lloret et al., 2008).
- **Organización del documento**
  El enfoque basado en la organización de los documentos, se centra en la idea preconcebida de la importancia de ciertas estructuras como son los títulos, subtítulos y encabezados. Estas estructuras reflejan las ideas más relevantes del documento y son la referencia para la extracción de *palabras principales*, en la heurística de búsqueda de frases con dichos términos principales (Teufel & Moens, 1997). Trabajos más recientes aplican esta idea a contextos de documentos actuales como es el correo electrónico (Carenini & Zhou, 2008); (Wan & McKeown, 2004). Otros trabajos amplían esta propuesta mediante el cálculo de la similitud semántica entre ciertas palabras del título del documento y sus oraciones contenidas, haciendo uso de la métrica de Jiang y Conrath (1997).
- **Posición del texto**
  La idea principal de este enfoque, se centra en la localización de frases en zonas del documento donde se puedan considerar de mayor importancia. Diversos trabajos han estudiado las estructuras de los documentos para determinar cuáles son las zonas cuyas frases presentaban una mayor relevancia informacional, (Kupiec, Pedersen, & Chen, 1995); (Teufel y Moens, 1997). Llegando a la consecución de que las principales zonas corresponden con los encabezados de secciones tales como “Resumen” o “Conclusiones”, así como al principio y final de un documento (Bawakid & Oussalah, 2008). Esta heurística al igual que las anteriores, sigue siendo utilizada pero no como un único enfoque de selección de segmentos de

texto, sino como apoyo a otros enfoques de selección. Por ejemplo, este enfoque de la posición del texto, es utilizado como elemento de pesado para el cálculo de la importancia de las frases (Bossard et al., 2008).

- **Expresiones señalizadoras**
  Este enfoque, se particulariza en un conjunto de palabras de referencia, que aportan información de la importancia de la frase valorada. Para ello se utilizan ciertos *corpus* de palabras denominadas *bonus* o *stigma*. Las palabras *bonus* referencian posibles frases con un contenido importante en su texto, al contrario que las palabras *stigma* (Rush et al. 1971); (Kupiec et al. 1995). Este enfoque es muy dependiente del dominio aplicado, de manera que ciertos *corpus* especializados en un dominio no deberán ser aplicables a otros dominios ya que obtendrán un peor rendimiento. Por ello, este enfoque tiene un alcance limitado, por lo que su uso no es muy amplio.
- **Aprendizaje Automático**
  Ciertos enfoques actuales, aplican técnicas y algoritmos basados en el Aprendizaje Automático. Estos requerirán de ciertos *corpus* de entrenamiento que permitan "aprender" el peso o relevancia asignables a los atributos o palabras contenidas en el texto de las frases. Para ello se aplican técnicas de clasificación bayesiana (Kupiec et al., 1995), atributos de representación de frases (Neto et al., 2002), modelos de regresión y árboles de decisión (Metzler & Kanungo, 2008), entre otros.

Todos los anteriores enfoques en los que se fundamenta las *técnicas basadas en la extracción*, comparten la ventaja de su simplicidad frente a los basados en *técnicas de abstracción*. Las técnicas basadas en extracción son aptas para textos no muy extensos, donde se pretende obtener las principales ideas o referencias de dicho documento, ya que por lo general los resúmenes reeditados como un único documento con estas técnicas pueden carecer de cohesión y coherencia es su resultado.

Sin embargo, las técnicas basadas en la abstracción, requieren un complejo análisis del lenguaje natural contenido en dichos documentos que permita seleccionar la información más relevante de cara a la rescritura completa del resumen final. Por tanto, para estas técnicas se hace imprescindible el análisis de las relaciones entre términos, así como su cohesión y coherencia. Algunos trabajos tratan de estudiar los términos adyacentes o cadenas léxicas con ciertas asociaciones de importancia entre ellas (Hearst, 1997); (Morris & Hirst, 1991). Trabajos más recientes traducen dichos términos en base a conceptos UMLS que posteriormente son encadenados en función de su pertenencia a un mismo tipo semántico (Reeve et al., 2007).

Otras mejoras, incorporan a dicho estudio entre términos ciertas relaciones de cohesión entre ellos como son: la repetición, sinonimia, hiperonimia, antonimia y holonimia. Siendo para ello necesario, el uso de recursos del conocimiento léxico en el entorno del lenguaje natural como WordNet. También es común para estos enfoques, utilizar grafos para la representación y análisis de la cohesión entre términos (Mani, 2001) (Agirre & Soroa, 2009) (Plaza, Díaz & Gervás, 2011). Al igual que las representaciones basadas en árboles, como estructuras para el análisis de la coherencia entre términos (Marcu, 1999); (Marcus, 2000) (Radev et al., 2000).

7.2 Generación de resúmenes basados en la *frecuencia de conceptos* y *agregación de frases*

En esta sección, se valora la capacidad semántica del Metatesauro UMLS para dar validez a la representación y expansión basada en conceptos UMLS (sección 6.1) del texto contenido en documentos especializados en el ámbito de la biomedicina. Esta representación de un documento, fundamenta la base inicial para la generación automática de su resumen mediante la aplicación de técnicas o enfoques basados en la extracción de información.

Particularmente en la idea propuesta en esta sección, se parte de un proceso inicial de *análisis* basado en la **abstracción** de un documento biomédico mediante su expansión en conceptos UMLS. A la que posteriormente, se le aplicará un proceso de *síntesis* basado en la **extracción** de segmentos de texto relevantes en función de sus características, como son la frecuencia de conceptos (CUIs) que las representan y la repetición de las frases textuales asociadas. De esta manera, aquí se combina la aplicación de ambas técnicas, en las fases de *análisis* y *síntesis.*

7.2.1. Procesado previo de los documentos médicos para su adaptación al resumen automatizado

A continuación, se enumeran y describen cada una de las fases previas necesarias para el desarrollo del proceso de generación de resúmenes automatizados propuesto, mediante la aplicación combinada de las técnicas de abstracción de conceptos y la extracción del texto relevante.

Filtrado de texto irrelevante

Antes de realizar el proceso de transformación del texto original de un documento a conceptos UMLS, es necesario desarrollar un paso previo de limpieza de aquellos elementos que no son de importancia para la generación de su resumen final (al igual que en el punto 6.1). Los elementos filtrados son por ejemplo, las cabeceras del documento, las etiquetas del lenguaje XML, además de aquellas etiquetas utilizadas para anonimizar los datos relativos al paciente o personal médico, como: [NAME], [DATE], [M.D.], etc.

A continuación, se muestra un ejemplo del documento correspondiente al *report32288*.

ENT CONSULT
HP
CHEST PAIN

Otorhinolaryngology
Consultation Report

CHIEF COMPLAINT: Right nasal obstruction.
HISTORY OF PRESENT ILLNESS: This patient had an episode of vomiting today and the vomitus came out his right nose. Since that time, he has had a feeling of it being swollen.
PHYSICAL EXAMINATION:
GENERAL: The patient is an alert male in no acute distress.
NOSE: The nasal cavity is open bilaterally with no foreign body identified.
IMPRESSION: Irritation secondary to the acid in his vomitus.

RECOMMENDATION: Ocean spray 2 sprays twice a day for 4 days.

Figura 7.1 Ejemplo documento médico original con filtrado previo. Report32288.

Transformación basada en conceptos

Una vez eliminados aquellos elementos textuales irrelevantes, posteriormente los documentos son transformados para expresar su información en base a conceptos UMLS, junto a sus tipos semánticos correspondientes. Esta transformación será la base inicial para el proceso de generación automático de resúmenes, en función de su frecuencia o repetición de conceptos.

Para realizar dicha transformación, este proceso se apoya en la herramienta MetaMap, la cual permite obtener los distintos conceptos (CUIs) que representan el contenido textual del documento. Los resultados ofrecidos por MetaMap son posteriormente tratados para generar una estructura compleja de *frases* y *subfrases*, a partir de sus conceptos UMLS asociados. Con este procesamiento, se consigue trasformar el documento médico original en una representación basada en conceptos y tipos semánticos, donde cada frase está conformada por sus diferentes subfrases que vienen dadas por las diferentes combinaciones de CUIs que las representan.

A continuación, se muestra como ejemplo la representación del *report32288.* Esta representación se compone del *identificador de frase* y *subfrase*; *negación*; *CUI*; *tipo semántico* y *frase textual* (se ha eliminado el *identificador de frase*, para su simplificación).

Report32288.txt

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0029892 bmod Otorhinolaryngology Consultation Report
3 0 C1269802 inpr Otorhinolaryngology Consultation Report
4 0 C0205090 spco Right nasal obstruction.
4 0 C0027429 fndg Right nasal obstruction.
5 0 C0030705 podg This patient
6 0 C0332189 tmco an episode of vomiting today
6 0 C0042963 sosy an episode of vomiting today
6 0 C0310367 antb an episode of vomiting today
7 0 C0332189 tmco an episode of vomiting today
7 0 C0042963 sosy an episode of vomiting today
7 0 C0750526 tmco an episode of vomiting today
8 0 C0332189 tmco an episode of vomiting today
8 0 C1963281 fndg an episode of vomiting today
8 0 C0310367 antb an episode of vomiting today
9 0 C0332189 tmco an episode of vomiting today
9 0 C1963281 fndg an episode of vomiting today
9 0 C0750526 tmco an episode of vomiting today
10 0 C0042965 bdsu the vomitus
11 0 C1547958 inpr the vomitus
12 0 C1608512 bdsu the vomitus
13 0 C0439787 spco out
14 0 C0849355 qlco out
15 0 C0205090 spco his right nose.
15 0 C0028429 bpoc his right nose.
16 0 C0205090 spco his right nose.
16 0 C1278896 bpoc his right nose.
17 0 C1711239 tmco Since
18 0 C0040223 tmco that time,
19 0 C0013987 menp a feeling of it
20 0 C1527305 menp a feeling of it
21 0 C0038999 fndg swollen.
22 0 C0030705 podg The patient
23 1 C0239110 fndg an alert male in no acute distress.
23 1 C0024554 fndg an alert male in no acute distress.

```
23 1 C0205178 tmco an alert male in no acute distress.
23 1 C0231303 menp an alert male in no acute distress.
24 1 C0239110 fndg an alert male in no acute distress.
24 1 C0086582 popg an alert male in no acute distress.
24 1 C0205178 tmco an alert male in no acute distress.
24 1 C0231303 menp an alert male in no acute distress.
25 1 C0239110 fndg an alert male in no acute distress.
25 1 C1706180 qlco an alert male in no acute distress.
25 1 C0205178 tmco an alert male in no acute distress.
25 1 C0231303 menp an alert male in no acute distress.
26 1 C0239110 fndg an alert male in no acute distress.
26 1 C1706428 qlco an alert male in no acute distress.
26 1 C0205178 tmco an alert male in no acute distress.
26 1 C0231303 menp an alert male in no acute distress.
27 1 C0239110 fndg an alert male in no acute distress.
27 1 C1706429 orga an alert male in no acute distress.
27 1 C0205178 tmco an alert male in no acute distress.
27 1 C0231303 menp an alert male in no acute distress.
28 0 C0027423 bsoj The nasal cavity
29 0 C1280672 bsoj The nasal cavity
30 0 C0175566 spco open
31 0 C0016542 inpo bilaterally with no foreign body
32 1 C0205396 qlco identified.
33 1 C1550043 fndg identified.
34 1 C1551388 ftcn identified.
35 0 C0441723 phpr Irritation
36 0 C1706307 fndg Irritation
37 0 C2700617 menp Irritation
38 0 C0175668 tmco secondary to the acid
38 0 C0001128 chem secondary to the acid
39 0 C0042965 bdsu in his vomitus.
40 0 C1547958 inpr in his vomitus.
41 0 C1608512 bdsu in his vomitus.
42 0 C0028814 geoa Ocean spray 2 sprays
42 0 C1704413 qnco Ocean spray 2 sprays
43 0 C0028814 geoa Ocean spray 2 sprays
43 0 C2003858 ftcn Ocean spray 2 sprays
44 0 C0585361 tmco twice a day for 4 days.
```

Figura 7.2 Representación basada en conceptos UMLS del report32288.

Eliminación de CUIs repetidos en cada frase

Una vez transformado el documento en una representación basada en conceptos UMLS para cada una de sus frases. Posteriormente, se realiza un proceso de eliminación de aquellos CUIs repetidos en cada frase analizada. Esta eliminación de CUIs repetidos viene dada por la necesidad de eliminar la multiplicidad provocada por las diferentes combinaciones de CUIs que conforman las variaciones de una misma frase en subfrases, siendo mayor dichas repeticiones cuanto superior es la longitud textual de dicha frase y más sinónimos existan para cada término. Al mismo tiempo, son eliminados todos aquellos conceptos asociados con la negación (con valor "1").

De esta manera, solo se representan los CUIs únicos que componen las diferentes variaciones de frases pero sin repeticiones, eliminando de esta forma la multiplicidad generada por la expansión del documento. A continuación, se muestra como ejemplo el *report32288*, sin repeticiones de CUIs para una misma frase.

Report32288-SinRepeticiones.txt

```
1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0029892 bmod Otorhinolaryngology Consultation Report
3 0 C1269802 inpr Otorhinolaryngology Consultation Report
4 0 C0205090 spco Right nasal obstruction.
4 0 C0027429 fndg Right nasal obstruction.
5 0 C0030705 podg This patient
6 0 C0332189 tmco an episode of vomiting today
6 0 C0042963 sosy an episode of vomiting today
6 0 C0310367 antb an episode of vomiting today
7 0 C0750526 tmco an episode of vomiting today
8 0 C1963281 fndg an episode of vomiting today
10 0 C0042965 bdsu the vomitus
11 0 C1547958 inpr the vomitus
12 0 C1608512 bdsu the vomitus
13 0 C0439787 spco out
14 0 C0849355 qlco out
15 0 C0205090 spco his right nose.
15 0 C0028429 bpoc his right nose.
16 0 C1278896 bpoc his right nose.
17 0 C1711239 tmco Since
18 0 C0040223 tmco that time,
19 0 C0013987 menp a feeling of it
20 0 C1527305 menp a feeling of it
21 0 C0038999 fndg swollen.
22 0 C0030705 podg The patient
28 0 C0027423 bsoj The nasal cavity
29 0 C1280672 bsoj The nasal cavity
30 0 C0175566 spco open
31 0 C0016542 inpo bilaterally with no foreign body
35 0 C0441723 phpr Irritation
36 0 C1706307 fndg Irritation
37 0 C2700617 menp Irritation
38 0 C0175668 tmco secondary to the acid
38 0 C0001128 chem secondary to the acid
39 0 C0042965 bdsu in his vomitus.
40 0 C1547958 inpr in his vomitus.
41 0 C1608512 bdsu in his vomitus.
42 0 C0028814 geoa Ocean spray 2 sprays
42 0 C1704413 qnco Ocean spray 2 sprays
43 0 C2003858 ftcn Ocean spray 2 sprays
44 0 C0585361 tmco twice a day for 4 days.
```

Figura 7.3 Representación del report32288, con eliminación de CUIs repetidos para una frase.

Filtrado de tipos semánticos

La anterior representación basada en conceptos UMLS y con eliminación de CUIs repetidos, requiere de un posterior proceso de filtrado para aquellos conceptos cuya capacidad informacional sea limitada e irrelevante. Dichos conceptos, si no fuesen filtrados, darían lugar a resúmenes finales formados por posibles CUIs de poco interés que degradarían la extracción de información, obviando posibles conceptos más relevantes. Es por ello que en esta etapa, se hace uso de los tipos "*semánticos específicos*" identificados en la sección 6.2, cuya relevancia informacional son de mayor interés.

Para algunos tipos semánticos, dada su especial importancia o características, se seleccionan todas las variaciones de frase para un mismo CUI, como son por ejemplo *elii* o *medd* (sección 6.2).

A partir de este proceso, se obtiene una representación basada en conceptos de un documento biomédico textual, para el que se ha realizado un proceso de eliminación de CUIs repetidos y el filtrado de aquellos conceptos pertenecientes a tipos semánticos

informacionalmente no relevantes. A continuación, se muestra el *report32288* filtrado semánticamente.

Report32288-Filtrado.txt

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0029892 bmod Otorhinolaryngology Consultation Report
4 0 C0027429 fndg Right nasal obstruction.
5 0 C0030705 podg This patient
6 0 C0042963 sosy an episode of vomiting today
6 0 C0310367 antb an episode of vomiting today
8 0 C1963281 fndg an episode of vomiting today
10 0 C0042965 bdsu the vomitus
12 0 C1608512 bdsu the vomitus
15 0 C0028429 bpoc his right nose.
16 0 C1278896 bpoc his right nose.
21 0 C0038999 fndg swollen.
22 0 C0030705 podg The patient
28 0 C0027423 bsoj The nasal cavity
29 0 C1280672 bsoj The nasal cavity
31 0 C0016542 inpo bilaterally with no foreign body
35 0 C0441723 phpr Irritation
36 0 C1706307 fndg Irritation
39 0 C0042965 bdsu in his vomitus.
41 0 C1608512 bdsu in his vomitus.

Figura 7.4 Representación del report32288, con eliminación de tipos semánticos irrelevantes.

7.2.2. Frecuencia de conceptos y repetición de frases

La representación basada en conceptos de un documento biomédico, da un nuevo enfoque a su contenido textual en referencia a su semántica contenida y simbolizada por el Metatesauro UMLS como recurso del conocimiento biomédico. Esta representación se analizará a través de la frecuencia de sus conceptos y la repetición de frases asociadas, reflejando así su capacidad de caracterizar el contexto referenciado.

Para la generación del resumen de un documento basado en la **frecuencia de sus conceptos**, se realiza una evaluación de las repeticiones de cada concepto referenciado en diferentes frases. Siendo este resultado ordenado de mayor a menor frecuencia de repetición para cada concepto, junto a sus frases asociadas. A continuación se muestra el *report32288*, ordenado por la frecuencia de sus conceptos.

Report32288-Filtrado-ord.txt

C1608512 bdsu the vomitus
C1608512 bdsu in his vomitus.
C0042965 bdsu the vomitus
C0042965 bdsu in his vomitus.
C0030705 podg This patient
C0030705 podg The patient
C2926613 clna CHEST PAIN
C1963281 fndg an episode of vomiting today
C1706307 fndg Irritation
C1280672 bsoj The nasal cavity
C1278896 bpoc his right nose.
C0441723 phpr Irritation
C0310367 antb an episode of vomiting today
C0042963 sosy an episode of vomiting today
C0038999 fndg swollen
C0029892 bmod Otorhinolaryngology Consultation Report
C0028429 bpoc his right nose
C0027429 fndg Right nasal obstruction.
C0027423 bsoj The nasal cavity
C0016542 inpo bilaterally with no foreign body

Figura 7.5 Representación del report32288, ordenado por frecuencia de conceptos.

Para la representación textual de cada concepto cuya repetición se localiza en diferentes frases del documento (Figura 7.5), se selecciona la frase de mayor longitud o contenido informacional asociada, junto con el número de repeticiones de dicho concepto o CUI (Figura 7.6). Esta nueva representación, se compone del *CUI*, su *tipo semántico*, la *frase de mayor contenido informacional* y el *número de repeticiones* de dicho CUI (Figura 7.6)

Report32288-Densidad.txt

C1608512 bdsu in his vomitus. (Rep. 2)
C0042965 bdsu in his vomitus. (Rep. 2)
C0030705 podg This patient (Rep. 2)
C2926613 clna CHEST PAIN (Rep. 1)
C1963281 fndg an episode of vomiting today (Rep. 1)
C1706307 fndg Irritation (Rep. 1)
C1280672 bsoj The nasal cavity (Rep. 1)
C1278896 bpoc his right nose (Rep. 1)
C0441723 phpr Irritation (Rep. 1)
C0310367 antb an episode of vomiting today (Rep. 1)
C0042963 sosy an episode of vomiting today (Rep. 1)
C0038999 fndg swollen (Rep. 1).
C0029892 bmod Otorhinolaryngology Consultation Report (Rep. 1)
C0028429 bpoc his right nose (Rep. 1)
C0027429 fndg Right nasal obstruction. (Rep. 1)
C0027423 bsoj The nasal cavity (Rep. 1)
C0016542 inpo bilaterally with no foreign body (Rep. 1)

Figura 7.6 Representación de la frecuencia de repetición de cada concepto en el report32288.

Esta representación del documento médico basado en la frecuencia de sus conceptos, es la base para determinar que conceptos y frases asociadas son las de mayor interés en el resumen final. Para ello, se procederá a valorar la distribución de la frecuencia de conceptos y la posterior agregación de frases repetidas de la anterior representación.

7.2.3. *Distribución de frecuencia de conceptos (densidad cualitativa)*

En esta primera fase se evalúa la incidencia de cada concepto en el documento, de manera que aquellos conceptos (CUIs) con mayor repetición, destacarán por su mayor relevancia informacional en la generación automática del resumen de dicho documento. Descartando por otro lado, aquellos con poca o nula relevancia.

Para poder discernir los conceptos (junto con sus frases asociadas) de mayor relevancia, se seleccionan aquellos cuya repetición esta dentro de un **porcentaje de distribución**. Este porcentaje o punto de corte seleccionado, corresponde a los **dos primeros cuartiles de mayor frecuencia** de repetición de conceptos en el documento.

A continuación, se muestra en la siguiente figura, las repeticiones de conceptos para el Report32288. Que viene conformado por sus *CUIs*, *tipos semánticos*, *frases de mayor contenido informacional* y el *número de repeticiones* de cada CUI dentro del documento. Junto con el punto de corte correspondiente a la descripción anterior. Como se observa en la figura, el objetivo es contabilizar el número de repeticiones de cada CUI en todas las frases del report.

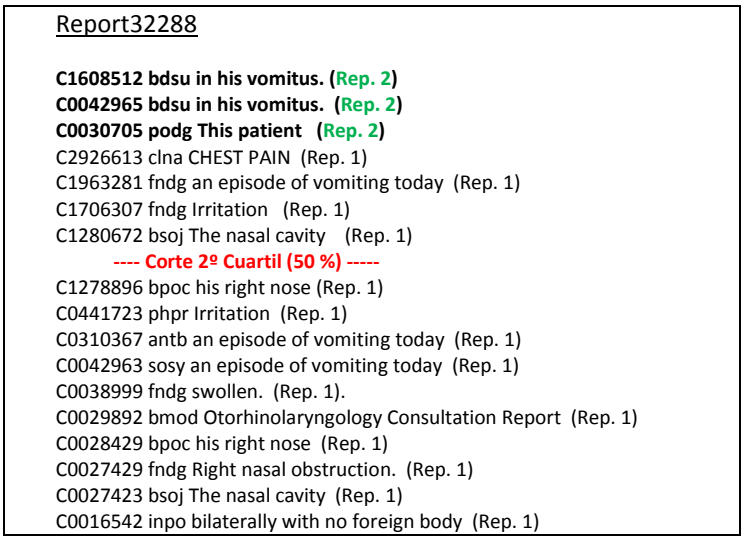

Figura 7.7 Repeticiones de cada concepto (CUI) para el report32288.

Para ampliar esta explicación, se presenta también como caso de ejemplo el Report32496, el cual contiene tres niveles de repetición de CUIS, al contrario del caso anterior en el que solo existían dos niveles de repetición de CUIs.

Report32496

C0521421 bpoc Right face ear pain, (Rep. 3)
C0030705 podg This patient (Rep. 3)
C0013443 bpoc Right face ear pain, (Rep. 3)
C1962975 fndg right lateral face pain. (Rep. 2)
C1289033 bpoc Right face ear pain, (Rep. 2)
C1281591 bpoc Right face ear pain, (Rep. 2)
C0229298 bpoc Right face ear pain, (Rep. 2)
C0030193 sosy noted pain on the right side of his (Rep. 2)
---- Corte 2º Cuartil (50 %) -----
C0015468 sosy right lateral face pain. (Rep. 2)
C0015450 blor Right face ear pain, (Rep. 2)
C2926613 clna CHEST PAIN (Rep. 1)
C1517034 bpoc The external canals, (Rep. 1)
C1444656 fndg indicated. (Rep. 1)
C1316572 clna noted pain on the right side of his (Rep. 1)
C1281592 bpoc Neck - There (Rep. 1)
C1268972 bpoc middle ear (Rep. 1)
C1265570 fndg are within normal limits. (Rep. 1)
C0743626 fndg unknown etiology. (Rep. 1)
C0439048 fndg a very poor historian. (Rep. 1)
C0230025 blor noted pain on the right side of his (Rep. 1)
C0163712 orch relates (Rep. 1)
C0041445 bpoc tympanic membranes, (Rep. 1)
C0029892 bmod Otorhinolaryngology Consultation Report (Rep. 1)
C0027530 blor Neck - There (Rep. 1)
C0026649 orgf good facial nerve movement. (Rep. 1)
C0015462 bpoc good facial nerve movement. (Rep. 1)
C0013456 sosy Right face ear pain, (Rep. 1)
C0013455 bsoj middle ear (Rep. 1)

Figura 7.8 Repeticiones de cada concepto (CUI) para el report32496.

Para dar mayor claridad a lo expresado anteriormente, en la siguiente figura se representa el número de frases (eje de abscisas) en las que aparece cada nivel de repetición de conceptos (eje de ordenadas) ordenado descendentemente. Así, se puede observar con más detalle el punto de corte de los dos primeros cuartiles, comentado anteriormente

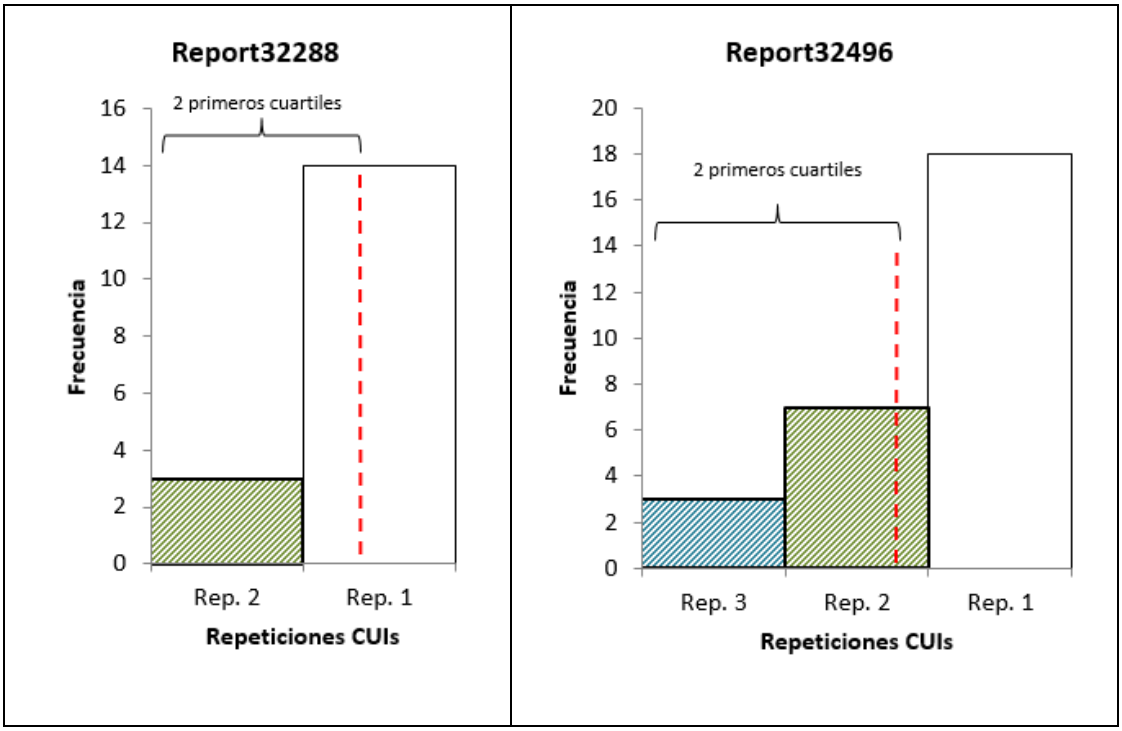

Figura 7.9 Distribución de la frecuencia de conceptos (repeticiones de cuis), para ambos ejemplos.

Como se puede observar en los dos ejemplos de la figura anterior (Figura 7.9), dependiendo del menor nivel de repetición de conceptos, sobre el que se sitúa el corte de de los dos primeros cuartiles, se tomará una acción para la selección de los conceptos (junto con sus frecuencias y frases) que conformarán parte del resumen, de la siguiente manera:

- De forma **general** para todos los casos, se seleccionarán todos aquellos conceptos (CUIs junto con su frase) que se encuentren en el menor nivel de repetición de conceptos y superiores niveles. Así por ejemplo, para el report32496, si el menor nivel de frecuencia final sobre el que recae el corte del segundo cuartil se encuentra en nivel "2". Se seleccionarán todos aquellos CUIs (junto con su frase) desde los mayores niveles de frecuencia, hasta el último concepto perteneciente a la frecuencia de nivel "2".

- De forma **particular**, en aquellos documentos con muy poca información y con el objeto de no incluir CUIs poco relevantes. Si el menor nivel de frecuencia final dentro del segundo cuartil recae en el nivel "1". Se descartará el menor nivel de frecuencia y el **porcentaje de distribución** o **corte,** se establecerá en el nivel de frecuencia superior siguiente. Seleccionando a continuación, todos aquellos CUIs (junto con su frase) desde los mayores niveles de frecuencia, hasta el nuevo nivel de corte, como ocurre para el report32288.

El siguiente proceso de valoración corresponderá a la a*gregación de frases o densidad cuantitativa*, en el que se agregarán las repeticiones de CUIs para una misma frase textual repetida, este proceso se explica a continuación.

7.2.4. *Agregación de frases (densidad cuantitativa)*

Esta sección, corresponde con el último proceso de valoración del documento, antes de la generación del resultado final del resumen. Aquí se realiza la agregación de todos aquellos valores cualitativos obtenidos para una misma frase repetida. Por tanto, en la agregación de frases, se acumulan las frecuencias de CUIs para una misma frase repetida.

Es necesario tener en cuenta que las frases se pueden repetir en el mismo o distintos grupos semánticos. Por tanto, será necesario eliminar aquellas posibles frases textuales repetidas, seleccionando únicamente aquella frase (entre las repetidas) cuya frecuencia sea mayor. En el caso de que exista igualdad entre los valores de frecuencia de las frases repetidas, se seleccionará la correspondiente al grupo semántico con más relevancia (esta situación es indiferente dentro de un mismo grupo, pudiéndose eliminar cualquiera frase repetida)

Así por ejemplo, tenemos como resultado para los dos casos expuestos anteriormente, las siguientes agregaciones de frases o densidades cuantitativas, para cada grupo semántico.

Report32496	
LIVB:	**This patient (Rep. 3)**
ANAT:	**Right face ear pain (Rep. 14)**
DISO:	**right lateral face pain (Rep. 4)** **noted pain on the right side of his (Rep. 2)**

Figura 7.10 Agregación de frases report32796. Acumulación de frecuencia para frases repetidas.

Report32288	
LIVB :	**This patient (Rep. 2)**
ANAT:	**in his vomitus (Rep. 4)**
DISO:	**an episode of vomiting today (Rep. 3)**

Figura 7.11 Agregación de frases report32288. Acumulación de frecuencia para frases repetidas.

Conviene observar también para los casos particulares, en los que el menor nivel de repetición de conceptos, sobre el que se sitúa el corte de distribución corresponde con el nivel de repetición "1" (como ocurre en el r*eport32288*). **Se incluirán aquellas repeticiones de frases** existentes en el nivel de frecuencia igual a "1", siempre y cuando el **valor de agregación** de dichas frases **sea mayor al mínimo valor de frecuencia de CUIs** para el que se establece el punto de corte (es decir, el inmediatamente superior).

Por tanto, para el caso del *report32288*, se añadirán las frases repetidas del nivel de frecuencia "1" (frases en negrita de la Figura 7.12), cuyo valor de agregación de frase sea mayor al mínimo nivel de frecuencia preestablecido (nivel "2", para el ejemplo del r*eport32288*). En este ejemplo, corresponde a la frase "**an episode of vomiting today (Rep.3)**".

Es decir, se han añadido frases con cierta densidad cuantitativa superior al mínimo nivel de frecuencia de CUIs.

Esto es de especial importancia y se actúa de la misma manera, para los casos más desfavorables y menos probables donde la frecuencia de conceptos para toda la distribución es únicamente "1". Por tanto, para estos casos más desfavorables únicamente se añadirán frases que se repitan al menos una vez, es decir, se añadirán frases con cierta densidad cuantitativa superior al mínimo nivel de frecuencia de CUIs, que este caso sería de "1".

Para su mejor comprensión, se muestra como ejemplo el *report32288*. Donde para este caso se, descartan los conceptos para el nivel de frecuencia "1". Pero se incluyen aquellas frases repetidas (frases en negrita) cuyo valor de agregación es mayor al mínimo nivel de frecuencia preestablecido (nivel "2" en este ejemplo).

Report32288

C1608512 bdsu in his vomitus. (Rep. 2)
C0042965 bdsu in his vomitus. (Rep. 2)
C0030705 podg This patient (Rep. 2)
C2926613 clna CHEST PAIN (Rep. 1)
C1963281 fndg **an episode of vomiting today (Rep. 1)**
C1706307 fndg Irritation (Rep. 1)
C1280672 bsoj The nasal cavity (Rep. 1)
---- Corte 2º Cuartil (50 %) -----
C1278896 bpoc his right nose (Rep. 1)
C0441723 phpr Irritation (Rep. 1)
C0310367 antb **an episode of vomiting today (Rep. 1)**
C0042963 sosy **an episode of vomiting today (Rep. 1)**
C0038999 fndg swollen. (Rep. 1).
C0029892 bmod Otorhinolaryngology Consultation Report (Rep. 1)
C0028429 bpoc his right nose (Rep. 1)
C0027429 fndg Right nasal obstruction. (Rep. 1)
C0027423 bsoj The nasal cavity (Rep. 1)
C0016542 inpo bilaterally with no foreign body (Rep. 1)

Figura 7.12 Repeticiones de conceptos y agregación de frases para el report32288.

Estos casos más desfavorables, suelen corresponder con documentos médicos con poca información y muy inconexa.

7.2.5. Filtrado por densidad.

Esta última fase tiene como objeto, eliminar en cada grupo semántico aquellas frases finales con bajos niveles de densidad total (densidad cualitativa+cuantitativa) con respecto al conjunto de frases representadas en dicho grupo. Se procederá a eliminar dentro de cada grupo semántico, aquellas frases para las cuales su valor de agregación de frase sea inferior al 50% del máximo valor de agregación de frase en dicho grupo.

Así por ejemplo, para el *report32496*, si el máximo valor de agregación en un grupo semántico es "14", solo se mostrarían aquellas frases cuyo valor sea igual o superior al 50% de este máximo (es decir aquellas frases cuyo valor de agregación sea igual o mayor de "7"). (Ver ejemplo en Apéndice B, para *report32095*)

Por último, las frases seleccionadas finalmente, serán ordenadas de mayor a menor nivel frecuencia y asociadas a los grupos semánticos que pertenezcan, para la generación automática final del resumen.

A continuación, se muestran los resúmenes para ambos ejemplos.

Resumen del documento "Report32496":

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Right face ear pain (14)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **right lateral face pain (4)**
- **noted pain on the right side of his (2)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **This patient (3)**

Figura 7.13 Resumen final del report32496.

Resumen del documento "Report32288"

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **in his vomitus (4)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **an episode of vomiting today (3)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **This patient (2)**

Figura 7.14 Resumen final del report32288

En el **Apéndice B**, se muestra a modo de ejemplo, una selección de documentos médicos extraídos del TREC, a los que se les aplicaron el proceso automático de generación de resúmenes presentado.

7.2.6. Resultados

A modo de referencia y para comprobar que los resultados obtenidos en este trabajo no eran completamente distantes a un simple proceso de **contabilización de frecuencias de palabras** (sección 7.1). Se ha aplicado una comparativa entre un sencillo proceso de contabilización de frecuencias de palabras, con respecto a la densidad de conceptos y agregación de frases aquí presentado.

De los resultados obtenidos se puede observar, como en los resúmenes automáticos aplicados en ambos ejemplos, se refleja de una manera más completa la importancia de la información contenida. Es decir, sus resúmenes muestran de qué hablan o se refieren dichos documentos.

En ambos ejemplos, dado que corresponden a dos casos sencillos que fueron seleccionados para facilitar la comprensión de las explicaciones. Se puede observar que ciertas palabras coinciden en ambos resultados. Sin embargo, los resultados mostrados por el sistema de generación de resúmenes propuesto, ofrecen un mayor grado de significación semántica y una clasificación por grupos semánticos de los distintos extractos textuales.

Report32496	**Report32288**
5 ear	**2 nasal**
5 face	**2 nose**
4 pain	**2 patient**
4 patient	**2 right**
4 right	**2 vomitus**
2 etiology	----------
2 history	1 acid
----------	1 acute
1 adenopathy	1 cavity
1 cervical	1 chest
1 chest	1 distress
1 disease	1 examination
1 distress	1 illness
1 dizziness	1 irritation
1 hearing	1 male
1 male	1 obstruction
1 mass	1 otorhinolaryngology
1 membranes	1 swollen
1 neck	1 vomiting
1 otorhinolaryngology	
1 tinnitus	
1 tympanic	

Figura 7.15 Proceso de contabilización de frecuencias de palabras para anteriores ejemplos.

En casos de mayor complejidad y volumen de información, (como el *report32095*, incluido en el Apéndice B). Se observa como la información que aporta el resumen final automatizado aquí propuesto, tiene un mayor interés y significado semántico para un

usuario final o aplicación que requiriera de los elementos que representen la información principal de dicho documento médico.

A continuación, se muestra la comparativa del caso complejo *report32095*, seleccionado entre los diferentes casos desarrollados y mostrados en el Apéndice B.

Report32095

13 history
10 patient
9 pain
8 chest
6 artery
6 mg
5 coronary
5 daily
5 disease
3 circumflex
3 distal
3 epigastric
3 graft
3 heard
3 hypertension
3 left
3 left-sided
3 marginal
3 mouth
3 nitroglycerin
3 normal
3 occluded
3 proximal
3 stenosis
3 stenting
2 allergies
2 blood
2 branch
2 cardiovascular
2 circulation
2 complains
2 congestive
2 consultation
2 diabetes
2 distress
2 failure
2 grafts
2 heart
2 medications
2 neck
2 occlusion
2 plavix
2 saphenous
2 stent
2 stented
2 vein
---
1 vomiting
Resto con Rep. "1"
…

Resumen del documento "Report32095":

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **total occlusion of the LAD (11)**
- **Proximal circumflex stenosis (9)**
- **admitted with a history of chest pain (8)**
- **Coronary artery disease (6)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **The patient complains of epigastric distress (10)**
- **a black male, (5)**

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)
- **Plavix 75 mg. (7)**
- **taking three to four nitroglycerin (6)**
- **Lipitor 20 mg by mouth daily. (6)**
- **Protonix 40 mg daily. (6)**
- **Lovenox 40 mg subcu daily. (5)**
- **Metoprolol 200 mg by mouth daily. (5)**

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **concerned about his chest pain (6)**
- **Blood pressure 160/97. (3)**

DISPOSITIVOS MÉDICO, DISPENSADOR DE FÁRMACOS: (DEVI)
- **stented with a Taxus stent (5)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **the remaining saphenous vein grafts to the left-sided (5)**
- **The left internal mammary artery graft to the (4)**
- **normal vesicular breathing heard all over the lung (3)**

PROCEDIMIENTOS TERAPEÚTICOS Y DIAGNÓSTICOS: (PROC)
- **status post stenting** (3)
- **Coronary artery bypass graft about eight years ago (3)**

Figura 7.16 Comparativa frecuencia de palabras, con densidad de conceptos y agregación para un caso complejo, report32095.

7.3 Conclusiones

En esta sección, se muestra un nuevo ámbito de aplicación a la representación basada en conceptos UMLS aportada previamente en este trabajo. Esta representación muestra la base fundamental para la generación automática de resúmenes de documentación biomédica.

Para esta generación de resúmenes basadas en conceptos, se han combinado procesos fundamentales asociados a dos diferentes técnicas como son la **abstracción** y **extracción** de información. Para cada una de estas técnicas, se ha seleccionado un proceso representativo de cada una de ellas, tanto para la fase de análisis como de síntesis de los documentos. De manera que para la generación automática de resúmenes en este trabajo, se aplica un proceso de abstracción del documento basada en conceptos UMLS en su fase de análisis. Combinado con un proceso de extracción de información, basado en densidades de conceptos y frases para la fase de síntesis.

Esta operativa, de extracción de información, permite determinar y definir la importancia del texto final del resumen, a partir de la frecuencia de repetición de un mismo concepto en distintas frases (denominado como *densidad cualitativa*) y la posterior agregación de las frecuencias de repetición de conceptos asociados a una misma representación de frase (denominado como *densidad cuantitativa*).

Tras las pruebas realizadas, los resultados obtenidos son ciertamente satisfactorios ya que como se ha podido ver en los ejemplos anteriores, el sistema propuesto aporta al resumen automático un mayor sentido semántico, además de eliminar aquella información no relevante para el resumen y clasificarla en grupos semánticos.

Con ello, se demuestra nuevamente la validez de la representación basada en conceptos UMLS de los documentos médicos y la capacidad de las técnicas aplicadas para obtener resúmenes automáticos que reflejen la idea principal o la información fundamental de la que trata dicho documento.

Esta última parte del trabajo aquí presentado, pone de manifiesto un paso previo a nuevas mejoras que se podrán incorporar a esta investigación, aplicada en una nueva fase de síntesis del resumen que permita la "generación" de un documento nuevamente reescrito con dichas premisas.

Capítulo 8

8. Conclusiones

La recuperación de documentos médicos a través del procesamiento del lenguaje natural, es lo suficientemente importante y complejo como para dedicarle una atención especial a esta área de investigación. Es por ello, que muchos trabajos publicados abordan el asunto de las métricas de similitud semántica en un contexto teórico (formado por pares de conceptos independientes y cerrados), mediante el apoyo de algunos recursos contenidos en el Metatesauro UMLS. Sin embargo, ninguno de estos trabajos, centra su estudio en un contexto real de recuperación de información biomédica.

Por esta razón, en el trabajo desarrollado en esta tesis, se ha propuesto un nuevo estudio experimental para la evaluación del comportamiento de las métricas *Intrinsic IC-Path* y *Path* en un entorno real de documentación médica, empleando como soporte el recurso UMLS.

Todos los trabajos anteriores se centran en el rendimiento de métricas de similitud semántica, tomando como conjuntos de pruebas pares de conceptos limitados y reducidos (29 pares de conceptos, 101, 430, etc.). Por este motivo, se toman 101,712 documentos médicos (*reports*) recogidos en el desafío del TREC Medical Records Track 2011. Estos *reports* reales de-identificados han sido calificados por expertos médicos a partir de unas palabras claves de búsqueda (*topics*). La dificultad de la evaluación realizada en este trabajo reside en que, a diferencia de la comparación entre pares de conceptos aislados utilizados en los trabajos previos, en este trabajo se ha procedido a comparar *reports* (media de 500 palabras aprox.) contra *topics* (media de 10 palabras).

Para poder realizar este novedoso trabajo experimental de evaluación, surge la necesidad de crear un método específico de recuperación de información basado en la parametrización del Metatesauro UMLS que agregue las similitudes de ambos elementos (*matriz de similitud*) en un único resultado final ("*Relevancia*"/"*No Relevancia*") que se enfrentará con los juicios de relevancia de los expertos del TREC para evaluar el rendimiento de cada una de las métricas. La implementación de este sistema "*ad-hoc*" ha provocado que se haya realizado, en la primera parte del trabajo, un estudio exhaustivo y parametrización del recurso UMLS con el objetivo de obtener una óptima cobertura en los resultados ofrecidos por las distintas métricas de similitud semántica.

Además, este estudio exhaustivo sobre la infraestructura UMLS ha permitido identificar una serie de problemas y diferencias de resultados en los trabajos previos de evaluación de las métricas de similitud semántica en contextos teóricos. A continuación, se enumeran las conclusiones de esta primera parte de la tesis que deberán servir como base de futuros trabajos en este campo.

- **Las versiones del Metatesuro UMLS** en la evaluación de las métricas de similitud semántica han de ser las mismas, ya que **su impacto en los resultados es muy notable.**
- **Las parametrizaciones del Metatesauro UMLS** (diferentes recursos y relaciones aplicadas) **afectan drásticamente a los resultados** aportados por las métricas de similitud semántica. Obteniendo los mejores resultados, cuando se aplican únicamente las **relaciones jerárquicas directas e indirectas** entre conceptos y se usa la **totalidad de recursos** contenidos en el Metatesauro UMLS.
- **Diferentes medidas de correlación** aplicadas a los resultados en trabajos previos, **ofrecen dispares resultados y distintas interpretaciones** de las evaluaciones de las distintas **métricas de similitud semántica.**

De esta manera, se establece un marco confiable de optimización del Metatesauro UMLS en la evaluación de las diferentes métricas de similitud semántica bajo un mismo criterio. Sin embargo, este marco es evaluado únicamente sobre contextos teóricos, formados por conjuntos de pares de conceptos independientes, los cuales pueden condicionar los resultados de las métricas.

Por ello, surge la necesidad de proponer un novedoso sistema de recuperación de información que integre el uso óptimo de la infraestructura UMLS en la aplicación de las métricas de similitud semántica sobre un contexto real de documentación biomédica (basado en el repositorio del TREC). Este sistema permitirá valorar el alcance real de las principales métricas (*Path* e *Intrinsic IC-Path*) sobre un marco único y confiable, dando lugar a las siguientes conclusiones finales:

- Las **variaciones** existentes en los resultados **de las métricas de similitud semántica en contextos teóricos**, **desaparecen en contextos reales**, donde la cantidad de información crece y está relacionada.
    - En contextos teóricos se han usado diversos conjuntos de pares de conceptos aislados y cerrados.
    - En contextos reales, los grandes volúmenes de información expresados en lenguaje natural, son representados por colecciones de conceptos interrelacionados. (*report* y *topic* se relacionan).

- **Ambas métricas muestran un comportamiento similar en un contexto real de información,** al contrario de lo que ocurría en contextos teóricos. De manera que no se justifica la utilización de métricas con un mayor coste computacional como *Intrinsic IC-Path*, en entornos reales.
    - *Path* (F-Measure=0.430).
    - *Intrinsic IC-Path* (F-Measure=0.427).

- La **clasificación de los tipos semánticos** (Genéricos y Específicos), se muestra **necesaria en un contexto real** de recuperación de información.

- Este sistema es **aplicable a cualquier proceso de búsqueda de similitud semántica en documentos textuales asociados al domino biomédico** (historias, resúmenes, notas, informes, imágenes médicas, etc.).

Por último, se ha propuesto un sistema de generación automática de resúmenes sobre documentos médicos, como paso a dos nuevos planteamientos. El primero de ellos, surge como necesidad para validar la utilidad de la representación basada en conceptos de un documento médico presentada en este trabajo, en otros contextos o aplicaciones (como es la generación de resúmenes). El segundo planteamiento, surge como un paso previo a posibles futuras mejoras del sistema de recuperación de información definido y evaluado en el desarrollo de esta tesis.

El sistema de generación automática de resúmenes propuesto refleja las siguientes características:

- Reafirma la **validez de la propuesta** realizada en esta tesis, de la **representación basada en conceptos UMLS de los documentos médicos**, como aplicación en otros contextos de trabajo o necesidades finales.
- La generación de **resúmenes basados en la densidad de conceptos y agregación de frases**, mejoran los resultados propuestos por técnicas tradicionales, aportando un **mayor sentido semántico** al resumen y una **clasificación por grupos semánticos** de los extractos textuales.
- Aporta y refleja la **información principal de un documento médico**, de manera que puede ser de **gran utilidad en futuras mejoras del sistema de recuperación de información** propuesto.

Capítulo 9

9. Aportaciones

Como se ha referido previamente, para la consecución de esta tesis, ha sido necesario identificar las principales características del Metatesauro UMLS que mejor optimicen los resultados de las métricas de similitud semántica. Así como, especificar un sistema de recuperación de información que formalice el uso del Metatesauro UMLS mediante matrices de similitud semántica, en un entorno real de documentos médicos (TREC Medical Records Track 2011).

Las aportaciones finales que se extraen del trabajo e investigación realizada en esta tesis son las siguientes:

- Un marco unificado que permite una comparación fiable de las métricas bajo un mismo contexto teórico de evaluación. En esta tesis se pone de manifiesto las diferencias existentes en los resultados ofrecidos por otros autores en trabajos previos de evaluación de las métricas de similitud semántica en términos de: versiones del Metatesuro UMLS, su correcta parametrización y las distintas medidas de correlación empleadas en la evaluación de resultados.

- Un conjunto de mejoras en la parametrización del Metatesauro UMLS que permite la optimización de los resultados ofrecidos por las distintas métricas de similitud semántica.

- Un sistema de recuperación de información, basado en el Metatesauro UMLS y las métricas de similitud semántica, aplicado a contextos reales de documentación biomédica.

- Un marco de evaluación de las principales métricas de similitud semántica (*Path* e *Intrinsic IC-Path)* en entornos reales de documentación biomédica.

- Un sistema de generación automática de resúmenes de utilidad en documentos médicos basados en una representación de conceptos UMLS, mediante la aplicación de densidad de conceptos y agregación de frases.

Capítulo 10

10. Trabajos futuros

De los capítulos anteriores se puede determinar, que la mejora del rendimiento de ambas métricas de similitud semántica en un entorno real, no tienen el mismo impacto que en entornos cerrados. Por lo que se hace necesario, una mejora en trabajos futuros del comportamiento del sistema de recuperación propuesto en el proceso de esta evaluación, así como la realización de otros estudios complementarios.

- Estudiar con profundidad la estructura interna del Metatesauro UMLS ya que puede que no refleje adecuadamente el conocimiento del dominio, para ciertas consultas tal y como las métricas pretender interpretar. Ciertas consultas pueden mostrar inexactitud u errores en la estructura del Metatesauro UMLS. (Como ocurre en las consultas, 123 y 134)

- Comprobar el posible beneficio de eliminar o independizar aquellas subfrases del tema de búsqueda (*topic*) que no están semánticamente relacionadas.

- Debido a la gran cantidad de información contenida en un documento médico (*report*), ésta puede ser completamente dispar. Sería conveniente filtrar y agrupar su contenido común por diferentes aspectos (*subjects*) en un mismo paciente. Como enfermedades, síntomas, medicaciones, pruebas, etc. Es decir, creando resúmenes de los diferentes aspectos más generales de un paciente. Consiguiendo resumir o clasificar semánticamente un documento médico en la principal información o aspectos tratados de un paciente, se podría mejorar la recuperación de información, eliminando todos aquellos elementos del documento médico, menos relevantes.

- Creación de "*clusters*" de documentos médicos a partir de la información semántica contenida en el propio documento o en los anteriores resúmenes generados.

- Unificar los registros médicos de una única visita en solo documento (ciertos criterios de la consulta pueden estar en distintos documentos del paciente), para posteriormente proceder a su filtrado y agrupamiento por aspectos.

- Utilización códigos ICD9-CM, en el proceso de recuperación, junto al texto libre incorporado en cada documento médico.

Finalmente, se puede decir que ante la problemática de establecer métodos y técnicas efectivas para la recuperación de documentos en entornos reales tan complejos y ambiguos como el de la biomedicina. La aplicación de las métricas de similitud semántica en la representación basada en conceptos tanto de documentos y consultas, han ofrecido unos resultados altamente satisfactorios. Si bien, es necesario una ampliación de este estudio como finalidad de mejora del sistema de recuperación aquí propuesto.

Capítulo 11

11. Trabajos publicados

Las aportaciones resultantes del trabajo de esta tesis, han sido publicadas tanto en una revista de carácter científico, como presentadas en congresos especializados:

Revistas científicas:

- Alonso, I., Contreras, D., ***Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach***.
  International Journal - *Expert Systems with Applications* - Elsevier
  Accepted manuscript (unedited version) available online: 28-SEP-2015
  http://dx.doi.org/10.1016/j.eswa.2015.09.028

  Factor de Impacto JCR: 2.240 (2014).
  Ranked #1 among Artificial Intelligence journals by Google Scholar
  h5-index:91; h5-median:107

Congresos:

- Romero, F.P., Olivas, J.A., Serrano-Guerrero, J., Caballero, I., Oruezábal, M.J., Alonso, I., Contreras, D., ***Aplicación de la lógica borrosa para la calificación de información clínica bajo criterios de calidad de datos***. XVII CONGRESO ESPAÑOL SOBRE TECNOLOGIAS Y LÓGICA FUZZY (ESTYLF 2014), ISBN: 978-84-15688-76-1, pp. 523-528 Zaragoza, España. 5-7 de febrero 2014.

- Alonso, I., Romero, F.P., Contreras, D., Olivas, J.A., ***Una Aplicación de la Similitud Semántica en la mejora de la selección de Información Clínica***. Congreso Español de Informática (CEDI 2013) - (Actas de la Multiconferencia CAEPIA 2013 - IV Simposio sobre Lógica Fuzzy y Soft Computing, LFSC), ISBN: 978-84-695-8348-7, pp. 1173-1181, Madrid, España, 17-20 de Septiembre 2013.

- Alonso, I., Contreras, D., Romero, F.P., ***Aplicación y medida de la similitud semántica presente en registros clínicos electrónicos para la mejora en estudios de cohortes***, XVI Congreso Nacional de Informática de la Salud. (INFORMED 2013), ISBN: 978-84695-7444-7, pp. 270-274 Madrid, España. 9-11 de abril 2013.

- Alonso, I., Contreras, D., Romero, F.P., ***Impacto de las relaciones jerárquicas inter/intra-recursos en la similitud semántica entre conceptos del dominio biomédico***. XVI Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF 2012). ISBN: 978-84-615-6653-2, pp. 438-442. Valladolid, España. 1-3 de febrero 2012.

- Romero, F.P., Ferreira-Satler, M., Alonso, I., Olivas, J.A., Contreras, D., ***Una aproximación a la representación semántica de documentos basada en medidas de similitud entre conceptos***. 14th Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2011). ISBN: 978-3-642-25273-0. La Laguna, Tenerife, España. 7-11 de noviembre 2011.

- Alonso, I., Romero, F.P., ***Técnicas de Expansión de Consultas aplicadas a repositorios Web de Historias Clínicas Electrónicas***. Congreso Español de Informática (CEDI 2010) - (III Simposio sobre Lógica Fuzzy y Soft Computing, LFSC-EUSFLAT), ISBN: 978-84-92812-65-3, pp. 401-408. Valencia, España. 7-10 de septiembre 2010.

- Alonso, I., Romero, F.P., ***Recuperación eficiente de información en repositorios de documentos PHR mediante el uso de ontologías del dominio***. 5ª Conferencia Ibérica de Sistemas y Tecnologías de Información - Doctoral Simposium, (CISTI 2010), Santiago de Compostela, España. 16-19 de junio 2010.

Bibliografía

Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 33-41). Association for Computational Linguistics.

Alonso, I., & Contreras, D. (2015). Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. Expert Systems with Applications. Elsevier (Artículo Aceptado) 19-Sep-2015.

Alonso, I., & Romero, F.P. (2010). Técnicas de Expansión de Consultas aplicadas a repositorios Web de Historias Clínicas Electrónicas. Congreso Español de Informática (CEDI 2010) - (III Simposio sobre Lógica Fuzzy y Soft Computing, LFSC-EUSFLAT), ISBN: 978-84-92812-65-3, pp. 401-408. Valencia, España. 7-10 de septiembre 2010.

Alonso, I., Contreras, D., Romero, F.P. (2012). Impacto de las relaciones jerárquicas inter/intra-recursos en la similitud semántica entre conceptos del dominio biomédico. XVI Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF 2012). ISBN: 978-84-615-6653-2, pp. 438-442. Valladolid, España.

Alonso, I., Contreras, D., Romero, F.P. (2013). Aplicación y medida de la similitud semántica presente en registros clínicos electrónicos para la mejora en estudios de cohortes, XVI Congreso Nacional de Informática de la Salud. (INFORMED 2013), ISBN: 978-84695-7444-7, pp. 270-274 Madrid, España.

Alonso, I., Romero, F.P., Contreras, D., Olivas, J.A. (2013) Una Aplicación de la Similitud Semántica en la mejora de la selección de Información Clínica. Congreso Español de Informática (CEDI 2013) - (Actas de la Multiconferencia CAEPIA 2013 - IV Simposio sobre Lógica Fuzzy y Soft Computing, LFSC), ISBN: 978-84-695-8348-7, pp. 1173-1181, Madrid, España.

Alpi, K. M. (2005). Expert searching in public health. Journal of the Medical Library Association, 93(1), 97.

Al-Mubaid, H., & Nguyen, H. (2006). A cluster-based approach for semantic similarity in the biomedical domain. In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE (pp. 2713-2717).

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the American Medical Informatics Association Symposium 2001 (pp. 17-21).

Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229-236.

Aronson, A. R., Rindflesch, T. C., & Browne, A. C. (1994, October). Exploiting a Large Thesaurus for Information Retrieval. In RIAO (Vol. 94, pp. 197-216).

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. In Proceedings of the American Medical Informatics Association Annual Fall Symposium (p. 485).

Attar, R., & Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. Journal of the ACM (JACM), 24(3), 397-417.

Babashzadeh, A., Huang, J., & Daoud, M. (2013, July). Exploiting semantics for improving clinical information retrieval. In Proceedings of the 36th international Association for Computing Machinery's Special Interest Group on Information Retrieval Conference on Research and development in information retrieval (pp. 801-804). ACM SIGIR 2013.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In Computational linguistics and intelligent text processing (pp. 136-145). Springer Berlin Heidelberg.

Banko, M., & Vanderwende, L. (2004). Using n-grams to understand the nature of summaries. In Proceedings of North American Chapter of the Association for Computational Linguistics, Human Language Technologies (HLT-NAACL Association for Computational Linguistics). Short Papers (pp. 1-4).

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics, 44(1), 118-125.

Bawakid, A., & Oussalah, M. (2008). A semantic summarization system: University of Birmingham at TAC 2008. In Proceedings of the first text analysis conference.

Benjamins, V. R., Contreras, J., Casanovas, P., Ayuso, M., Becue, M., Lemus, L., & Urios, C. (2004). Ontologies of professional legal knowledge as the basis for intelligent it support for judges. Artificial Intelligence and Law, 12(4), 359-378.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. Information processing & management, 43(4), 866-886.

Bodenreider, O. (2001). Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In Proceedings of the American Medical Informatics Association Symposium (p. 57).

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), D267-D270.

Bodenreider, O., & Burgun, A. (2005). Biomedical ontologies. In Medical Informatics (pp. 211-236). Springer US.

Bodenreider, O., & McCray, A. T. (2003). Exploring semantic groups through visual approaches. Journal of biomedical informat-ics, 36(6), 414-432.

Bossard, A., Généreux, M., & Poibeau, T. (2008). Description of the LIPN System at TAC 2008: Summarizing Information and Opinions. In Proceedings of the 1st Text Analysis Conference 2008 (pp. 282-291).

Burgun, A., & Bodenreider, O. (2001). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In Proceedings of the North American Chapter of the Association for Computational Linguistics 2001; Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations" (pp. 77-82).

Burgun, A., & Bodenreider, O. (2001). Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. Studies in health technology and informatics, 84(0 1), 171.

Carenini, G., Ng, R. T., & Zhou, X. (2008). Summarizing Emails with Conversational Cohesion and Subjectivity. In Proceedings of the Association for Computational Linguistics - Human Language Technologies (Vol. 8, pp. 353-361).

Carpineto, C., De Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS), 19(1), 1-27.

Caviedes, J. E., & Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the UMLS. Journal of biomedical informatics, 37(2), 77-85.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological review, 82(6), 407.

Crouch, C. J., & Yang, B. (1992, June). Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 77-88). ACM.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002, May). Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web (pp. 325-332). ACM.

Dalianis, H., Hassel, M., Henriksson, A., & Skeppstedt, M. (2012, October). Stockholm epr corpus: A clinical database used to improve health care. In Swedish Language Technology Conference (pp. 17-18).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. JAsIs, 41(6), 391-407.

Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in biology and medicine, 39(4), 396-403.

Efthimiadis, E. N. (1996). Query expansion. Annual review of information science and technology, 31, 121-187.

Efthimiadis, E. N. (2000). Interactive query expansion: a user□based evaluation in a relevance feedback environment. Journal of the American Society for Information Science, 51(11), 989-1003.

Erdogan, H., Erdem, E., & Bodenreider, O. (2010). Exploiting UMLS semantics for checking semantic consistency among UMLS concepts. Proceedings of MedInfo, 342-356.

Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, Massachusetts: MIT Press.

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of biomedical informatics, 35(4), 222-235.

Garla, V. N., & Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BioMedCentral Bioinformatics, 13(1), 261.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. arXiv preprint cmp-lg/9808002.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. Computer, IEEE Computer. 33(11), 29-36.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 23(1), 33-64.

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., & Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. Journal of Biomedical Semantics, 5(6).

Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994, January). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In SIGIR'94 (pp. 192-201). Springer London.

Hersh, W. R., & Hickam, D. H. (1998). How well do physicians use electronic information retrieval systems?: A framework for investigation and systematic review. Jama, 280(15), 1347-1352.

Hersh, W., Price, S., & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In Proceedings of the AMIA Symposium (p. 344). American Medical Informatics Association.

Hersh, W., & Voorhees, E. (2009). TREC genomics special issue overview. Information Retrieval, 12(1), 1-15.

Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., & Taylor, R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Affairs, 24(5), 1103-1117.

Hoffman, S. (2010). Electronic health records and research: Privacy versus scientific priorities. The American Journal of Bioethics, 10(9), 19-20.

Hoffman, Sharona, and Andy Podgurski. "Improving health care outcomes through personalized comparisons of treatment effectiveness based on electronic health records." The Journal of Law, Medicine & Ethics 39.3 (2011): 425-436.

Hotho, A., Staab, S., & Stumme, G. (2003, November). Ontologies improve text document clustering. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 541-544). IEEE.

Hovy, E., & Lin, C. Y. (1998). Automated text summarization and the SUMMARIST system. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (pp. 197-214). Association for Computational Linguistics.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., & Rhee, S. Y. (2008). Big data: The future of biocuration. Nature, 455(7209), 47-50.

Hsu, M. H., Tsai, M. F., & Chen, H. H. (2006). Query expansion with conceptnet and wordnet: An intrinsic comparison. In Information Retrieval Technology (pp. 1-13). Springer Berlin Heidelberg.

Jiang, JJ., Conrath, DW. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics;19–33.

Jing, H. (2002). Using hidden Markov modeling to decompose human-written summaries. Computational linguistics, 28(4), 527-543.

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006, May). Generating query substitutions. In Proceedings of the 15th international conference on World Wide Web (pp. 387-396). ACM.

Kalpathy-Cramer, J., de Herrera, A. G. S., Demner-Fushman, D., Antani, S., Bedrick, S., & Müller, H. (2015). Evaluating performance of biomedical image retrieval systems—An overview of the medical image retrieval task at ImageCLEF 2004–2013. Computerized Medical Imaging and Graphics, 39, 55-61.

Kashyap, V. (2003). The UMLS® Semantic Network and the Semantic Web. In AMIA Annual Symposium Proceedings (Vol. 2003, p. 351). American Medical Informatics Association.

Klingberg, T. (2008). The Overflowing Brain: Information Overload and the Limits of Working Memory: Information Overload and the Limits of Working Memory. Oxford University Press.

Kuhn, K. A., & Giuse, D. A. (2001). From hospital information systems to health information systems problems, challenges, perspectives. Methods Inf Med, 40(4), 275-287.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73). ACM.

Kurtz, C., Beaulieu, C. F., Napel, S., & Rubin, D. L. (2014). A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations. Journal of biomedical informatics, 49, 227-244.

Lin, D. (1998). An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine learning (Vol. 98, pp. 296-304).

Liu, H., Lussier, Y. A., & Friedman, C. (2001). Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. Journal of biomedical informatics, 34(4), 249-261.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, The MIT Press, Cambridge, MA, pages 265–283.

Lloret, E., Ferrández, O., Munoz, R., & Palomar, M. (2008, June). A Text Summarization Approach under the Influence of Textual Entailment. In Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science in Conjunction with the 10th International Conference on Enterprise Information Systems (pp. 22-31).

Lu, Z., Kim, W., & Wilbur, W. J. (2009). Evaluation of query expansion using MeSH in PubMed. Information retrieval, 12(1), 69-80.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

Mani, I. (2001). Summarization evaluation: An overview. In Proceedings of the NAACL 2001 Workshop on Automatic Summarization.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. Advances in automatic text summarization, The MIT Press, 123-136.

Marcu, D. (2000). The theory and practice of discourse parsing and summarization. MIT press.

Martinez, D., Otegi, A., Soroa, A., & Agirre, E. (2014). Improving search over Electronic Health Records using UMLS-based query expansion through random walks. Journal of biomedical informatics.

Mandala, R., Tokunaga, T., & Tanaka, H. (2000). Query expansion using heterogeneous thesauri. Information Processing & Management, 36(3), 361-378.

McCray, A. T., Aronson, A. R., Browne, A. C., Rindflesch, T. C., Razi, A., & Srinivasan, S. (1993). UMLS knowledge for biomedical language processing. Bulletin of the Medical Library Association, 81(2), 184.

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In Proceedings of the Annual Symposium on Computer Application in Medical Care (p. 235). American Medical Informatics Association.

McInnes, B. T., Pedersen, T., & Pakhomov, S. V. (2009). UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In American Medical Informatics Association Annual Symposium Proceedings (Vol. 2009, p. 431).

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics, 46(6), 1116-1124.

Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., & Hripcsak, G. (2006). Inter-patient distance metrics using SNOMED CT defining relationships. Journal of biomedical informatics, 39(6), 697-705.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

Metzler, D., & Kanungo, T. (2008). Machine learned sentence selection strategies for query-biased summarization. In Proceedings of the Special Interest Group on Information Retrieval Conference, Learning to Rank for Information Retrieval Workshop (pp. 40-47).

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35, 128-44.

Middleton, B., Bloomrosen, M., Dente, M. A., Hashmat, B., Koppel, R., Overhage, J. M., & Zhang, J. (2013). Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. Journal of the American Medical Informatics Association, 20(e1), e2-e8.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and cognitive processes, 6(1), 1-28.

Mitra, M., Singhal, A., & Buckley, C. (1998, August). Improving automatic query expansion. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 206-214). ACM.

Moldovan, D. I., & Mihalcea, R. (2000). Using wordnet and lexical operators to improve internet searches. IEEE Internet Computing, (1), 34-43.

Moldovan, D. I., & Mihalcea, R. (1999). Improving the search on the Internet by using WordNet and lexical operators. IEEE Internet Computing, 14(1).

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational linguistics, 17(1), 21-48.

Natarajan, K., Stein, D., Jain, S., & Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. International journal of medical informatics, 79(7), 515-522.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. In Advances in Artificial Intelligence (pp. 205-215). Springer Berlin Heidelberg.

Nguyen, H. A., & Al-Mubaid, H. (2006). New ontology-based semantic similarity measure for the biomedical domain. In Granular Computing, 2006 IEEE International Conference on (pp. 623-628).

Paice, C. D. (1980). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (pp. 172-191).

Paice, C. D. (1990). Constructing literature abstracts by computer: techniques and prospects. Information Processing & Management, 26(1), 171-186.

Patwardhan, S., & Pedersen, T. (2006, April). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together (Vol. 1501, pp. 1-8).

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics, 40(3), 288-299.

Plaza, L., & Díaz, A. (2010). Retrieval of similar electronic health records using UMLS concept graphs. In Natural Language Processing and Information Systems (pp. 296-303). Springer Berlin Heidelberg.

Plaza, L., Díaz, A., & Gervás, P. (2011). A semantic graph-based approach to biomedical summarisation. In Artificial intelligence in medicine, 53(1), 1-14.

Prokosch, H. U., & Ganslandt, T. (2009). Perspectives for medical informatics: Reusing the electronic medical record for clinical re-search. Methods of Information in Medicine, 48, 38-44.

Qi, Y., & Laquerre, P. F. (2012). Retrieving Medical Records with sennamed: NEC Labs America at TREC 2012 Medical Records Track. NEC LABORATORIES AMERICA INC PRINCETON NJ.

Qiu, Y., & Frei, H. P. (1993, July). Concept based query expansion. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 160-169). ACM.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1), 17-30.

Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938.

Reeve, L. H., Han, H., & Brooks, A. D. (2007). The use of domain-specific concepts in biomedical text summarization. Information Processing & Management, 43(6), 1765-1776.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. 448 – 453.

Rindflesch, T. C., & Aronson, A. R. (1994). Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In Proceedings of the Annual Symposium on Computer Application in Medical Care (p. 240). American Medical Informatics Association.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y & Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. In AMIA Annual Symposium Proceedings (Vol. 2007, p. 625). American Medical Informatics Association.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information science, 27(3), 129-146.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. The SMART retrieval system: experiments in automatic document processing. 1971:313–323.

Roque, F. S., Jensen, P. B., et al. (2011). Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology, 7(8), e1002141.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. Proceedings of the 13th International Conference on World Wide Web, 13-19.

Roth, L., & Hole, W. T. (2000). Managing name ambiguity in the UMLS metathesaurus. In Proceedings of the AMIA Symposium (p. 1124). American Medical Informatics Association.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the Association for Computing Machinery's, 8(10), 627-633.

Rush, J. E., Salvador, R., & Zamora, A. (1971). Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, *22*(4), 260-274.

Saggion, H., Lloret, E., & Palomar, M. (2010). Using text summaries for predicting rating scales. In Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), Lisbon, Portugal (pp. 44-51).

Salton, G. (1971). The SMART retrieval system—experiments in automatic document processing.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. Readings in information retrieval, 24(5), 355-363.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of biomedical informatics, 44(5), 749-759.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. Knowledge-Based Systems, 24(2), 297-303.

Sánchez, D., Batet, M., & Valls, A. (2010). Web-based semantic similarity: an evaluation in the biomedical domain. International journal of software and informatics, 4(1), 39-52.

Schütze, H. (1998). Automatic word sense discrimination. Computational linguistics, 24(1), 97-123.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.

Shin, K., Han, S. Y., Gelbukh, A., & Park, J. (2004). Advanced relevance feedback query expansion strategy for information retrieval in medline. In Progress in Pattern Recognition, Image Analysis and Applications (pp. 425-431). Springer Berlin Heidelberg.

Sittig, D. F., Wright, A., Simonaitis, L., Carpenter, J. D., Allen, G. O., Doebbeling, B. N., & Middleton, B. (2010). The state of the art in clinical knowledge management: an inventory of tools and techniques. International journal of medical informatics, 79(1), 44-57.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In European Conference on Artificial Intelligence. ECAI'04 (Vol. 16, p. 1089).

Selberg, E. (1997). Information Retrieval Advances using Relevance Feedback. UW Dept. of CSE General Exam.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

Sparck Jones, K. (1999). Automatic summarizing: factors and directions. Advances in automatic text summarization, MIT Press, Cambridge, pages 1–12.

Srinivasan, P. (1996). Optimal document-indexing vocabulary for MEDLINE. Information Processing & Management, 32(5), 503-514.

Stapley, B. J., & Benoit, G. (2000, January). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In Pac Symp Biocomput (Vol. 5, pp. 529-540).

Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. In Proceedings of the Association for Computational Linguistics. (Vol. 97, No. 1997, pp. 58-65).

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 2-10). ACM.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141-188.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In Special Interest Group on Information Retrieval, SIGIR'94 (pp. 61-69). Springer London.

Voorhees, E., & Tong, R. (2011, November). Overview of the TREC 2011 medical records track. In The Twentieth Text REtrieval Conference Proceedings TREC.

Wan, S., & McKeown, K. (2004). Generating overview summaries of ongoing email thread discussions. In Proceedings of the 20th international conference on Computational Linguistics (p. 549). Association for Computational Linguistics.

Wang, L., Ranjan, R., Kołodziej, J., Zomaya, A., & Alem, L. (2015). Software Tools and Techniques for Big Data Computing in Healthcare Clouds. Future Generation Computer Systems, 43, 38-39.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

Xu, J., & Croft, W. B. (1996, August). Query expansion using local and global document analysis. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 4-11). ACM.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems (TOIS), 18(1), 79-112.

Xu, X., Zhu, W., Zhang, X., Hu, X., & Song, I. Y. (2006). A comparison of local analysis, global analysis and ontology-based query expansion strategies for bio-medical literature search. In Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on (Vol. 4, pp. 3441-3446). IEEE

Zeng, Q. T., Redd, D., Rindflesch, T., & Nebeker, J. (2012). Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In American Medical Informatics Association Annual Symposium Proceedings (Vol. 2012, p. 1050).

Zhou, Z., Wang, Y., & Gu, J. (2008, December). A new model of information content for semantic similarity in WordNet. In Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on (Vol. 3, pp. 85-89). IEEE.

Zhu, D., Wu, S., Carterette, B., & Liu, H. (2014). Using large clinical corpora for query expansion in text-based cohort identification. Journal of biomedical informatics.

Zhu, W., Xu, X., Hu, X., Song, I. Y., & Allen, R. B. (2006, May). Using UMLS-based Re-Weighting Terms as a Query Expansion Strategy. In GrC (pp. 217-222).

Apéndice A

TOPIC 101: Patients with hearing loss

1001 1 0 C0030705 **podg** Patients with hearing loss - **Patients**
1001 1 0 C0011053 dsyn Patients with hearing loss - Deafness
1002 1 0 C0030705 **podg** Patients with hearing loss - **Patients**
1002 1 0 C0018772 fndg Patients with hearing loss - Hearing Loss, Partial
1003 1 0 C0030705 **podg** Patients with hearing loss - **Patients**
1003 1 0 C1384666 fndg Patients with hearing loss - hearing impairment

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
dsyn – Disease or Syndrome (DISO) fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB)

TOPIC 102: Patients with complicated GERD who receive endoscopy

1001 1 0 C0030705 **podg** Patients with complicated GERD - **Patients**
1001 1 0 C0231242 **ftcn** Patients with complicated GERD - **Complicated**
1001 1 0 C0017168 dsyn Patients with complicated GERD - Gastroesophageal reflux disease
1002 2 0 C1514756 **qlco** receive - **Receive**
1003 3 0 C0014245 diap endoscopy - Endoscopy (procedure)
1004 3 0 C1552424 **hcro** endoscopy - **Clinic / Center - Endoscopy**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
dsyn – Disease or Syndrome (DISO) diap – Diagnostic Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC) **qlco** – Qualitative Concept (CONC) **hcro** – Health Care Related Organization (ORGA)

TOPIC 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis.

1001 1 0 C0870668 **podg** Hospitalized patients - **Hospitalized patients**
1002 2 0 C0332293 topp treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Treated with
1002 2 0 C0343401 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - MRSA - Methicillin resistant Staphylococcus aureus infection
1002 2 0 C0014118 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Endocarditis
1003 2 0 C0332293 topp treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis- Treated with
1003 2 0 C1265292 bact treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Methicillin resistant Staphylococcus aureus (organism)
1003 2 0 C0014118 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Endocarditis
1004 2 0 C1522326 **ftcn** treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - **Treating**
1004 2 0 C0343401 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - MRSA - Methicillin resistant Staphylococcus aureus infection
1004 2 0 C0014118 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Endocarditis
1005 2 0 C1522326 **ftcn** treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - **Treating**
1005 2 0 C1265292 bact treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Methicillin resistant Staphylococcus aureus (organism)
1005 2 0 C0014118 dsyn treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis - Endocarditis

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
topp – Thrapeutic or Preventive Procedure (PROC) dsyn – Disease or Syndrome (DISO) bact – Bacterium (LIVB)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC)

TOPIC 104: Patients diagnosed with localized prostate cancer and treated with robotic surgery.

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1002 2 0 C0796563 neop diagnosed with localized prostate cancer - Localized Malignant Neoplasm
1002 2 0 C0033572 bpoc diagnosed with localized prostate cancer - Prostate
1003 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1003 2 0 C0796563 neop diagnosed with localized prostate cancer - Localized Malignant Neoplasm
1003 2 0 C1278980 bpoc diagnosed with localized prostate cancer - Entire prostate
1004 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1004 2 0 C1334407 neop diagnosed with localized prostate cancer - Localized Carcinoma
1004 2 0 C0033572 bpoc diagnosed with localized prostate cancer - Prostate
1005 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1005 2 0 C1334407 neop diagnosed with localized prostate cancer - Localized Carcinoma
1005 2 0 C1278980 bpoc diagnosed with localized prostate cancer - Entire prostate
1006 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1006 2 0 C0392752 **spco** diagnosed with localized prostate cáncer - **Localized**
1006 2 0 C0376358 neop diagnosed with localized prostate cancer - Malignant neoplasm of prostate
1007 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1007 2 0 C0392752 **spco** diagnosed with localized prostate cancer - **Localized**
1007 2 0 C0600139 neop diagnosed with localized prostate cancer - Prostate carcinoma
1008 2 0 C0011900 fndg diagnosed with localized prostate cancer - Diagnosis
1008 2 0 C0392752 **spco** diagnosed with localized prostate cancer - **Localized**
1008 2 0 C2984325 **ftcn** diagnosed with localized prostate cáncer - **Prostate Cancer Pathway**
1009 3 0 C0332293 topp treated with robotic surgery - Treated with
1009 3 0 C0035785 ocdi treated with robotic surgery - Robotics
1009 3 0 C0038894 bmod treated with robotic surgery - Surgery specialty
1010 3 0 C0332293 topp treated with robotic surgery - Treated with
1010 3 0 C0035785 ocdi treated with robotic surgery - Robotics
1010 3 0 C0038895 **ftcn** treated with robotic surgery - **Surgical aspects**
1011 3 0 C0332293 topptreated with robotic surgery - Treated with
1011 3 0 C0035785 ocdi treated with robotic surgery - Robotics
1011 3 0 C0543467 diap treated with robotic surgery - Operative Surgical Procedures

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
fndg – Finding (DISO) neop – Neoplastic Process (DISO) bpoc – Body Part, Organ, or Organ Component (ANAT) topp – Thrapeutic or Preventive Procedure (PROC) ocdi – Occupation or Discipline (OCCU) bmod – Biomedical Occupation or Discipline (OCCU) diap – Diagnostic Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC) **spco** – Spatial Concept (CONC)

TOPIC 105: Patients with dementia

1001 1 0 C0030705 **podg** Patients with dementia - **Patients**
1001 1 0 C0011265 mobd Patients with dementia - Presenile dementia
1002 1 0 C0030705 **podg** Patients with dementia - **Patients**
1002 1 0 C0497327 mobd Patients with dementia - Dementia

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
modb - Mental or Behavioral Dysfunction (DISO)	**podg** – Patient or Disabled Group (LIVB)

TOPIC 106: Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer

1001 1 0 C0030705 **podg**	Patients	**-Patients**
1002 2 0 C0032743 diap	positron emission tomography (PET)	-Positron-Emission Tomography
1003 3 0 C0024485 diap	magnetic resonance imaging (MRI),	-Magnetic Resonance Imaging
1004 3 0 C1552358 **prog**	magnetic resonance imaging (MRI),	**-Radiologic Technologist-Magnetic Resonance Imaging**
1005 4 0 C0040405 diap	computed tomography (CT) for staging	-X-Ray Computed Tomography
1005 4 0 C0332305 **ftcn**	computed tomography (CT) for staging	**-With staging**
1006 4 0 C1552357 **prog**	computed tomography (CT)for staging	**-Radiologic Technologist-Computed Tomography**
1006 4 0 C0332305 **ftcn**	computed tomography (CT) for staging	**-With staging**
1007 5 0 C0150369 hlca	monitoring of cancer	-Preventive monitoring
1007 5 0 C0006826 neop	monitoring of cancer	-Malignant Neoplasms
1008 5 0 C0150369 hlca	monitoring of cancer	-Preventive monitoring
1008 5 0 C0998265 euka	monitoring of cancer	-Cancer Genus
1009 5 0 C0150369 hlca	monitoring of cancer	-reventive monitoring
1009 5 0 C1306459 fndg	monitoring of cancer	-Primary malignant neoplasm
1010 5 0 C1516647 **resa**	monitoring of cancer	**-Clinical Trials, Monitoring**
1010 5 0 C0006826 neop	monitoring of cancer	-Malignant Neoplasms
1011 5 0 C1516647 **resa**	monitoring of cancer	**-Clinical Trials, Monitoring**
1011 5 0 C0998265 euka	monitoring of cancer	-Cancer Genus
1012 5 0 C1516647 **resa**	monitoring of cancer	**-Clinical Trials, Monitoring**
1012 5 0 C1306459 fndg	monitoring of cancer	-Primary malignant neoplasm

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
diap – Diagnostic Procedure (PROC) hlca - Health Care Activity (PROC) neop – Neoplastic Process (DISO) euka - Eukaryote (LIVB) fndg - Finding (DISO)	**podg** – Patient or Disabled Group (LIVB) **prog** – Professional or Occupational Group (LIVB9 **ftcn** – Functional Concept (CONC) **resa** – Research Activity PROC

TOPIC 107: Patients with ductal carcinoma in situ (DCIS)

1001 1 0 C0030705 **podg** Patients with ductal carcinoma in situ - **Patients**
1001 1 0 C0007124 neop Patients with ductal carcinoma in situ - Noninfiltrating Intraductal Carcinoma

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
neop – Neoplastic Process (DISO)	**podg** – Patient or Disabled Group (LIVB)

TOPIC 108: Patients treated for vascular claudication surgically

1001 1 0 C0030705 **podg** Patients **Patients**
1002 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1002 2 0 C0005847 bpoc treated for vascular claudication surgically - Blood Vessel
1002 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1002 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1003 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1003 2 0 C0005847 bpoc treated for vascular claudication surgically - Blood Vessel
1003 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1003 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1004 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1004 2 0 C1558950 fndg treated for vascular claudication surgically - Adverse Event Associated with Vascular
1004 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1004 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1005 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1005 2 0 C1558950 fndg treated for vascular claudication surgically - Adverse Event Associated with Vascular
1005 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1005 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1006 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1006 2 0 C1801960 **qlco** treated for vascular claudication surgically - **Vascular**
1006 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1006 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1007 2 0 C0332293 topp treated for vascular claudication surgically - Treated with
1007 2 0 C1801960 **qlco** treated for vascular claudication surgically - **Vascular**
1007 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1007 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1008 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1008 2 0 C0005847 bpoc treated for vascular claudication surgically - Blood Vessel
1008 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1008 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1009 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1009 2 0 C0005847 bpoc treated for vascular claudication surgically - Blood Vessel
1009 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1009 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1010 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1010 2 0 C1558950 fndg treated for vascular claudication surgically - Adverse Event Associated with Vascular
1010 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1010 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1011 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1011 2 0 C1558950 fndg treated for vascular claudication surgically - Adverse Event Associated with Vascular
1011 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1011 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1012 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1012 2 0 C1801960 **qlco** treated for vascular claudication surgically - **Vascular**
1012 2 0 C0311395 fndg treated for vascular claudication surgically - Lameness
1012 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures
1013 2 0 C1522326 **ftcn** treated for vascular claudication surgically - **Treating**
1013 2 0 C1801960 **qlco** treated for vascular claudication surgically - **Vascular**
1013 2 0 C1456822 dsyn treated for vascular claudication surgically - Claudication (finding)
1013 2 0 C0543467 diap treated for vascular claudication surgically - Operative Surgical Procedures

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
topp – Thrapeutic or Preventive Procedure (PROC) bpoc – Body Part, Organ, or Organ Component (ANAT) fndg – Finding (DISO) diap – Diagnostic Procedure (PROC) dsyn – Disease or Syndrome (DISO)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC) **ftcn** – Functional Concept (CONC)

TOPIC 109: Women with osteopenia

1001 1 0 C0043210 popg Women with osteopenia - Woman
1001 1 0 C0029453 patf Women with osteopenia - Osteopenia

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
popg – Population Group (LIVB) patf – Pathologic Function (DISO)	

TOPIC 110: Patients being discharged from the hospital on hemodialysis

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0438953 fndg discharged from the hospital - Discharged from hospital
1003 3 0 C0019004 topp on hemodialysis - Hemodialysis
1004 3 0 C1524112 **ftcn** on hemodialysis - **Drug Administration via Hemodialysis**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
fndg – Finding (DISO) topp – Thrapeutic or Preventive Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC)

TOPIC 111: Patients with chronic back pain who receive an intraspinal pain-medicine pump

1001 1 0 C0030705 **podg** Patients with chronic back pain - **Patients**
1001 1 0 C0740418 sosy Patients with chronic back pain - Chronic back pain
1002 2 0 C1514756 **qlco** receive - **Receive**
1003 3 0 C1283188 **spco** an intraspinal pain-medicine pump - **Intraspinal**
1003 3 0 C3263701 bmod an intraspinal pain-medicine pump - (No hay resultado) … pain Medicine
1003 3 0 C0182537 medd an intraspinal pain-medicine pump - pump (device)
1004 3 0 C1283188 **spco** an intraspinal pain-medicine pump - **Intraspinal**
1004 3 0 C3263701 bmod an intraspinal pain-medicine pump - (No hay resultado) … pain Medicine
1004 3 0 C1706421 mnob an intraspinal pain-medicine pump - Pump Device Component

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
sosy – Sign or Symptom (DISO) bmod – Biomedical Occupation or Discipline (OCCU) medd – Medical Device (DEVI) mnob – Manufactured Object (OBJC)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC) **spco** – Spatial Concept (CONC)

TOPIC 112: Female patients with breast cancer with mastectomies during admission

1001 1 0 C0007104 neop Female patients with breast cancer - Female Breast Carcinoma
1001 1 0 C0030705 **podg** Female patients with breast cancer - **Patients**
1002 1 0 C0235653 neop Female patients with breast cancer - Malignant neoplasm of female breast
1002 1 0 C0030705 **podg** Female patients with breast cancer - **Patients**
1003 2 0 C0024881 topp with mastectomies - Mastectomy
1004 3 0 C0184666 hlca during admission - Hospital admission
1005 3 0 C0809949 hlca during admission - Admission activity

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
neop – Neoplastic Process (DISO) topp – Thrapeutic or Preventive Procedure (PROC) hlca - Health Care Activity (PROC)	**podg** – Patient or Disabled Group (LIVB)

TOPIC 113: Adult patients who received colonoscopies during admission which revealed adenocarcinoma

1001 1 0 C0001675 aggp Adult patients – Adult
1001 1 0 C0030705 **podg** Adult patients –**Patients**
1002 1 0 C1706450 **humn** Adult patients –**Legal Adult**
1002 1 0 C0030705 **podg** Adult patients – **Patients**
1003 2 0 C1514756 **qlco** received – **Receive**
1004 3 0 C0009378 diap colonoscopies during admission – colonoscopy
1004 3 0 C0585045 **tmco** colonoscopies during admission –**During admission**
1005 4 0 C0443289 **qlco** revealed – **Revealed**
1006 5 0 C0001418 neop adenocarcinoma –Adenocarcinoma

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) diap – Diagnostic Procedure (PROC) neop – Neoplastic Process (DISO)	**podg** – Patient or Disabled Group (LIVB) **humn** – Human (LIVB) **qlco** – Qualitative Concept (CONC) **tmco** – Temporal Concept (CONC)

TOPIC 114: Adult patients discharged home with palliative care / home hospice

1001 1 0 C0001675 aggp	Adult patients	- Adult
1001 1 0 C0030705 **podg**	Adult patients	- **Patients**
1002 1 0 C1706450 **humn**	Adult patients	- **Legal Adult**
1002 1 0 C0030705 **podg**	Adult patients	- **Patients**
1003 2 0 C0012621 bdsu	discharged	- Body Fluid Discharge
1004 2 0 C0030685 hlca	discharged	- Patient Discharge
1005 2 0 C2926602 bdsu	discharged	- Discharge, body substance
1006 3 0 C0204977 hlca	home with palliative care / home hospice	- Home care of patient
1006 3 0 C1285530 **qlco**	home with palliative care / home hospice	- **Palliative**
1006 3 0 C0019947 hcro	home with palliative care / home hospice	- hospice environment
1007 3 0 C0204977 hlca	home with palliative care / home hospice	- Home care of patient
1007 3 0 C1285530 **qlco**	home with palliative care / home hospice	- **Palliative**
1007 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1008 3 0 C0204977 hlca	home with palliative care / home hospice	- Home care of patient
1008 3 0 C1555456 **idcn**	home with palliative care / home hospice	- **Act Code - Palliative**
1008 3 0 C0019947 hcro	home with palliative care / home hospice	- hospice environment
1009 3 0 C0204977 hlca	home with palliative care / home hospice	- Home care of patient
1009 3 0 C1555456 **idcn**	home with palliative care / home hospice	- **Act Code - Palliative**
1009 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1010 3 0 C0994454 hlca	home with palliative care / home hospice	- Home care aspects
1010 3 0 C1285530 **qlco**	home with palliative care / home hospice	- **Palliative**
1010 3 0 C0019947 hcro	home with palliative care / home hospice	- hospice environment
1011 3 0 C0994454 hlca	home with palliative care / home hospice	- Home care aspects
1011 3 0 C1285530 **qlco**	home with palliative care / home hospice	- **Palliative**
1011 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1012 3 0 C0994454 hlca	home with palliative care / home hospice	- Home care aspects
1012 3 0 C1555456 **idcn**	home with palliative care / home hospice	- **Act Code - Palliative**
1012 3 0 C0019947 hcro	home with palliative care / home hospice	- hospice environment
1013 3 0 C0994454 hlca	home with palliative care / home hospice	- Home care aspects
1013 3 0 C1555456 **idcn**	home with palliative care / home hospice	- **Act Code - Palliative**
1013 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1014 3 0 C0442519 **cnce**	home with palliative care / home hospice	- **Home environment**
1014 3 0 C0030231 hlca	home with palliative care / home hospice	- Palliative Care
1014 3 0 C0019947 hcro	home with palliative care / home hospice	- hospice environment
1015 3 0 C0442519 **cnce**	with palliative care / home hospice	- **Home environment**
1015 3 0 C0030231 hlca	home with palliative care / home hospice	- Palliative Care
1015 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1016 3 0 C0442519 **cnce**	home with palliative care / home hospice	- **Home environment**
1016 3 0 C1552336 **inpr**	home with palliative care / home hospice	-**Palliative/Hospice-NUCCProviderCodes**
1016 3 0 C1947933 **acty**	home with palliative care / home hospice	- **care activity**
1017 3 0 C0442519 **cnce**	home with palliative care / home hospice	- **Home environment**
1017 3 0 C1285530 **qlco**	home with palliative care / home hospice	- **Palliative**
1017 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care
1018 3 0 C0442519 **cnce**	home with palliative care / home hospice	- **Home environment**
1018 3 0 C1555456 **idcn**	home with palliative care / home hospice	- **Act Code - Palliative**
1018 3 0 C0085555 hlca	home with palliative care / home hospice	- Hospice Care

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) bdsu – Body Substance (ANAT) hlca - Health Care Activity (PROC) hcro - Health Care Related Organization (ORGA)	**podg** – Patient or Disabled Group (LIVB) **humn** – Human (LIVB) **qlco** – Qualitative Concept (CONC) **idcn** – Idea or Concept (CONC) **cnce** – Conceptual Entity (CONC) **inpr** – Intellectual Product (CONC) **acty** – Activity (ACTI)

TOPIC 115: Adult patients who are admitted with an asthma exacerbation

1001 1 0 C0001675 aggp	Adult patients	- Adult
1001 1 0 C0030705 **podg**	Adult patients	- **Patients**
1002 1 0 C1706450 **humn**	Adult patients	- **Legal Adult**
1002 1 0 C0030705 **podg**	Adult patients	- **Patients**
1003 2 0 C0184666 hlca	admitted with an asthma exacerbation	- Hospital admission
1003 2 0 C0349790 fndg	admitted with an asthma exacerbation	- Exacerbation of asthma

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) hlca - Health Care Activity (PROC) fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB) **humn** – Human (LIVB)

TOPIC 116: Patients received methotrexate for cancer treatment while in the hospital

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C1514756 **qlco**	received	- **Receive**
1003 3 0 C0746573 topp	methotrexate for cancer treatment	- METHOTREXATE TREATMENT
1003 3 0 C0006826 neop	methotrexate for cancer treatment	- Malignant Neoplasms
1004 3 0 C0746573 topp	methotrexate for cancer treatment	- METHOTREXATE TREATMENT
1004 3 0 C0998265 euka	methotrexate for cancer treatment	- Cancer Genus
1005 3 0 C0746573 topp	methotrexate for cancer treatment	- METHOTREXATE TREATMENT
1005 3 0 C1306459 fndg	methotrexate for cancer treatment	- Primary malignant neoplasm
1006 3 0 C0025677 orch	methotrexate for cancer treatment	- Methotrexate
1006 3 0 C0920425 topp	methotrexate for cancer treatment	- cancer treatment
1007 4 0 C0019994 hcro	while in the hospital	- Hospitals
1008 4 0 C1510665 **qlco**	while in the hospital	- **Hospital environment**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
topp – Thrapeutic or Preventive Procedure (PROC) neop – Neoplastic Process (DISO)euka fndg – Finding (DISO) orch – Organic Chemical CHEM hcro - Health Care Related Organization (ORGA)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC)

TOPIC 117: Patients with Post-traumatic Stress Disorder

1001 1 0 C0030705 **podg**	Patients with Post-traumatic Stress Disorder	- **Patients**
1001 1 0 C0038436 mobd	Patients with Post-traumatic Stress Disorder	- Post-Traumatic Stress Disorder

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
mobd – Mental or Behavioral Dysfunction DISO	**podg** – Patient or Disabled Group (LIVB)

TOPIC 118: Adults who received a coronary stent during an admission

1001 1 0 C0001675 aggp Adults - Adult
1002 2 0 C1514756 **qlco** received - **Receive**
1003 3 0 C0687568 medd a coronary stent during an admission - Stents, Vascular, Coronary
1003 3 0 C0184666 hlca a coronary stent during an admission - Hospital admission
1004 3 0 C0687568 medd a coronary stent during an admission - Stents, Vascular, Coronary
1004 3 0 C0809949 hlca a coronary stent during an admission - Admission activity

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) medd – Medical Device (DEVI) hlca - Health Care Activity (PROC)	**qlco** – Qualitative Concept (CONC)

TOPIC 119: Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes

1001 1 0 C0001675 aggp Adult patients - Adult
1001 1 0 C0030705 **podg** Adult patients - **Patients**
1002 1 0 C1706450 **humn** Adult patients - **Legal Adult**
1002 1 0 C0030705 **podg** Adult patients - **Patients**
1003 2 0 C0449450 **idcn** presented to the emergency room - **Presentation**
1003 2 0 C0562508 hcro presented to the emergency room - Accident and Emergency department
1004 3 0 C0860062 dsyn with anion gap acidosis - Anion gap acidosis
1005 4 0 C0175668 **tmco** secondary to insulin dependent diabetes - **Secondary to**
1005 4 0 C0011854 dsyn secondary to insulin dependent diabetes - Diabetes Mellitus, Insulin-Dependent

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) hcro - Health Care Related Organization (ORGA) dsyn – Disease or Syndrome (DISO)	**podg** – Patient or Disabled Group (LIVB) **humn** – Human (LIVB) **idcn** – Idea or Concept (CONC) **tmco** – Temporal Concept (CONC)

TOPIC 120: Patients admitted for treatment of CHF exacerbation

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0184666 hlca admitted for treatment of CHF exacerbation - Hospital admission
1002 2 0 C0039798 **ftcn** admitted for treatment of CHF exacerbation - **therapeutic aspects**
1002 2 0 C0018802 dsyn admitted for treatment of CHF exacerbation - Congestive heart failure
1003 2 0 C0184666 hlca admitted for treatment of CHF exacerbation - Hospital admission
1003 2 0 C0087111 topp admitted for treatment of CHF exacerbation - Therapeutic procedure
1003 2 0 C0018802 dsyn admitted for treatment of CHF exacerbation - Congestive heart failure
1004 2 0 C0184666 hlca admitted for treatment of CHF exacerbation - Hospital admission
1004 2 0 C1522326 **ftcn** admitted for treatment of CHF exacerbation - **Treating**
1004 2 0 C0018802 dsyn admitted for treatment of CHF exacerbation - Congestive heart failure
1005 2 0 C0184666 hlca admitted for treatment of CHF exacerbation - Hospital admission
1005 2 0 C1533734 topp admitted for treatment of CHF exacerbation - Administration procedure
1005 2 0 C0018802 dsyn admitted for treatment of CHF exacerbation - Congestive heart failure
1006 2 0 C0184666 hlca admitted for treatment of CHF exacerbation - Hospital admission
1006 2 0 C1705169 **cnce** admitted for treatment of CHF exacerbation - **Biomaterial Treatment**
1006 2 0 C0018802 dsyn admitted for treatment of CHF exacerbation - Congestive heart failure

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca - Health Care Activity (PROC) dsyn – Disease or Syndrome (DISO) topp – Thrapeutic or Preventive Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC) **cnce** – Conceptual Entity (CONC)

TOPIC 121: Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix

1001 1 0 C0030705 **podg**	Patients with CAD	- **Patients**
1002 2 0 C0449450 **idcn**	presented to the Emergency Department	- **Presentation**
1002 2 0 C0562508 hcro	presented to the Emergency Department	- Accident and Emergency department
1003 3 0 C0948089 dsyn	with Acute Coronary Syndrome	- Acute Coronary Syndrome
1004 4 0 C1442162 **cnce**	given	- **GIVEN**
1005 4 0 C1550718 **idcn**	given	- **Entity Name Part Type - given**
1006 4 0 C3244317 **inpr**	given	- **Given name**
1007 5 0 C0633084 orch	Plavix	- Plavix

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hcro - Health Care Related Organization (ORGA) dsyn – Disease or Syndrome (DISO) orch – Organic Chemical CHEM	**podg** – Patient or Disabled Group (LIVB) **idcn** – Idea or Concept (CONC) **cnce** – Conceptual Entity (CONC) **inpr** – Intellectual Product (CONC)

TOPIC 122: Patients who received total parenteral nutrition while in the hospital

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C1514756 **qlco**	received	- **Receive**
1003 3 0 C0030548 topp	total parenteral nutrition	- Parenteral Nutrition, Total
1004 4 0 C0019994 hcro	while in the hospital	- Hospitals
1005 4 0 C1510665 **qlco**	while in the hospital	- **Hospital environment**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
topp – Thrapeutic or Preventive Procedure (PROC) hcro - Health Care Related Organization (ORGA)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC)

TOPIC 123: Diabetic patients who received diabetic education in the hospital

1001 1 0 C0241863 fndg	Diabetic patients	- Diabetic
1001 1 0 C0030705 **podg**	Diabetic patients	- **Patients**
1002 2 0 C1514756 **qlco**	received	- **Receive**
1003 3 0 C0204935 hlca	diabetic education in the hospital	- Diabetic education
1003 3 0 C0019994 hcro	diabetic education in the hospital	- Hospitals
1004 3 0 C0204935 hlca	diabetic education in the hospital	- Diabetic education
1004 3 0 C1510665 **qlco**	diabetic education in the hospital	- **Hospital environment**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
fndg – Finding (DISO) hlca - Health Care Activity (PROC) hcro - Health Care Related Organization (ORGA)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC)

TOPIC 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma

1001 1 0 C0030705	**podg**	Patients	- **Patients**
1002 2 0 C0150312	**qnco**	present to the hospital	- **Present**
1002 2 0 C0019994	hcro	present to the hospital	- Hospitals
1003 2 0 C0150312	**qnco**	present to the hospital	- **Present**
1003 2 0 C1510665	**qlco**	present to the hospital	- **Hospital environment**
1004 2 0 C0449450	**idcn**	present to the hospital	- **Presentation**
1004 2 0 C0019994	hcro	present to the hospital	- Hospitals
1005 2 0 C0449450	**idcn**	present to the hospital	- **Presentation**
1005 2 0 C1510665	**qlco**	present to the hospital	- **Hospital environment**
1006 3 0 C0332189	**tmco**	with episodes	- Episode of
1007 4 0 C0205178	**tmco**	of acute loss	- **acute**
1007 4 0 C1517945	**qnco**	of acute loss	- **Loss**
1008 5 0 C0042789	orgf	of vision	- Vision
1009 5 0 C2707266	**clna**	of vision	- **Vision:-:Point in time:^Patient**:
1010 6 0 C0175668	**tmco**	secondary to glaucoma	- Secondary to
1010 6 0 C0017601	dsyn	secondary to glaucoma	- Glaucoma
1011 6 0 C0175668	**tmco**	secondary to glaucoma	- **Secondary to**
1011 6 0 C0997768	euka	secondary to glaucoma	- Glaucoma (eukaryote)
1012 6 0 C0175668	**tmco**	secondary to glaucoma	- **Secondary to**
1012 6 0 C1962986	fndg	secondary to glaucoma	- Glaucoma Adverse Event

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hcro – Health Care Related Organization (ORGA) orgf – Organism Function PHYS dsyn – Disease or Syndrome (DISO) euka– Eukaryote (LIVB) fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB) **qnco** – Quantitative Concept CONC **qlco** – Qualitative Concept (CONC) **idcn** – Idea or Concept (CONC) **tmco** – Temporal Concept (CONC) **clna** – Clinical Attribute (PHYS)

TOPIC 125: Patients co-infected with Hepatitis C and HIV

1001 1 0 C0030705	**podg**	Patients	- **Patients**
1002 2 0 C0439663	**ftcn**	co-infected with Hepatitis C	-**Infected**
1002 2 0 C0019196	dsyn	co-infected with Hepatitis C	- Hepatitis C
1003 2 0 C0439663	**ftcn**	co-infected with Hepatitis C	- **Infected**
1003 2 0 C0220847	virs	co-infected with Hepatitis C	- Hepatitis C virus
1004 3 0 C0019682	virs	HIV	- HIV
1005 3 0 C0019693	dsyn	HIV	- HIV Infections
1006 3 0 C0019699	fndg	HIV	- HIV Seropositivity

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
dsyn – Disease or Syndrome (DISO) virs – Virus LIVB fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB) **ftcn** – Functional Concept (CONC)

TOPIC 126: Patients admitted with a diagnosis of multiple sclerosis

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0332133 fndg admitted with a diagnosis of multiple sclerosis - Admitting diagnosis
1002 2 0 C0026769 dsyn admitted with a diagnosis of multiple sclerosis - Multiple Sclerosis
1003 2 0 C0332133 fndg admitted with a diagnosis of multiple sclerosis - Admitting diagnosis
1003 2 0 C1417325 gngm admitted with a diagnosis of multiple sclerosis - MS gene
1004 2 0 C1555318 **inpr** admitted with a diagnosis of multiple sclerosis - **Act Code - admitting diagnosis**
1004 2 0 C0026769 dsyn admitted with a diagnosis of multiple sclerosis - Multiple Sclerosis
1005 2 0 C1555318 **inpr** admitted with a diagnosis of multiple sclerosis - **Act Code - admitting diagnosis**
1005 2 0 C1417325 gngm admitted with a diagnosis of multiple sclerosis - MS gene

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
fndg – Finding (DISO) dsyn – Disease or Syndrome (DISO) gngm – Gene or Genome GENE	**podg** – Patient or Disabled Group (LIVB) **inpr** – Intellectual Product (CONC)

TOPIC 127: Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0184666 hlca admitted with morbid obesity - Hospital admission
1002 2 0 C0028756 dsyn admitted with morbid obesity - Obesity, Morbid
1003 3 0 C0277555 dsyn secondary diseases of diabetes - Secondary disease
1003 3 0 C0011847 dsyn secondary diseases of diabetes - Diabetes
1004 3 0 C0277555 dsyn secondary diseases of diabetes - Secondary disease
1004 3 0 C0011849 dsyn secondary diseases of diabetes - Diabetes
1005 4 0 C0020538 dsyn hypertension - Hypertensive disease
1006 4 0 C1963138 fndg hypertension - Hypertension Adverse Event

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC) dsyn – Disease or Syndrome (DISO) fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB)

TOPIC 128: Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0184666 hlca admitted for hip - Hospital admission
1002 2 0 C0019552 bpoc admitted for hip - Hip region structure
1003 2 0 C0184666 hlca admitted for hip - Hospital admission
1003 2 0 C0022122 bpoc admitted for hip -Bone structure of ischium
1004 3 0 C0187769 topp knee surgery - Operative procedure on knee
1005 4 0 C0332293 topp treated with anti-coagulant medications - Treated with
1005 4 0 C0003280 phsu treated with anti-coagulant medications - Anticoagulants
1005 4 0 C0013227 phsu treated with anti-coagulant medications - Pharmaceutical Preparations
1006 4 0 C0332293 topp treated with anti-coagulant medications - Treated with
1006 4 0 C0003280 phsu treated with anti-coagulant medications - Anticoagulants
1006 4 0 C0802604 **clna** treated with anti-coagulant medications - **Medications:Presence or Identity:Duration of the study:^Patient:Nominal**
1007 4 0 C0332293 topp treated with anti-coagulant medications - Treated with
1007 4 0 C0003280 phsu treated with anti-coagulant medications - Anticoagulants
1007 4 0 C2598133 **clna** treated with anti-coagulant medications - **Medications:-:Point in time:^Patient:-**
1008 4 0 C0332293 topp treated with anti-coagulant medications - Treated with

1008 4 0 C0848112 topp treated with anti-coagulant medications - on anti-coagulants
1008 4 0 C0013227 phsu treated with anti-coagulant medications - Pharmaceutical Preparations
1009 4 0 C0332293 topp treated with anti-coagulant medications - Treated with
1009 4 0 C0848112 topp treated with anti-coagulant medications - on anti-coagulants
1009 4 0 C0802604 **clna** treated with anti-coagulant medications -**Medications:Presence or Identity:Duration of the study:^Patient:Nominal**
1010 4 0 C0332293 topp treated with anti-coagulant medications - Treated with
1010 4 0 C0848112 topp treated with anti-coagulant medications - on anti-coagulants
1010 4 0 C2598133 **clna** treated with anti-coagulant medications - **Medications:-:Point in time:^Patient:-**
1011 5 0 C0687676 **tmco** post-op - **Post**
1012 5 0 C1704687 mnob post-op - Post Device Component

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC) bpoc – Body Part, Organ, or Organ Component (ANAT) topp – Thrapeutic or Preventive Procedure (PROC) phsu – Pharmacologic Substance (CHEM) mnob – Manufactured Object OBJC	**podg** – Patient or Disabled Group (LIVB) **clna** – Clinical Attribute (PHYS) **tmco** – Temporal Concept (CONC)

TOPIC 129: Patients admitted with chest pain and assessed with CT angiography

1001 1 0 C0030705 **podg** Patients - **Patients**
1002 2 0 C0184666 hlca admitted with chest pain - Hospital admission
1002 2 0 C0008031 sosy admitted with chest pain - Chest Pain
1003 2 0 C0184666 hlca admitted with chest pain - Hospital admission
1003 2 0 C2926613 **clna** admitted with chest pain - **Chest pain:Finding:Point in time:^Patient:Ordinal**
1004 3 0 C1516048 **acty** assessed with CT angiography - **Assessed**
1004 3 0 C1536105 diap assessed with CT angiography - CT angiography

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC) sosy – Sign or Symptom (DISO) diap – Diagnostic Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **clna** – Clinical Attribute (PHYS) **acty** – Activity (ACTI)

TOPIC 130: Children admitted with cerebral palsy who received physical therapy

1001 1 0 C0008059 aggp Children - Child
1002 1 0 C0680063 **famg** Children - **Offspring**
1003 2 0 C0184666 hlca admitted with cerebral palsy - Hospital admission
1003 2 0 C0007789 dsyn admitted with cerebral palsy - Cerebral Palsy
1004 3 0 C1514756 **qlco** received - **Receive**
1005 4 0 C0699718 bmod physical therapy - Physical therapy (field)
1006 4 0 C0949766 topp physical therapy - Physical therapy

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
aggp – Age Group (LIVB) hlca – Health Care Activity (PROC) dsyn – Disease or Syndrome (DISO) bmod – Biomedical Occupation or Discipline (OCCU) topp – Thrapeutic or Preventive Procedure (PROC)	**famg** – Family Group (LIVB) **qlco** – Qualitative Concept (CONC)

TOPIC 131: Patients who underwent minimally invasive abdominal surgery

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C0205281 **qlco**	minimally invasive abdominal surgery	- **Invasive**
1002 2 0 C0000726 blor	minimally invasive abdominal surgery	- Abdomen
1002 2 0 C0038894 bmod	minimally invasive abdominal surgery	- Surgery specialty
1003 2 0 C0205281 **qlco**	minimally invasive abdominal surgery	- **Invasive**
1003 2 0 C0000726 blor	minimally invasive abdominal surgery	- Abdomen
1003 2 0 C0038895 **ftcn**	minimally invasive abdominal surgery	- **Surgical aspects**
1004 2 0 C0205281 **qlco**	minimally invasive abdominal surgery	- **Invasive**
1004 2 0 C0000726 blor	minimally invasive abdominal surgery	- Abdomen
1004 2 0 C0543467 diap	minimally invasive abdominal surgery	- Operative Surgical Procedures

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
blor – Body Location or Region (ANAT(	**podg** – Patient or Disabled Group (LIVB)
bmod – Biomedical Occupation or Discipline (OCCU)	**qlco** – Qualitative Concept (CONC)
diap – Diagnostic Procedure (PROC)	**ftcn** – Functional Concept (CONC)

TOPIC 132: Patients admitted for surgery of the cervical spine for fusion or discectomy

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C0184666 hlca	admitted for surgery of the cervical spine	- Hospital admission
1002 2 0 C0742216 topp	admitted for surgery of the cervical spine	- CERVICAL SPINE SURGERY
1003 3 0 C0332466 **ftcn**	for fusion	- **Fused structure**
1004 3 0 C1293131 topp	for fusion	- Fusion procedure
1005 4 0 C0206078 topp	discectomy	– Diskectomy

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC)	**podg** – Patient or Disabled Group (LIVB)
topp – Thrapeutic or Preventive Procedure (PROC)	**ftcn** – Functional Concept (CONC)

TOPIC 133: Patients admitted for care who take herbal products for osteoarthritis

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C0184666 hlca	admitted for care	- Hospital admission
1002 2 0 C1947933 **acty**	admitted for care	- **care activity**
1003 3 0 C1515187 hlca	take	- Take
1004 4 0 C3146288 **qlco**	herbal products for osteoarthritis	- **Herbal**
1004 4 0 C1514468 **enty**	herbal products for osteoarthritis	- **product**
1004 4 0 C0029408 dsyn	herbal products for osteoarthritis	- Degenerative polyarthritis

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC)	**podg** – Patient or Disabled Group (LIVB)
dsyn – Disease or Syndrome (DISO)	**acty** – Activity (ACTI)
	qlco – Qualitative Concept (CONC)
	enty – Entity OBJC

TOPIC 134: Patients admitted with chronic seizure disorder to control seizure activity

1001 1 0 C0030705 **podg**	Patients	- **Patients**
1002 2 0 C0184666 hlca	admitted with chronic seizure disorder to	- Hospital admission
1002 2 0 C0008679 dsyn	admitted with chronic seizure disorder to	- Chronic disease
1002 2 0 C0036572 sosy	admitted with chronic seizure disorder to	- Seizures
1003 2 0 C0184666 hlca	admitted with chronic seizure disorder to	- Hospital admission
1003 2 0 C0008679 dsyn	admitted with chronic seizure disorder to	- Chronic disease
1003 2 0 C1959629 fndg	admitted with chronic seizure disorder to	- Seizure Adverse Event
1004 2 0 C0184666 hlca	admitted with chronic seizure disorder to	- Hospital admission
1004 2 0 C0205191 **tmco**	admitted with chronic seizure disorder to	- **chronic**
1004 2 0 C0014544 dsyn	admitted with chronic seizure disorder to	- Epilepsy
1005 3 0 C0009932 **grup**	control	- **Control Groups**
1006 3 0 C0243148 **qlco**	control	- **control aspects**
1007 3 0 C1550141 **sbst**	control	- **control substance**
1008 3 0 C1882979 **cnce**	control	- **Scientific Control**
1009 3 0 C2587213 **ftcn**	control	- **Control function**
1010 4 0 C1148454 fndg	seizure activity	- Seizure activity

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
hlca – Health Care Activity (PROC) dsyn – Disease or Syndrome (DISO) sosy – Sign or Symptom (DISO) fndg – Finding (DISO)	**podg** – Patient or Disabled Group (LIVB) **grup** – Group LIVB **qlco** – Qualitative Concept (CONC) **sbst** – Substance OBJC **cnce** – Conceptual Entity CONC **ftcn** – Functional Concept (CONC) **tmco** – Temporal Concept (CONC)

TOPIC 135: Cancer patients with liver metastasis treated in the hospital who underwent a procedure

1001 1 0 C0030705 **podg**	Cancer patients with liver metastasis	- **Patients**
1001 1 0 C0494165 neop	Cancer patients with liver metastasis	- Secondary malignant neoplasm of liver
1002 2 0 C0332293 topp	treated in the hospital	- Treated with
1002 2 0 C0019994 hcro	treated in the hospital	- Hospitals
1003 2 0 C0332293 topp	treated in the hospital	- Treated with
1003 2 0 C1510665 **qlco**	0treated in the hospital	- **Hospital environment**
1004 2 0 C1522326 **ftcn**	treated in the hospital	- **Treating**
1004 2 0 C0019994 hcro	treated in the hospital	- Hospitals
1005 2 0 C1522326 **ftcn**	treated in the hospital	- **Treating**
1005 2 0 C1510665 **qlco**	treated in the hospital	- **Hospital environment**
1006 3 0 C0184661 diap	a procedure	- Interventional procedure
1007 3 0 C1948041 topp	a procedure	- Surgical and medical procedures
1008 3 0 C2700391 **acty**	a procedure	- **Procedure (set of actions)**

TIPOS SEMÁNTICOS ESPECÍFICOS	TIPOS SEMÁNTICOS GENÉRICOS
neop – Neoplastic Process (DISO) topp – Thrapeutic or Preventive Procedure (PROC) hcro – Health Care Related Organization (ORGA) diap – Diagnostic Procedure (PROC)	**podg** – Patient or Disabled Group (LIVB) **qlco** – Qualitative Concept (CONC) **ftcn** – Functional Concept (CONC) **acty** – Activity (ACTI)

Apéndice B

Report32496

- **Documento médico original con filtrado previo**

CHEST PAIN

Otorhinolaryngology
Consultation Report

Right ear and right lateral face pain.

This patient is a very poor historian. He relates that approximately four days ago he noted pain on the right side of his face and ear. Over the course of the four days, this has gotten progressively better. The patient has had no history of ear disease. There is no tinnitus, there is no dizziness, and there is questionable hearing loss.

The patient is an alert male in no acute distress. Ears - The external canals, tympanic membranes, and middle ear are within normal limits. Face –There is good facial nerve movement. There is no palpable mass in his face. Neck - There is no cervical adenopathy. Right face ear pain, unknown etiology. Since the patient has improved since onset and there is no obvious etiology, I feel at this time no further evaluation or treatment is indicated. I did ask him to return to the office in one week if there has been no resolution of the problem. I will let you know if he comes to the office.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0029892 bmod Otorhinolaryngology Consultation Report
4 0 C0229298 bpoc Right ear
5 0 C1289033 bpoc Right ear
6 0 C0015468 sosy right lateral face pain.
7 0 C1962975 fndg right lateral face pain.
8 0 C0030705 podg This patient
9 0 C0439048 fndg a very poor historian.
10 0 C0163712 orch relates
12 0 C1316572 clna noted pain on the right side of his face
12 0 C0030193 sosy noted pain on the right side of his face
12 0 C0230025 blor noted pain on the right side of his face
16 0 C0013443 bpoc ear.
17 0 C0521421 bpoc ear.
21 0 C0030705 podg The patient
37 0 C0013443 bpoc Ears -
38 0 C0521421 bpoc Ears -
40 0 C1517034 bpoc The external canals,
41 0 C0041445 bpoc tympanic membranes,
42 0 C0013455 bsoj middle ear
43 0 C1268972 bpoc middle ear
44 0 C1265570 fndg are within normal limits.
45 0 C0015450 blor Face - There
46 0 C1281591 bpoc Face - There
47 0 C0015462 bpoc good facial nerve movement.
47 0 C0026649 orgf good facial nerve movement.
52 0 C0027530 blor Neck - There
53 0 C1281592 bpoc Neck - There
55 0 C0229298 bpoc Right face ear pain,
55 0 C0015450 blor Right face ear pain,
55 0 C0030193 sosy Right face ear pain,

56 0 C1281591 bpoc Right face ear pain,
57 0 C1289033 bpoc Right face ear pain,
59 0 C0015468 sosy Right face ear pain,
59 0 C0013443 bpoc Right face ear pain,
60 0 C0521421 bpoc Right face ear pain,
61 0 C1962975 fndg Right face ear pain,
63 0 C0013456 sosy Right face ear pain,
65 0 C0743626 fndg unknown etiology.
66 0 C0030705 podg the patient
84 0 C1444656 fndg indicated.

- **Distribución de Freciencia de Conceptos (Densidad Cualitativa):**

C0521421 bpoc Right face ear pain, (3)
C0030705 podg This patient (3)
C0013443 bpoc Right face ear pain, (3)
C1962975 fndg right lateral face pain. (2)
C1289033 bpoc Right face ear pain, (2)
C1281591 bpoc Right face ear pain, (2)
C0229298 bpoc Right face ear pain, (2)
C0030193 sosy noted pain on the right side of his (2)
---------- Corte 2º Cuartil (50 %) ----------
C0015468 sosy right lateral face pain. (2)
C0015450 blor Right face ear pain, (2)

"Report32496" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	**This patient** (3)
ANAT - Parte del cuerpo/Organo (bpoc):	**Right face ear pain** (3)
ANAT - Localización o región del cuerpo (blor):	**Right face ear pain** (2)
DISO - Localización del trastorno (fndg):	**right lateral face pain** (2)
DISO - Signo o Sintoma (sosy):	**noted pain on the right side of his** (2)
DISO - Signo o Sintoma (sosy):	**right lateral face pain** (2)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report32496" (Densidad Cuantitativa + Cualitativa)
****

LIVB - Tipo de paciente (podg):	**This patient (3)**
ANAT - Parte del cuerpo/Organo (bpoc):	**Right face ear pain (14)**
ANAT - Localización o región del cuerpo (blor):	**Right face ear pain (14)**
DISO - Localización del trastorno (fndg):	**right lateral face pain (4)**
DISO - Signo o Sintoma (sosy):	**noted pain on the right side of his (2)**
DISO - Signo o Sintoma (sosy):	**right lateral face pain (4)**

****

Resumen Final:

Resumen del documento "Report32496":

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Right face ear pain (14)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **right lateral face pain (4)**
- **noted pain on the right side of his (2)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **This patient (3)**

Report32288

- **Documento médico original con filtrado previo**

CHEST PAIN

Otorhinolaryngology
Consultation Report

Right nasal obstruction.
This patient had an episode of vomiting today and the vomitus came out his right nose. Since that time, he has had a feeling of it being swollen.
The patient is an alert male in no acute distress. The nasal cavity is open bilaterally with no foreign body identified. Irritation secondary to the acid in his vomitus. Ocean spray 2 sprays twice a day for 4 days.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0029892 bmod Otorhinolaryngology Consultation Report
4 0 C0027429 fndg Right nasal obstruction.
5 0 C0030705 podg This patient
6 0 C0042963 sosy an episode of vomiting today
6 0 C0310367 antb an episode of vomiting today
8 0 C1963281 fndg an episode of vomiting today
10 0 C0042965 bdsu the vomitus
12 0 C1608512 bdsu the vomitus
15 0 C0028429 bpoc his right nose.
16 0 C1278896 bpoc his right nose.
21 0 C0038999 fndg swollen.
22 0 C0030705 podg The patient
28 0 C0027423 bsoj The nasal cavity
29 0 C1280672 bsoj The nasal cavity
31 0 C0016542 inpo bilaterally with no foreign body
35 0 C0441723 phpr Irritation
36 0 C1706307 fndg Irritation
39 0 C0042965 bdsu in his vomitus.
41 0 C1608512 bdsu in his vomitus.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C1608512 bdsu in his vomitus. (2)
C0042965 bdsu in his vomitus. (2)
C0030705 podg This patient (2)
C2926613 clna CHEST PAIN (1)
C1963281 fndg an episode of vomiting today (1)
C1706307 fndg Irritation (1)
C1280672 bsoj The nasal cavity (1)
--------- ---------- ***Corte 2º Cuartil (50 %)*** ----------
C1278896 bpoc his right nose (1).
C0441723 phpr Irritation (1)
C0310367 antb an episode of vomiting today (1)
C0042963 sosy an episode of vomiting today (1)
C0038999 fndg swollen (1).
C0029892 bmod Otorhinolaryngology Consultation Report (1)
C0028429 bpoc his right nose (1)
C0027429 fndg Right nasal obstruction. (1)
C0027423 bsoj The nasal cavity (1)
C0016542 inpo bilaterally with no foreign body (1)

"Report32288" (Densidad Cualitativa)
****
LIVB - Tipo de paciente (podg): **This patient** (2)
ANAT - Substancia del cuerpo (bdsu): **in his vomitus** (2)
****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report32288" (Densidad Cualitativa + Cuantitativa)
****
LIVB - Tipo de paciente (podg): **This patient (2)**
ANAT - Substancia del cuerpo (bdsu): **in his vomitus (4)**
DISO - Localización del trastorno (fndg): **an episode of vomiting today (3)**
****

Resumen Final:

Resumen del documento "Report32288":

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **in his vomitus (4)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **an episode of vomiting today (3)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **This patient (2)**

Report33510

- **Documento médico original con filtrado previo**

CHEST PAIN

Gastroenterology
Operative Report

COLONOSCOPY TO CECUM.

Abdominal pain.
Versed 2 mg and fentanyl 100.
After explanation of the risks of the procedure, perforation, hemorrhage, and alternatives, the scope was introduced into the rectum and sigmoid all the way to the cecum. There were no other lesions seen. Then, the scope was withdrawn terminating the procedure. Normal colon except for mild hemorrhoids. High-fiber diet.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0017163 bmod Gastroenterology Operative Report
6 0 C0009378 diap COLONOSCOPY TO CECUM.
6 0 C0007531 bpoc COLONOSCOPY TO CECUM.
7 0 C1278925 bpoc COLONOSCOPY TO CECUM.
10 0 C0000737 sosy Abdominal pain.
11 0 C0042553 orch Versed
13 0 C2346927 elii 2 mg
18 0 C0015846 orch fentanyl 100.
21 0 C0184661 diap of the procedure,
22 0 C1948041 topp of the procedure,
24 0 C0549099 fndg perforation,
25 0 C1881710 phpr perforation,
26 0 C0019080 patf hemorrhage,
29 0 C0034896 bpoc introduced into the rectum
30 0 C1278926 bpoc introduced into the rectum
31 0 C0227391 bpoc sigmoid
32 0 C0007531 bpoc all the way to the cecum.
33 0 C1278925 bpoc all the way to the cecum.
38 0 C0424092 fndg withdrawn
41 0 C2825032 mobd withdrawn
44 0 C0184661 diap the procedure.
45 0 C1948041 topp the procedure.
47 0 C0940500 fndg Normal colon except for mild hemorrhoids.
47 0 C0019112 acab Normal colon except for mild hemorrhoids.
48 0 C1963134 fndg Normal colon except for mild hemorrhoids.
49 0 C0301568 topp High-fiber diet.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C1948041 topp of the procedure, (2)
C1278925 bpoc all the way to the cecum. (2)
C0184661 diap of the procedure, (2)
C0007531 bpoc COLONOSCOPY TO CECUM. (2)
C2926613 clna CHEST PAIN (1)
C2825032 mobd withdrawn (1)
C2346927 elii 2 mg (1)
C1963134 fndg Normal colon except for mild hemorrhoids. (1)
C1881710 phpr perforation, (1)
---------- Corte 2º Cuartil (50 %) ----------
C1278926 bpoc introduced into the rectum (1)
C0940500 fndg Normal colon except for mild hemorrhoids. (1)
C0549099 fndg perforation, (1)
C0424092 fndg withdrawn (1)
C0301568 topp High-fiber diet. (1)
C0227391 bpoc sigmoid (1)

C0042553 orch Versed (1)
C0034896 bpoc introduced into the rectum (1)
C0019112 acab Normal colon except for mild hemorrhoids. (1)
C0019080 patf hemorrhage, (1)
C0017163 bmod Gastroenterology Operative Report (1)
C0015846 orch fentanyl 100. (1)
C0009378 diap COLONOSCOPY TO CECUM. (1)
C0008031 sosy CHEST PAIN (1)

"Report33510" (Densidad Cualitativa)
****
ANAT - Parte del cuerpo/Organo (bpoc): **all the way to the cecum** (2)
ANAT - Parte del cuerpo/Organo (bpoc): **COLONOSCOPY TO CECUM** (2)
PROC - Procedimiento terapeútico (topp): **of the procedure** (2)
PROC - Procedimiento de diagnóstico (diap): **of the procedure** (2)
****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report33510" (Densidad Cualitativa + Cuantitativa)
****
ANAT - Parte del cuerpo/Organo (bpoc): **all the way to the cecum (2)**
ANAT - Parte del cuerpo/Organo (bpoc): **COLONOSCOPY TO CECUM (3)**
DISO - Localización del trastorno (fndg): **Normal colon except for mild hemorrhoids. (3)**
DISO - Anormalidad adquirida (acab): **Normal colon except for mild hemorrhoids. (3)**
PROC - Procedimiento terapeútico (topp): **of the procedure (4)**
PROC - Procedimiento de diagnóstico (diap): **of the procedure (4)**
****

Resumen Final:

Resumen del documento "Report33510":

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **of the procedure (4)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **COLONOSCOPY TO CECUM (3)**
- **all the way to the cecum (2)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **Normal colon except for mild hemorrhoids. (3)**

Report32095

- **Documento médico original con filtrado previo**

CHEST PAIN

CARDIOLOGY

This in 70s-year-old gentleman was admitted with a history of chest pain and epigastric pain. Therefore, for evaluation of the chest pain, this consultation is obtained. As per the history, the patient says that he has been having left-sided chest pain with radiation to the left arm. He thought taking nitroglycerin would relieve his pain and he was taking three to four nitroglycerin for the last about two days without much relief, but the patient also complains of abdominal pain. The patient has a known history of coronary artery disease but no associated nausea, vomiting or diaphoresis. The patient had a cardiac catheterization done on Jun 16 2006, which revealed total occlusion of the LAD with multiple sequential 70% to 80% stenosis in the proximal mid vessel. The diagonal artery lateral from the LAD was occluded. The circumflex had approximately 80% stenosis subtending a large distal obtuse marginal circulation. The intermediate ramus and smaller obtuse marginal arteries were occluded. The right coronary provided an acute marginal collateral to the distal LAD and had a patent saphenous vein graft to the PDA and posterolateral branch. The left internal mammary artery graft to the LAD was occluded as was the remaining saphenous vein grafts to the left-sided circulation. The patient had stenting done on the same day. Proximal circumflex stenosis was directly stented with a Taxus stent and also the distal LAD was dilated and sequentially stented with a Taxus stent and good flow was established. No history of any congestive heart failure.

Past medical history is significant for:
1. History of CVA.
2. Hypertension.
3. Coronary artery disease.
4. Coronary artery bypass graft about eight years ago.
5. Arthritis.
6. Nasal surgery.
7. Status post hemorrhoidectomy.
8. Status post stenting.
No history of any diabetes mellitus.

Current medications. The patient is on:
1. Aspirin 325 mg by mouth daily.
2. Lipitor 20 mg by mouth daily.
3. Plavix 75 mg.
4. Lovenox 40 mg subcu daily.
5. Metoprolol 200 mg by mouth daily.
6. Nitroglycerin one-inch every six hours.
7. Protonix 40 mg daily.
8. Albuterol on a p.r.n. basis.

Mother had congestive heart failure. There is no history of any diabetes or kidney disease. Denies any history of hypertension. He lives with his son.
The patient denies tobacco or alcohol use. The patient complains of epigastric distress. The rest of the review of systems is unremarkable.
Blood pressure 160/97. Afebrile. Pulse is 50. General: The patient is a black male, age of about 75. He is normally built for his and not in any acute distress, but the patient is concerned about his chest pain and the epigastric pain. Neck: No distended neck veins. No bruits of the carotids. Chest: Both hemithoraces are equal. Emphysematous changes of the chest are noted. The lungs reveal normal vesicular breathing heard all over the lung field. Cardiovascular system: S1 and S2 are heard. There is no murmur or gallop and no thrill felt. The apical impulse is within the left midclavicular line. The abdomen is soft, bowel sounds are heard.
Extremities: No pedal edema. The Homans sign is negative. Neurological examination is within normal limits.

EKG was normal.
Chest x-ray - no active disease.
Other lab work: Hemoglobin 13.5 and hematocrit 40.2. BNP is 191. Troponin I - one set is negative. Potassium 4.2. BUN 19 and creatinine 1.7. Blood sugar is 96.

1. Possible angina.
2. Coronary artery disease, status post CABG with occlusion of the left-sided grafts and a patent RCA and posterior descending branch, status post stenting of the LAD and the proximal circumflex.

3. Hypertension.
4. Hyperlipidemia.
Continue beta blocker and Plavix.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0007189 bmod CARDIOLOGY
7 0 C0024902 sosy admitted with a history of chest pain
10 0 C0262926 fndg admitted with a history of chest pain
12 0 C2004062 fndg admitted with a history of chest pain
13 0 C0232493 sosy epigastric pain.
17 0 C0008031 sosy of the chest pain,
18 0 C2926613 clna of the chest pain,
24 0 C0262926 fndg per the history,
26 0 C2004062 fndg per the history,
27 0 C0030705 podg the patient says
27 0 C0080151 dsyn the patient says
28 0 C0541828 sosy left-sided chest pain with radiation
28 0 C0034519 npop left-sided chest pain with radiation
30 0 C0851346 npop left-sided chest pain with radiation
31 0 C1522449 topp left-sided chest pain with radiation
32 0 C1524020 topp left-sided chest pain with radiation
34 0 C0230347 bpoc to the left arm.
35 0 C1268256 bpoc to the left arm.
36 0 C0017887 orch thought taking nitroglycerin
37 0 C0885378 orch thought taking nitroglycerin
41 0 C0030193 sosy his pain
42 0 C0017887 orch taking three to four nitroglycerin
43 0 C0885378 orch taking three to four nitroglycerin
49 0 C0030705 podg the patient also
50 0 C0000737 sosy complains of abdominal pain.
51 0 C0030705 podg The patient
52 0 C1881055 clna a known history of coronary artery disease
59 0 C0030705 podg The patient
60 0 C0018795 diap a cardiac catheterization
64 0 C0011382 ortf total occlusion of the LAD
65 0 C0028778 patf total occlusion of the LAD
67 0 C1110554 patf total occlusion of the LAD
68 0 C1882137 phpr total occlusion of the LAD
78 0 C0009814 patf to 80% stenosis
80 0 C0947637 patf to 80% stenosis
81 0 C1261287 anab to 80% stenosis
83 0 C0005847 bpoc in the proximal mid vessel.
84 0 C1282006 bpoc The diagonal artery
86 0 C0028778 patf occluded.
89 0 C0009814 patf approximately 80% stenosis
91 0 C0947637 patf approximately 80% stenosis
92 0 C1261287 anab approximately 80% stenosis
94 0 C1305142 blor a large distal obtuse marginal circulation.
95 0 C1516559 orgf a large distal obtuse marginal circulation.
98 0 C2827755 lbtr The intermediate ramus
99 0 C0226001 bpoc smaller obtuse marginal arteries
99 0 C1305142 blor smaller obtuse marginal arteries
100 0 C1282822 bpoc smaller obtuse marginal arteries
101 0 C0028778 patf occluded.
103 0 C0018787 bpoc The right coronary
105 0 C1305794 blor an acute marginal collateral to the distal LAD
105 0 C1275670 bpoc an acute marginal collateral to the distal LAD
107 0 C0729538 bpoc a patent saphenous vein graft to the PDA
107 0 C0013274 cgab a patent saphenous vein graft to the PDA
108 0 C0226047 bpoc a patent saphenous vein graft to the PDA
111 0 C1253959 bpoc posterolateral branch.
116 0 C0226276 bpoc The left internal mammary artery graft to the LAD
117 0 C0332835 tisu The left internal mammary artery graft to the LAD
118 0 C1961139 topp The left internal mammary artery graft to the LAD
119 0 C0028778 patf occluded
121 0 C0729538 bpoc the remaining saphenous vein grafts to the left-sided circulation.
122 0 C1516559 orgf the remaining saphenous vein grafts to the left-sided circulation.
123 0 C0030705 podg The patient
124 0 C0038257 medd stenting

125 0 C2348535 topp stenting
130 0 C0009814 patf Proximal circumflex stenosis
132 0 C0947637 patf Proximal circumflex stenosis
133 0 C1261287 anab Proximal circumflex stenosis
136 0 C0038257 medd stented with a Taxus stent
139 0 C0038257 medd stented with a Taxus stent
140 0 C0806140 npop good flow
144 0 C0262926 fndg Past medical history
145 0 C0455458 fndg Past medical history
152 0 C0038454 dsyn History of CVA.
155 0 C0262926 fndg History of CVA.
157 0 C2004062 fndg History of CVA.
161 0 C0020538 dsyn Hypertension.
162 0 C1963138 fndg Hypertension.
166 0 C0010054 dsyn Coronary artery disease.
167 0 C0010068 dsyn Coronary artery disease.
168 0 C1956346 dsyn Coronary artery disease.
170 0 C0010055 topp Coronary artery bypass graft about eight years ago.
171 0 C1260596 medd Coronary artery bypass graft about eight years ago.
174 0 C0003864 dsyn Arthritis.
175 0 C0188970 topp Nasal surgery.
176 0 C0019108 topp Status post hemorrhoidectomy.
177 0 C0038257 medd Status post stenting.
178 0 C2348535 topp Status post stenting.
180 0 C0746467 fndg Current medications.
181 0 C0030705 podg The patient
187 0 C0992015 clnd Aspirin 325 mg by mouth daily.
191 0 C0593906 orch Lipitor 20 mg by mouth daily.
192 0 C2346927 elii Lipitor 20 mg by mouth daily.
196 0 C0633084 orch Plavix 75 mg.
197 0 C2346927 elii Plavix 75 mg.
200 0 C2346927 elii Lovenox 40 mg subcu daily.
203 0 C0025859 orch Metoprolol 200 mg by mouth daily.
204 0 C2346927 elii Metoprolol 200 mg by mouth daily.
205 0 C0017887 orch Nitroglycerin one-inch
208 0 C0885378 orch Nitroglycerin one-inch
212 0 C0876139 orch Protonix 40 mg daily.
213 0 C2346927 elii Protonix 40 mg daily.
214 0 C0001927 orch Albuterol on a p.r.n. basis.
214 0 C0378444 orch Albuterol on a p.r.n. basis.
215 0 C0026591 famg Mother
216 0 C0018802 dsyn congestive heart failure.
225 0 C0557132 fndg lives with his son.
227 0 C0030705 podg The patient
236 0 C0030705 podg The patient complains of epigastric distress.
236 0 C0221490 sosy The patient complains of epigastric distress.
237 0 C0488564 clna The rest of the review of systems
238 0 C0488565 clna The rest of the review of systems
244 0 C0005823 orgf Blood pressure 160/97.
246 0 C1271104 fndg Blood pressure 160/97.
248 0 C1272641 fndg Blood pressure 160/97.
250 0 C0277797 fndg Afebrile.
252 0 C0232117 clna Pulse
254 0 C1947910 phpr Pulse
257 0 C0030705 podg The patient
258 0 C0005680 popg a black male,
258 0 C0024554 fndg a black male,
259 0 C0086582 popg a black male,
263 0 C0027567 popg a black male,
268 0 C0085756 popg a black male,
286 0 C0030705 podg the patient
287 0 C0008031 sosy concerned about his chest pain
288 0 C2926613 clna concerned about his chest pain
291 0 C0232493 sosy the epigastric pain.
292 0 C0027530 blor Neck
293 0 C1281592 bpoc Neck
298 0 C0817096 blor Chest
299 0 C1527391 blor Chest
300 0 C0934569 blor hemithoraces
301 0 C1827591 blor hemithoraces
304 0 C0558236 fndg Emphysematous changes of the chest
305 0 C1316572 clna noted.
307 0 C0024109 bpoc The lungs
309 0 C0231857 sosy normal vesicular breathing heard all over the lung field.
309 0 C0234725 fndg normal vesicular breathing heard all over the lung field.
309 0 C0225759 blor normal vesicular breathing heard all over the lung field.

310 0 C0007226 bdsy Cardiovascular system
311 0 C1269562 bdsy Cardiovascular system
312 0 C1179705 blor S2
316 0 C1455844 fndg heard.
318 0 C0232269 fndg no thrill
319 0 C2678517 fndg no thrill
323 0 C0738480 blor is within the left midclavicular line.
324 0 C0000726 blor The abdomen
325 0 C0230168 blor The abdomen
326 0 C1281594 bpoc The abdomen
327 0 C0232693 fndg soft, bowel sounds
329 0 C1455844 fndg heard.
330 0 C0015385 bpoc Extremities
331 0 C0278454 bpoc Extremities
333 0 C0231781 sosy The Homans sign
335 0 C1513916 fndg negative.
338 0 C0027853 diap Neurological examination
339 0 C1265570 fndg is within normal limits.
340 0 C0013798 fndg EKG
341 0 C1623258 diap EKG
345 0 C0039985 diap Chest x-ray -
346 0 C1114557 clna Chest x-ray -
348 0 C1955473 fndg Other lab work
350 0 C0018935 lbpr hematocrit 40.2.
351 0 C0518014 lbtr hematocrit 40.2.
352 0 C1095989 lbpr BNP
360 0 C1513916 fndg negative.
363 0 C0032821 elii Potassium 4.2.
365 0 C0597277 elii Potassium 4.2.
366 0 C0005845 lbpr BUN 19
369 0 C1561535 fndg creatinine 1.7.
374 0 C0002962 sosy Possible angina.
380 0 C0010054 dsyn Coronary artery disease,
381 0 C0010068 dsyn Coronary artery disease,
382 0 C1956346 dsyn Coronary artery disease,
383 0 C0010055 topp status post CABG
384 0 C0011382 ortf with occlusion
385 0 C0028778 patf with occlusion
387 0 C1110554 patf with occlusion
388 0 C1882137 phpr with occlusion
390 0 C0332835 tisu of the left-sided grafts
396 0 C1253959 bpoc posterior descending branch,
406 0 C0038257 medd status post stenting
407 0 C2348535 topp status post stenting
412 0 C0020538 dsyn Hypertension.
413 0 C1963138 fndg Hypertension.
415 0 C0020473 dsyn Hyperlipidemia.
418 0 C0633084 orch Plavix.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C0030705 podg The patient complains of epigastric distress. (10)
C2346927 elii Lipitor 20 mg by mouth daily. (5)
C2346927 elii Plavix 75 mg. (5)
C2346927 elii Lovenox 40 mg subcu daily. (5)
C2346927 elii Metoprolol 200 mg by mouth daily. (5)
C2346927 elii Protonix 40 mg daily. (5)
C0038257 medd stented with a Taxus stent (5)
C0028778 patf total occlusion of the LAD (5)
C0262926 fndg admitted with a history of chest pain (4)
C2926613 clna concerned about his chest pain (3)
C2348535 topp status post stenting (3)
C2004062 fndg admitted with a history of chest pain (3)
C1261287 anab Proximal circumflex stenosis (3)
C0947637 patf Proximal circumflex stenosis (3)
C0885378 orch taking three to four nitroglycerin (3)
C0017887 orch taking three to four nitroglycerin (3)
C0009814 patf Proximal circumflex stenosis (3)
C0008031 sosy concerned about his chest pain (3)
C1963138 fndg Hypertension. (2)
C1956346 dsyn Coronary artery disease. (2)
C1882137 phpr total occlusion of the LAD (2)
C1516559 orgf the remaining saphenous vein grafts to the left-sided (2)
C1305142 blor a large distal obtuse marginal circulation. (2)

C1253959 bpoc posterior descending branch, (2)
C1110554 patf total occlusion of the LAD (2)
C0729538 bpoc the remaining saphenous vein grafts to the left-sided (2)
C0633084 orch Plavix 75 mg. (2)
C0332835 tisu The left internal mammary artery graft to the (2)
C0232493 sosy the epigastric pain. (2)
C0020538 dsyn Hypertension. (2)
C0011382 ortf total occlusion of the LAD (2)
C0010068 dsyn Coronary artery disease. (2)
C0010055 topp Coronary artery bypass graft about eight years ago. (2)
C0010054 dsyn Coronary artery disease. (2)
C2827755 lbtr The intermediate ramus (1)
C2678517 fndg no thrill (1)
---------- Corte 2º Cuartil (50 %) ----------
C1961139 topp The left internal mammary artery graft to the (1)
C1955473 fndg Other lab work (1)
C1947910 phpr Pulse (1)
C1881055 clna a known history of coronary artery disease (1)
C1827591 blor hemithoraces (1)
C1623258 diap EKG (1)
C1561535 fndg creatinine 1.7. (1)
C1527391 blor Chest (1)
C1524020 topp left-sided chest pain with radiation (1)
C1522449 topp left-sided chest pain with radiation (1)
C1316572 clna noted. (1)
C1305794 blor an acute marginal collateral to the distal LAD (1)
C1282822 bpoc smaller obtuse marginal arteries (1)
C1282006 bpoc The diagonal artery (1)
C1281594 bpoc The abdomen (1)
C1281592 bpoc Neck (1)
C1275670 bpoc an acute marginal collateral to the distal LAD (1)
C1272641 fndg Blood pressure 160/97. (1)
C1271104 fndg Blood pressure 160/97. (1)
C1269562 bdsy Cardiovascular system (1)
C1268256 bpoc to the left arm. (1)
C1265570 fndg is within normal limits. (1)
1260596 medd Coronary artery bypass graft about eight years ago. (1)
C1179705 blor S2 (1)
C1114557 clna Chest x-ray - (1)
C1095989 lbpr BNP (1)
C0992015 clnd Aspirin 325 mg by mouth daily. (1)
C0934569 blor hemithoraces (1)
C0876139 orch Protonix 40 mg daily. (1)
C0851346 npop left-sided chest pain with radiation (1)
C0817096 blor Chest (1)
C0806140 npop good flow (1)
C0746467 fndg Current medications. (1)
C0738480 blor is within the left midclavicular line. (1)
C0597277 elii Potassium 4.2. (1)
C0593906 orch Lipitor 20 mg by mouth daily. (1)
C0558236 fndg Emphysematous changes of the chest (1)
C0557132 fndg lives with his son. (1)
C0541828 sosy left-sided chest pain with radiation (1)
C0518014 lbtr hematocrit 40.2. (1)
C0488565 clna The rest of the review of systems (1)
C0488564 clna The rest of the review of systems (1)
C0455458 fndg Past medical history (1)
C0378444 orch Albuterol on a p.r.n. basis. (1)
C0278454 bpoc Extremities (1)
C0277797 fndg Afebrile. (1)
C0234725 fndg normal vesicular breathing heard all over the lung (1)
C0232693 fndg soft, bowel sounds (1)
C0232269 fndg no thrill (1)
C0232117 clna Pulse (1)
C0231857 sosy normal vesicular breathing heard all over the lung (1)
C0231781 sosy The Homans sign (1)
C0230347 bpoc to the left arm. (1)
C0230168 blor The abdomen (1)
C0226276 bpoc The left internal mammary artery graft to the (1)
C0226047 bpoc the remaining saphenous vein grafts to the left-sided (1)
C0226001 bpoc smaller obtuse marginal arteries (1)
C0225759 blor normal vesicular breathing heard all over the lung (1)
C0188970 topp Nasal surgery. (1)
C0086582 popg a black male, (1)
C0085756 popg a black male, (1)
C0080151 dsyn the patient says (1)

C0039985 diap Chest x-ray - (1)
C0038454 dsyn History of CVA. (1)
C0034519 npop left-sided chest pain with radiation (1)
C0032821 elii Potassium 4.2. (1)
C0030193 sosy his pain (1)
C0027853 diap Neurological examination (1)
C0027567 popg a black male, (1)
C0027530 blor Neck (1)
C0026591 famg Mother (1)
C0025859 orch Metoprolol 200 mg by mouth daily. (1)
C0024902 sosy admitted with a history of chest pain (1)
C0024554 fndg a black male, (1)
C0024109 bpoc The lungs (1)
C0020473 dsyn Hyperlipidemia. (1)
C0019108 topp Status post hemorrhoidectomy. (1)
C0018935 lbpr hematocrit 40.2. (1)
C0018802 dsyn congestive heart failure. (1)
C0018795 diap a cardiac catheterization (1)
C0018787 bpoc The right coronary (1)
C0015385 bpoc Extremities (1)
C0013798 fndg EKG (1)
C0013274 cgab a patent saphenous vein graft to the PDA (1)
C0007226 bdsy Cardiovascular system (1)
C0007189 bmod CARDIOLOGY (1)
C0005847 bpoc in the proximal mid vessel. (1)
C0005845 lbpr BUN 19 (1)
C0005823 orgf Blood pressure 160/97. (1)
C0005680 popg a black male, (1)
C0003864 dsyn Arthritis. (1)
C0002962 sosy Possible angina. (1)
C0001927 orch Albuterol on a p.r.n. basis. (1)
C0000737 sosy complains of abdominal pain. (1)

"Report32095" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	The patient complains of epigastric distress (10)
ANAT - Parte del cuerpo/Organo (bpoc):	posterior descending branch (2)
ANAT - Parte del cuerpo/Organo (bpoc):	the remaining saphenous vein grafts to the left-sided (2)
ANAT - Tejido (tisu):	The left internal mammary artery graft to the (2)
ANAT - Localización o región del cuerpo (blor):	a large distal obtuse marginal circulation (2)
DISO - Localización del trastorno (fndg):	admitted with a history of chest pain (4)
DISO - Localización del trastorno (fndg):	Hypertension (2)
DISO - Función Patológica (patf):	total occlusion of the LAD (5)
DISO - Función Patológica (patf):	Proximal circumflex stenosis (3)
DISO - Anormalidad Anatómica (anab):	Proximal circumflex stenosis (3)
DISO - Signo o Sintoma (sosy):	concerned about his chest pain (3)
DISO - Signo o Sintoma (sosy):	the epigastric pain (2)
DISO - Enfermedad o síndrome (dsyn):	Coronary artery disease (2)
DISO - Enfermedad o síndrome (dsyn):	Hypertension (2)
PHYS - Atributo Clínico (clna):	concerned about his chest pain (3)
PHYS - Función Orgánica (orgf):	the remaining saphenous vein grafts to the left-sided (3)
PHYS - Órgano o Tejido (ortf):	total occlusion of the LAD (2)
PROC - Procedimiento terapeútico (topp):	status post stenting (3)
PROC - Procedimiento terapeútico (topp):	Coronary artery bypass graft about eight years ago (2)
CHEM - Elemento químico/fármaco (elii):	Lipitor 20 mg by mouth daily. (5)
CHEM - Elemento químico/fármaco (elii):	Plavix 75 mg. (5)
CHEM - Elemento químico/fármaco (elii):	Lovenox 40 mg subcu daily. (5)
CHEM - Elemento químico/fármaco (elii):	Metoprolol 200 mg by mouth daily. (5)
CHEM - Elemento químico/fármaco (elii):	Protonix 40 mg daily. (5)
CHEM - Química Orgánica/fármaco (orch):	taking three to four nitroglycerin (3)
CHEM - Química Orgánica/fármaco (orch):	Plavix 75 mg. (2)
DEVI - Dispositivo médico (medd):	stented with a Taxus stent (5)
PHEN - Fenomeno o preoceso (phpr):	total occlusion of the LAD (2)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report32095" (Densidad Cualitativa + Cuantitativa)
****

LIVB - Tipo de paciente (podg):	**The patient complains of epigastric distress** (10)
LIVB - Tipo de paciente (podg):	**a black male, (5)**
ANAT - Parte del cuerpo/Organo (bpoc):	**posterior descending branch** (2)
ANAT - Parte del cuerpo/Organo (bpoc):	**the remaining saphenous vein grafts to the left-sided (5)**
ANAT - Tejido (tisu):	**The left internal mammary artery graft to the** (4)
ANAT - Localización o región del cuerpo (blor):	**a large distal obtuse marginal circulation** (2)
ANAT - Localización o región del cuerpo (blor):	**normal vesicular breathing heard all over the lung (3)**
DISO - Localización del trastorno (fndg):	**admitted with a history of chest pain (8)**
DISO - Localización del trastorno (fndg):	**Hypertension (4)**
DISO - Localización del trastorno (fndg):	**Blood pressure 160/97. (3)**
DISO - Localización del trastorno (fndg):	**normal vesicular breathing heard all over the lung (3)**
DISO - Localización del trastorno (fndg):	**a black male, (5)**
DISO - Función Patológica (patf):	**total occlusion of the LAD (11)**
DISO - Función Patológica (patf):	**Proximal circumflex stenosis (9)**
DISO - Anormalidad Anatómica (anab):	**Proximal circumflex stenosis (9)**
DISO - Signo o Sintoma (sosy):	**concerned about his chest pain (6)**
DISO - Signo o Sintoma (sosy):	**the epigastric pain** (2)
DISO - Signo o Sintoma (sosy):	**left-sided chest pain with radiation (5)**
DISO - Signo o Sintoma (sosy):	**normal vesicular breathing heard all over the lung (3)**
DISO - Enfermedad o síndrome (dsyn):	**Coronary artery disease (6)**
DISO - Enfermedad o síndrome (dsyn):	**Hypertension (4)**
PHYS - Atributo Clínico (clna):	**concerned about his chest pain (6)**
PHYS - Función Orgánica (orgf):	**the remaining saphenous vein grafts to the left-sided (4)**
PHYS - Órgano o Tejido (ortf):	**total occlusion of the LAD (11)**
PHYS - Función Orgánica (orgf):	**Blood pressure 160/97. (3)**
PROC - Procedimiento terapeútico (topp):	**status post stenting** (3)
PROC - Procedimiento terapeútico (topp):	**Coronary artery bypass graft about eight years ago** (3)
PROC - Procedimiento terapeútico (topp):	**left-sided chest pain with radiation (5)**
CHEM - Elemento químico/fármaco (elii):	**Lipitor 20 mg by mouth daily.** (6)
CHEM - Elemento químico/fármaco (elii):	**Plavix 75 mg. (7)**
CHEM - Elemento químico/fármaco (elii):	**Lovenox 40 mg subcu daily.** (5)
CHEM - Elemento químico/fármaco (elii):	**Metoprolol 200 mg by mouth daily**. (5)
CHEM - Elemento químico/fármaco (elii):	**Protonix 40 mg daily.** (6)
CHEM - Química Orgánica/fármaco (orch):	**taking three to four nitroglycerin (6)**
CHEM - Química Orgánica/fármaco (orch):	**Plavix 75 mg. (7)**
DEVI - Dispositivo médico (medd):	**stented with a Taxus stent** (5)
PHEN - Fenomeno o preoceso (phpr):	**total occlusion of the LAD (11)**

****

Resumen Final:

Resumen del documento "Report32095":

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **total occlusion of the LAD (11)**
- **Proximal circumflex stenosis (9)**
- **admitted with a history of chest pain (8)**
- **Coronary artery disease (6)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **The patient complains of epigastric distress (10)**
- **a black male, (5)**

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)
- **Plavix 75 mg. (7)**
- **taking three to four nitroglycerin (6)**
- **Lipitor 20 mg by mouth daily. (6)**
- **Protonix 40 mg daily. (6)**
- **Lovenox 40 mg subcu daily. (5)**
- **Metoprolol 200 mg by mouth daily. (5)**

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **concerned about his chest pain (6)**
- **Blood pressure 160/97. (3)**

DISPOSITIVOS MÉDICO, DISPENSADOR DE FÁRMACOS: (DEVI)
- **stented with a Taxus stent (5)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **the remaining saphenous vein grafts to the left-sided (5)**
- **The left internal mammary artery graft to the (4)**
- **normal vesicular breathing heard all over the lung (3)**

PROCEDIMIENTOS TERAPEÚTICOS Y DIAGNÓSTICOS: (PROC)
- **status post stenting** (3)
- **Coronary artery bypass graft about eight years ago (3)**

Report32186

- **Documento médico original con filtrado previo**

RAD

CHEST PAIN

PAIN.

Frontal view of chest obtained. Compared to prior study dated Jul 16 06. Median sternotomy wires are noted. Heart is mildly enlarged. Aorta is unremarkable. Lungs are clear.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

2 0 C0008031 sosy CHEST PAIN
3 0 C2926613 clna CHEST PAIN
4 0 C0030193 sosy PAIN.
5 0 C0817096 blor Frontal view of chest
6 0 C1527391 blor Frontal view of chest
9 0 C1282959 topp Median sternotomy wires
9 0 C0005978 medd Median sternotomy wires
10 0 C1316572 clna noted.
12 0 C0018787 bpoc Heart
13 0 C1281570 bpoc Heart
17 0 C1293134 topp enlarged.
18 0 C0003483 bpoc Aorta
19 0 C1278934 bpoc Aorta
21 0 C0024109 bpoc Lungs

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C2926613 clna CHEST PAIN (1)
C1527391 blor Frontal view of chest (1)
C1293134 topp enlarged. (1)
C1282959 topp Median sternotomy wires (1)
C1281570 bpoc Heart (1)
---------- Corte 2º Cuartil (50 %) ----------
C1278934 bpoc Aorta (1)
C0817096 blor Frontal view of chest (1)
C0030193 sosy PAIN. (1)
C0024109 bpoc Lungs (1)
C0018787 bpoc Heart (1)
C0008031 sosy CHEST PAIN (1)
C0005978 medd Median sternotomy wires (1)

"Report32186" (Densidad Cualitativa)
****

ANAT - Parte del cuerpo/Organo (bpoc):	Heart (1)
ANAT - Localización o región del cuerpo (blor):	Frontal view of chest (1)
PHYS - Atributo Clínico (clna):	CHEST PAIN (1)
PROC - Procedimiento terapeútico (topp):	enlarged (1)
PROC - Procedimiento terapeútico (topp):	Median sternotomy wires (1)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report32186" (Densidad Cualitativa + Cuantitativa)
****
ANAT - Parte del cuerpo/Organo (bpoc): **Heart (2)**
ANAT - Localización o región del cuerpo (blor): **Frontal view of chest (2)**
PHYS - Atributo Clínico (clna): **CHEST PAIN (2)**
PROC - Procedimiento terapeútico (topp): **Median sternotomy wires (2)**
****

Resumen Final:

Resumen del documento "Report32186":

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **CHEST PAIN (2)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Heart (2)**
- **Frontal view of chest (2)**

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **Median sternotomy wires (2)**

Report33471

- **Documento médico original con filtrado previo**

CHEST PAIN

Gastroenterology
Consultation Report

This 70s-year-old male was admitted with a history of abdominal pain and nausea. I was asked to see this patient for abdominal pain. The patient denies any hematemesis. No rectal bleeding.

1. Hypertension.
2. CABG.
3. Congestive heart failure.

1. Prilosec.
2. Aspirin.
3. Lipitor.
4. Plavix.
5. Lovenox.
6. Nitroglycerin patch.

A moderately-built male. VITAL SIGNS: Blood pressure of 130/80, pulse of 100, respirations 20. HEENT: Head normocephalic. No bruits. No cervical lymphadenopathy. Extraocular movements are normal. Conjunctivae pale. Ears normal. Throat: No pigmentation. No thyromegaly. No lymphadenopathy. CHEST: Moving with each respiration bilaterally symmetrically. HEART: Sinus rhythm. ABDOMEN: No palpable masses. RECTAL Guaiac negative. CNS: Within normal limits.

Abdominal pain for GI work-up.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0017163 bmod Gastroenterology Consultation Report
6 0 C0024554 fndg This 70s-year-old male
7 0 C0086582 popg This 70s-year-old male
16 0 C0000737 sosy admitted with a history of abdominal pain
19 0 C0262926 fndg admitted with a history of abdominal pain
21 0 C2004062 fndg admitted with a history of abdominal pain
22 0 C0027497 sosy nausea.
23 0 C1963179 fndg nausea.
26 0 C0042789 orgf see
28 0 C0030705 podg this patient for abdominal pain.
28 0 C0000737 sosy this patient for abdominal pain.
29 0 C0030705 podg The patient
39 0 C0020538 dsyn Hypertension.
40 0 C1963138 fndg Hypertension.
44 0 C0010055 topp CABG.
48 0 C0018802 dsyn Congestive heart failure.
52 0 C0700777 orch Prilosec.
56 0 C0004057 orch Aspirin.
60 0 C0593906 orch Lipitor.
62 0 C0633084 orch Plavix.
66 0 C0699222 clnd Nitroglycerin patch.
69 0 C0024554 fndg A moderately-built male.
70 0 C0086582 popg A moderately-built male.
85 0 C0518766 clna VITAL SIGNS
86 0 C0005823 orgf Blood pressure of 130/80,
87 0 C1271104 fndg Blood pressure of 130/80,
88 0 C1272641 fndg Blood pressure of 130/80,
90 0 C0232117 clna pulse of 100,
92 0 C1947910 phpr pulse of 100,
94 0 C1512338 blor HEENT
95 0 C0018670 blor Head

96 0 C1281590 bpoc Head
97 0 C1855201 fndg normocephalic.
100 0 C0026649 orgf Extraocular movements
104 0 C0009758 bpoc Conjunctivae pale.
104 0 C0030232 sosy Conjunctivae pale.
106 0 C0678215 fndg Conjunctivae pale.
108 0 C0013443 bpoc Ears
109 0 C0521421 bpoc Ears
114 0 C0031354 bpoc Throat
115 0 C0230069 blor Throat
116 0 C1280698 bpoc Throat
120 0 C0817096 blor CHEST
121 0 C1527391 blor CHEST
122 0 C0439837 fndg Moving with each respiration bilaterally symmetrically.
123 0 C0018787 bpoc HEART
124 0 C1281570 bpoc HEART
125 0 C0232201 fndg Sinus rhythm.
126 0 C0000726 blor ABDOMEN
127 0 C0230168 blor ABDOMEN
128 0 C1281594 bpoc ABDOMEN
132 0 C1513916 fndg Guaiac negative.
136 0 C0000726 blor Abdominal pain for GI work-up.
136 0 C0687713 sosy Abdominal pain for GI work-up.
136 0 C0750430 diap Abdominal pain for GI work-up.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C0086582 popg A moderately-built male. (2)
C0030705 podg this patient for abdominal pain. (2)
C0024554 fndg A moderately-built male. (2)
C0000737 sosy admitted with a history of abdominal pain (2)
C0000726 blor Abdominal pain for GI work-up. (2)
C2926613 clna CHEST PAIN (1)
C2004062 fndg admitted with a history of abdominal pain (1)
C1963179 fndg nausea. (1)
C1963138 fndg Hypertension. (1)
C1947910 phpr pulse of 100, (1)
C1855201 fndg normocephalic. (1)
C1527391 blor CHEST (1)
C1513916 fndg Guaiac negative. (1)
C1512338 blor HEENT (1)
C1281594 bpoc ABDOMEN (1)
C1281590 bpoc Head (1)
C1281570 bpoc HEART (1)
C1280698 bpoc Throat (1)
C1272641 fndg Blood pressure of 130/80, (1)
C1271104 fndg Blood pressure of 130/80, (1)
C0817096 blor CHEST (1)
C0750430 diap Abdominal pain for GI work-up. (1)
C0700777 orch Prilosec. (1)
---------- Corte 2º Cuartil (50 %) ----------
C0699222 clnd Nitroglycerin patch. (1)
C0687713 sosy Abdominal pain for GI work-up. (1)
C0678215 fndg Conjunctivae pale. (1)
C0633084 orch Plavix. (1)
C0593906 orch Lipitor. (1)
C0521421 bpoc Ears (1)
C0518766 clna VITAL SIGNS (1)
C0439837 fndg Moving with each respiration bilaterally symmetrically. (1)
C0262926 fndg admitted with a history of abdominal pain (1)
C0232201 fndg Sinus rhythm. (1)
C0232117 clna pulse of 100, (1)
C0230168 blor ABDOMEN (1)
C0230069 blor Throat (1)
C0042789 orgf see (1)
C0031354 bpoc Throat (1)
C0030232 sosy Conjunctivae pale. (1)
C0027497 sosy nausea. (1)
C0026649 orgf Extraocular movements (1)
C0020538 dsyn Hypertension. (1)
C0018802 dsyn Congestive heart failure. (1)
C0018787 bpoc HEART (1)
C0018670 blor Head (1)

C0017163 bmod Gastroenterology Consultation Report (1)
C0013443 bpoc Ears (1)
C0010055 topp CABG. (1)
C0009758 bpoc Conjunctivae pale. (1)
C0008031 sosy CHEST PAIN (1)
C0005823 orgf Blood pressure of 130/80, (1)

"Report33471" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	this patient for abdominal pain (2)
LIVB - Grupo de población (popg):	A moderately-built male (2)
ANAT - Localización o región del cuerpo (blor):	Abdominal pain for GI work-up (2)
DISO - Localización del trastorno (fndg):	A moderately-built male (2)
DISO - Signo o Sintoma (sosy):	admitted with a history of abdominal pain (2)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report33471" (Densidad Cualitativa + Cuantitativa)
****

LIVB - Tipo de paciente (podg):	this patient for abdominal pain (2)
LIVB - Grupo de población (popg):	**A moderately-built male (4)**
ANAT - Parte del cuerpo/Organo (bpoc):	**Conjunctivae pale. (3)**
ANAT - Localización o región del cuerpo (blor):	Abdominal pain for GI work-up (4)
ANAT - Localización o región del cuerpo (blor):	**Throat (3)**
DISO - Localización del trastorno (fndg):	**A moderately-built male (4)**
DISO - Localización del trastorno (fndg):	**Blood pressure of 130/80, (3)**
DISO - Localización del trastorno (fndg):	**Conjunctivae pale. (3)**
DISO - Localización del trastorno (fndg):	**admitted with a history of abdominal pain (4)**
DISO - Signo o Sintoma (sosy):	admitted with a history of abdominal pain (4)
DISO - Signo o Sintoma (sosy):	**Abdominal pain for GI work-up. (4)**
DISO - Signo o Sintoma (sosy):	**Conjunctivae pale. (3)**
PHYS - Función Orgánica (orgf):	**Blood pressure of 130/80, (3)**
PROC - Procedimiento de diagnóstico (diap):	**Abdominal pain for GI work-up. (4)**

****

Resumen Final:

Resumen del documento "Report33471":

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **admitted with a history of abdominal pain (4)**
- **Conjunctivae pale. (3)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **A moderately-built male (4)**
- this patient for abdominal pain (2)

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **Abdominal pain for GI work-up. (4)**

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **Blood pressure of 130/80, (3)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Throat (3)**

Report33511

- **Documento médico original con filtrado previo**

CHEST PAIN

Gastroenterology
Operative Report

Abdominal pain.
Cetacaine spray, Versed 1 mg, and fentanyl 100.
Pan endoscopy.
Following a Cetacaine spray, the scope was introduced to esophagus. The esophagus was normal in caliber. The scope was advanced to the stomach. The stomach was examined with anterior, posterior, and lesser and greater curvatures examined, which are normal. Duodenum is normal. The scope was
withdrawn, terminating the procedure. Normal endoscopy.

Proceed with colonoscopy.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C0017163 bmod Gastroenterology Operative Report
6 0 C0000737 sosy Abdominal pain.
7 0 C0055141 orch Cetacaine spray,
9 0 C0042553 orch Versed 1 mg,
10 0 C2346927 elii Versed 1 mg,
11 0 C0015846 orch fentanyl 100.
12 0 C0014245 diap Pan endoscopy.
13 0 C1552424 hcro Pan endoscopy.
15 0 C0055141 orch a Cetacaine spray,
18 0 C0014876 bpoc introduced to esophagus.
20 0 C1278919 bpoc introduced to esophagus.
21 0 C0014876 bpoc The esophagus
23 0 C1278919 bpoc The esophagus
28 0 C0038351 bpoc advanced to the stomach.
29 0 C1278920 bpoc advanced to the stomach.
30 0 C0038351 bpoc The stomach
31 0 C1278920 bpoc The stomach
32 0 C0332128 fndg examined with anterior, posterior,
34 0 C0227223 blor greater curvatures
35 0 C0332128 fndg examined
39 0 C0013303 bpoc Duodenum
40 0 C1278921 bpoc Duodenum
45 0 C0424092 fndg withdrawn
48 0 C2825032 mobd withdrawn
51 0 C0184661 diap the procedure.
52 0 C1948041 topp the procedure.
54 0 C0014245 diap Normal endoscopy.
55 0 C1552424 hcro Normal endoscopy.
60 0 C0009378 diap Proceed with colonoscopy.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C1552424 hcro Normal endoscopy. (2)
C1278920 bpoc advanced to the stomach. (2)
C1278919 bpoc introduced to esophagus. (2)
C0332128 fndg examined with anterior, posterior, (2)
C0055141 orch a Cetacaine spray, (2)
C0038351 bpoc advanced to the stomach. (2)
C0014876 bpoc introduced to esophagus. (2)
---------- Corte 2º Cuartil (50 %) ----------
C0014245 diap Normal endoscopy. (2)
--- resto (1's) ---

"Report33511" (Densidad Cualitativa)
****

ANAT - Parte del cuerpo/Organo (bpoc):	advanced to the stomach (2)
ANAT - Parte del cuerpo/Organo (bpoc):	introduced to esophagus (2)
DISO - Localización del trastorno (fndg):	examined with anterior, posterior (2)
PROC - Procedimiento de diagnóstico (diap):	Normal endoscopy (2)
CHEM - Química Orgánica/fármaco (orch):	a Cetacaine spray (2)
ORGA - Organización cuidados de la salud (hcro):	Normal endoscopy (2)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report33511" (Densidad Cualitativa + Cuantitativa)
****

ANAT - Parte del cuerpo/Organo (bpoc):	**advanced to the stomach (4)**
ANAT - Parte del cuerpo/Organo (bpoc):	**introduced to esophagus (4)**
DISO - Localización del trastorno (fndg):	examined with anterior, posterior (2)
PROC - Procedimiento de diagnóstico (diap):	**Normal endoscopy. (4)**
CHEM - Química Orgánica/fármaco (orch):	a Cetacaine spray (2)
ORGA - Organización cuidados de la salud (hcro):	**Normal endoscopy. (4)**

****

Resumen Final:

Resumen del documento "Report33511":

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)

- **advanced to the stomach (4)**
- **introduced to esophagus (4)**

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)

- **Normal endoscopy. (4)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)

- examined with anterior, posterior (2)

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)

- a Cetacaine spray (2)

Report33721

- **Documento médico original con filtrado previo**

CHEST PAIN

1. Chest pain.
2. Coronary artery disease status post coronary artery bypass graft surgery.
3. Hypertension.
4. Status post stent of the left anterior descending and circumflex.
5. Hyperlipidemia.
6. Bradycardia.
7. Hemorrhoids.
8. Hearing loss.

1. Ocean nasal spray 2 sprays q.4 hours as needed.
2. Senokot 2 tablets at bedtime.
3. Pentoxifylline 400 mg b.i.d.
4. Transderm nitro 0.4 mg daily, remove bedtime.
5. Toprol-XL 150 mg daily.
6. Hemorrhoidal topical cream per rectum as directed for one week.
7. Zetia 10 mg at bedtime.
8. Clopidogrel 75 mg daily.
9. Lipitor 20 mg daily.
10. Enteric-coated aspirin 325 mg daily.
11. Norvasc 5 mg daily.
12. Tylenol caplets 650 mg q.4 hours as needed for pain.

This is in 70s]-year-old male admitted because of chest and epigastric pain. The patient claims he has been having left-sided chest pain with radiation to the left arm. He took nitroglycerin over 2 days without much relief. He denied any vomiting, nausea, or diaphoresis. He had a prior cardiac catheterization on Jun 16 2006, which showed total occlusion of the LAD and multiple sequential 70% to 80% stenosis of the proximal mid vessel. The right coronary provided an acute marginal collateral to the distal LAD and a patent saphenous vein graft to the TDA and posterolateral branch. The circumflex stenosis was directly stented with Taxus stent as well as the LAD. The patient denied any history of congestive heart failure but he also has a history of CVA, hypertension, coronary artery disease, bypass graft surgery, arthritis, hemorrhoidectomy, and stenting as outlined above. He denied any alcohol and tobacco products. Chest x-ray on arrival showed no active disease. The patient was admitted with chest pain and to rule out angina versus myocardial infarction. He also had epigastric pain.
He was admitted for both cardiology and GI evaluations.

Initial electrolytes were normal. BUN and creatinine were 19 and 1.7. Hemoglobin and hematocrit 13.5 and 40.2, white count 4700. INR 1.0, BNP level was 191. Enzymes were negative for myocardial infarction. BUN and creatinine at discharge were 29 and 2.0. Cardiogram, first degree AV block with bradycardic response. Left ventricular hypertrophy and minor nonspecific ST and T-wave changes. Exercise stress test 1+ to 2+ mitral and tricuspid regurgitation. Dilated left ventricular ejection fraction of 35% to 40%. Overall picture was suggestive of ischemic cardiomyopathy.
The patient was admitted to Monitor Unit because of chest pain. He was placed on a prudent heart diet as well as enteric-coated aspirin, nitro paste, and Toprol. Serial cardiograms and enzymes were followed and a cardiac evaluation was requested to Dr. who did review prior cardiac catheterization records which he had, stenting of the LAD and circumflex as well as echocardiogram. There was no myocardial infarction but Dr. felt that there was a possibility of angina, and he was continued on cardiac regimen and eventually a stress test was done which was negative for ischemia.

The patient then had an evaluation from Dr. because of his epigastric pain and he did recommend a GI workup. The patient denied any hematemesis or rectal bleed. Endoscopy was done and was negative and colonoscopy was negative except for hemorrhoids. He was recommended hemorrhoidal cream along with a high-fiber diet. The patient also had suffered from hearing loss during this admission and ENT evaluation was requested to Dr. . He had complained of some right face and ear pain in addition to hearing loss. He was asked to return to Dr.'s office if there was no resolution of the problem postdischarge. At one point, the patient was also bradycardic and Dr. followed this closely and medications were adjusted according to heart rate. By the time of discharge, the patient denied any abdominal pain, any nausea or vomiting. He was very anxious to go home. His blood pressure was stable 153/91 which will continue to be followed, heart rate was 74. From GI standpoint, Dr. felt the patient was stable. He had no further chest pains and stress test was negative. Blood pressure was controlled and he will continue on a cardiac regimen at this time and follow with Dr. as an outpatient for his hearing loss. The patient will see Dr. within 2 weeks and Dr. within 2 weeks. No heavy lifting, bending, or straining. He will have a CBC and differential with a basic metabolic panel in one week. He will have a CAT scan of the abdomen as an outpatient.

He was encouraged oral fluids because of renal insufficiency. If he has any further chest pain or shortness of breath or rectal bleed, he is to return to the Emergency Department. He understood all instructions.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
6 0 C0008031 sosy Chest pain.
7 0 C2926613 clna Chest pain.
11 0 C0010054 dsyn Coronary artery disease status post coronary artery bypass graft surgery.
11 0 C0010055 topp Coronary artery disease status post coronary artery bypass graft surgery.
12 0 C0010068 dsyn Coronary artery disease status post coronary artery bypass graft surgery.
13 0 C1956346 dsyn Coronary artery disease status post coronary artery bypass graft surgery.
17 0 C0020538 dsyn Hypertension.
18 0 C1963138 fndg Hypertension.
20 0 C0038257 medd Status post stent
21 0 C1705817 medd Status post stent
27 0 C0020473 dsyn Hyperlipidemia.
28 0 C0428977 patf Bradycardia.
29 0 C0019112 acab Hemorrhoids.
30 0 C1963134 fndg Hemorrhoids.
31 0 C0011053 dsyn Hearing loss.
32 0 C0018772 fndg Hearing loss.
33 0 C1384666 fndg Hearing loss.
43 0 C1616946 clnd Senokot 2 tablets at bedtime.
47 0 C0988533 clnd Pentoxifylline 400 mg b.i.d.
47 0 C0152277 topp Pentoxifylline 400 mg b.i.d.
49 0 C0699246 orch Transderm nitro 0.4 mg daily,
50 0 C2346927 elii Transderm nitro 0.4 mg daily,
57 0 C0723783 orch Toprol-XL 150 mg daily.
58 0 C2346927 elii Toprol-XL 150 mg daily.
59 0 C0034896 bpoc Hemorrhoidal topical cream per rectum
65 0 C1170736 orch Zetia 10 mg at bedtime.
66 0 C2346927 elii Zetia 10 mg at bedtime.
67 0 C1124675 clnd Clopidogrel 75 mg daily.
68 0 C0593906 orch Lipitor 20 mg daily.
69 0 C2346927 elii Lipitor 20 mg daily.
70 0 C0992015 clnd Enteric-coated aspirin 325 mg daily.
71 0 C0162712 orch Norvasc 5 mg daily.
72 0 C2346927 elii Norvasc 5 mg daily.
73 0 C0982960 clnd Tylenol caplets 650 mg q.4 hours
74 0 C0030193 sosy needed for pain.
76 0 C0024554 fndg]-year-old male
77 0 C0086582 popg]-year-old male
86 0 C0817096 blor admitted because of chest
87 0 C1527391 blor admitted because of chest
88 0 C0232493 sosy epigastric pain.
89 0 C0030705 podg The patient
90 0 C0541828 sosy left-sided chest pain with radiation
90 0 C0034519 npop left-sided chest pain with radiation
92 0 C0851346 npop left-sided chest pain with radiation
93 0 C1522449 topp left-sided chest pain with radiation
94 0 C1524020 topp left-sided chest pain with radiation
96 0 C0230347 bpoc to the left arm.
97 0 C1268256 bpoc to the left arm.
100 0 C0017887 orch nitroglycerin over 2 days
101 0 C0885378 orch nitroglycerin over 2 days
112 0 C0018795 diap a prior cardiac catheterization on Jun 16 2006,
115 0 C0011382 ortf total occlusion of the LAD
116 0 C0028778 patf total occlusion of the LAD
118 0 C1110554 patf total occlusion of the LAD
119 0 C1882137 phpr total occlusion of the LAD
127 0 C0009814 patf multiple sequential 70% to 80% stenosis of the proximal mid vessel.
127 0 C0005847 bpoc multiple sequential 70% to 80% stenosis of the proximal mid vessel.
129 0 C0947637 patf multiple sequential 70% to 80% stenosis of the proximal mid vessel.
130 0 C1261287 anab multiple sequential 70% to 80% stenosis of the proximal mid vessel.
137 0 C0018787 bpoc The right coronary
139 0 C1305794 blor an acute marginal collateral to the distal LAD
139 0 C1275670 bpoc an acute marginal collateral to the distal LAD
141 0 C0729538 bpoc a patent saphenous vein graft to the TDA
143 0 C1253959 bpoc posterolateral branch.
148 0 C0009814 patf The circumflex stenosis

150 0 C0947637 patf The circumflex stenosis
151 0 C1261287 anab The circumflex stenosis
154 0 C0038257 medd stented with Taxus stent
155 0 C0030705 podg The patient
160 0 C0038454 dsyn a history of CVA,
163 0 C0262926 fndg a history of CVA,
165 0 C2004062 fndg a history of CVA,
166 0 C0020538 dsyn hypertension,
167 0 C1963138 fndg hypertension,
168 0 C0010054 dsyn coronary artery disease,
169 0 C0010068 dsyn coronary artery disease,
170 0 C1956346 dsyn coronary artery disease,
171 0 C1536078 topp bypass graft surgery,
172 0 C0332835 tisu bypass graft surgery,
173 0 C1961139 topp bypass graft surgery,
174 0 C0185098 topp bypass graft surgery,
174 0 C0038894 bmod bypass graft surgery,
176 0 C0543467 diap bypass graft surgery,
177 0 C0003864 dsyn arthritis,
178 0 C0019108 topp hemorrhoidectomy,
179 0 C0038257 medd stenting as outlined above.
180 0 C2348535 topp stenting as outlined above.
188 0 C0039985 diap Chest x-ray on arrival
190 0 C1114557 clna Chest x-ray on arrival
194 0 C0030705 podg The patient
195 0 C0008031 sosy admitted with chest pain
196 0 C2926613 clna admitted with chest pain
204 0 C0232493 sosy epigastric pain.
206 0 C0007189 bmod cardiology
216 0 C0005845 lbpr BUN
219 0 C1561535 fndg creatinine
224 0 C0018935 lbpr hematocrit 13.5
225 0 C0518014 lbtr hematocrit 13.5
227 0 C0007457 popg white count 4700.
229 0 C0043157 popg white count 4700.
233 0 C0525032 lbpr INR 1.0,
234 0 C1704538 topp INR 1.0,
235 0 C1095989 lbpr BNP level
257 0 C0005845 lbpr BUN
259 0 C0012621 bdsu creatinine at discharge
261 0 C2926602 bdsu creatinine at discharge
262 0 C1561535 fndg creatinine at discharge
268 0 C0849666 diap Cardiogram,
269 0 C0085614 dsyn first degree AV block with bradycardic response.
269 0 C0428977 patf first degree AV block with bradycardic response.
270 0 C1704632 fndg first degree AV block with bradycardic response.
273 0 C0149721 dsyn Left ventricular hypertrophy
274 0 C0232306 lbtr Left ventricular hypertrophy
275 0 C0026193 popg minor nonspecific ST
279 0 C0429103 fndg T-wave changes.
281 0 C0015260 diap Exercise stress test 1+
282 0 C0430120 diap Exercise stress test 1+
286 0 C0040961 dsyn tricuspid regurgitation.
288 0 C0428772 lbtr left ventricular ejection fraction of 35%
289 0 C0488728 clna left ventricular ejection fraction of 35%
295 0 C0349782 dsyn suggestive of ischemic cardiomyopathy.
296 0 C0030705 podg The patient
301 0 C0181904 medd Monitor
302 0 C0596972 medd Monitor
305 0 C0024902 sosy Unit because of chest pain.
307 0 C1704434 medd Unit because of chest pain.
310 0 C0018787 bpoc placed on a prudent heart diet
311 0 C1281570 bpoc placed on a prudent heart diet
316 0 C0718690 orch enteric-coated aspirin,
320 0 C0849666 diap Serial cardiograms
324 0 C0018787 bpoc a cardiac evaluation
335 0 C0018795 diap prior cardiac catheterization records
341 0 C0038257 medd stenting of the LAD
342 0 C2348535 topp stenting of the LAD
344 0 C0013516 diap echocardiogram.
353 0 C0002962 sosy a possibility of angina,
354 0 C0018787 bpoc continued on cardiac regimen
354 0 C0040808 topp continued on cardiac regimen
358 0 C0015260 diap a stress test
363 0 C0030705 podg The patient then
373 0 C0232493 sosy because of his epigastric pain

375 0 C0750430 diap a GI workup.
376 0 C0030705 podg The patient
382 0 C0014245 diap Endoscopy
383 0 C1552424 hcro Endoscopy
385 0 C1513916 fndg negative
388 0 C0009378 diap colonoscopy
390 0 C0019112 acab negative except for hemorrhoids.
391 0 C1963134 fndg negative except for hemorrhoids.
392 0 C1513916 fndg negative except for hemorrhoids.
399 0 C0034896 bpoc hemorrhoidal cream along with a high-fiber diet.
399 0 C0301568 topp hemorrhoidal cream along with a high-fiber diet.
405 0 C0030705 podg The patient also
406 0 C0683278 mobd suffered from hearing loss
406 0 C0011053 dsyn suffered from hearing loss
407 0 C0018772 fndg suffered from hearing loss
408 0 C1384666 fndg suffered from hearing loss
417 0 C0015450 blor complained of some right face
418 0 C1281591 bpoc complained of some right face
419 0 C0013456 sosy ear pain in addition to hearing loss.
419 0 C0011053 dsyn ear pain in addition to hearing loss.
420 0 C0018772 fndg ear pain in addition to hearing loss.
421 0 C1384666 fndg ear pain in addition to hearing loss.
431 0 C0030705 podg the patient
432 0 C0428977 patf also bradycardic
438 0 C0802604 clna medications
439 0 C2598133 clna medications
442 0 C0012621 bdsu of discharge,
444 0 C2926602 bdsu of discharge,
445 0 C0030705 podg the patient
457 0 C0005823 orgf His blood pressure
458 0 C1271104 fndg His blood pressure
459 0 C1272641 fndg His blood pressure
473 0 C0030705 podg the patient
482 0 C0005823 orgf Blood pressure
483 0 C1271104 fndg Blood pressure
484 0 C1272641 fndg Blood pressure
486 0 C0018787 bpoc continue on a cardiac regimen
486 0 C0040808 topp continue on a cardiac regimen
497 0 C0029921 podg as an outpatient
500 0 C0011053 dsyn for his hearing loss.
501 0 C0018772 fndg for his hearing loss.
502 0 C1384666 fndg for his hearing loss.
503 0 C0030705 podg The patient
504 0 C0042789 orgf see
519 0 C0009555 lbpr a CBC
521 0 C2237045 lbpr differential with a basic metabolic panel
525 0 C0412620 diap a CAT scan of the abdomen
526 0 C0029921 podg as an outpatient.
529 0 C0005889 bdsu oral fluids because of renal insufficiency.
529 0 C1565489 dsyn oral fluids because of renal insufficiency.
531 0 C1521806 tisu oral fluids because of renal insufficiency.
532 0 C0008031 sosy any further chest pain
533 0 C2926613 clna any further chest pain
534 0 C0013404 sosy shortness of breath
535 0 C2707305 clna shortness of breath
536 0 C0267596 patf rectal bleed,
538 0 C0562508 hcro return to the Emergency Department.
542 0 C3263700 clna all instructions.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C0030705 podg The patient also (11)
C2346927 elii Transderm nitro 0.4 mg daily, (5)
C2346927 elii Toprol-XL 150 mg daily.
C2346927 elii Zetia 10 mg at bedtime.
C2346927 elii Lipitor 20 mg daily.
C2346927 elii Norvasc 5 mg daily.
C0018787 bpoc placed on a prudent heart diet (5)
C2926613 clna admitted with chest pain (4)
C1384666 fndg ear pain in addition to hearing loss. (4)
C0038257 medd stenting as outlined above. (4)
C0018772 fndg ear pain in addition to hearing loss. (4)
C0011053 dsyn ear pain in addition to hearing loss. (4)
C0008031 sosy admitted with chest pain (4)

C0428977 patf first degree AV block with bradycardic response. (3)
C0232493 sosy because of his epigastric pain (3)
C2926602 bdsu creatinine at discharge (2)
C2348535 topp stenting as outlined above. (2)
C1963138 fndg Hypertension. (2)
C1963134 fndg negative except for hemorrhoids. (2)
C1956346 dsyn Coronary artery disease status post coronary artery bypass (2)
C1561535 fndg creatinine at discharge (2)
C1513916 fndg negative except for hemorrhoids. (2)
C1272641 fndg His blood pressure (2)
C1271104 fndg His blood pressure (2)
C1261287 anab multiple sequential 70% to 80% stenosis of the (2)
C0947637 patf multiple sequential 70% to 80% stenosis of the (2)
C0849666 diap Serial cardiograms (2)
C0040808 topp continue on a cardiac regimen (2)
C0034896 bpoc hemorrhoidal cream along with a high-fiber diet. (2)
C0029921 podg as an outpatient (2)
C0020538 dsyn Hypertension. (2)
C0019112 acab negative except for hemorrhoids. (2)
C0018795 diap a prior cardiac catheterization on Jun 16 2006, (2)
C0015260 diap Exercise stress test 1+ (2)
C0012621 bdsu creatinine at discharge (2)
C0010068 dsyn Coronary artery disease status post coronary artery bypass (2)
C0010054 dsyn Coronary artery disease status post coronary artery bypass (2)
C0009814 patf multiple sequential 70% to 80% stenosis of the (2)
C0005845 lbpr BUN (2)
---------- Corte 2º Cuartil (50 %) ----------
C0005823 orgf His blood pressure (2)
--- resto (1's) ---

"Report33721" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	The patient also (11)
LIVB - Tipo de paciente (podg):	as an outpatient (2)
ANAT - Parte del cuerpo/Organo (bpoc):	placed on a prudent heart diet (5)
ANAT - Parte del cuerpo/Organo (bpoc):	hemorrhoidal cream along with a high-fiber diet (2)
ANAT - Substancia del cuerpo (bdsu):	creatinine at discharge (2)
DISO - Función Patológica (patf):	multiple sequential 70% to 80% stenosis of the (2)
DISO - Anormalidad Anatómica (anab):	multiple sequential 70% to 80% stenosis of the (2)
DISO - Localización del trastorno (fndg):	ear pain in addition to hearing loss (4)
DISO - Localización del trastorno (fndg):	Hypertension (2)
DISO - Localización del trastorno (fndg):	negative except for hemorrhoids (2)
DISO - Localización del trastorno (fndg):	creatinine at discharge (2)
DISO - Localización del trastorno (fndg):	His blood pressure (2)
DISO - Función Patológica (patf):	first degree AV block with bradycardic response (3)
DISO - Signo o Sintoma (sosy):	admitted with chest pain (4)
DISO - Signo o Sintoma (sosy):	because of his epigastric pain (3)
DISO - Enfermedad o síndrome (dsyn):	ear pain in addition to hearing loss (4)
DISO - Enfermedad o síndrome (dsyn):	Coronary artery disease status post coronary artery bypass (2)
DISO - Enfermedad o síndrome (dsyn):	Hypertension (2)
DISO - Anormalidad adquirida (acab):	negative except for hemorrhoids (2)
PHYS - Atributo Clínico (clna):	admitted with chest pain (4)
PHYS - Función Orgánica (orgf):	His blood pressure(2)
PROC - Procedimiento terapeútico (topp):	stenting as outlined above (2)
PROC - Procedimiento terapeútico (topp):	continue on a cardiac regimen (2)
PROC - Procedimiento de diagnóstico (diap):	Serial cardiograms (2)
PROC - Procedimiento de diagnóstico (diap):	a prior cardiac catheterization on Jun 16 2006 (2)
PROC - Procedimiento de diagnóstico (diap):	Exercise stress test 1+ (2)
CHEM - Elemento químico/fármaco (elii):	Transderm nitro 0.4 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Toprol-XL 150 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Zetia 10 mg at bedtime (5)
CHEM - Elemento químico/fármaco (elii):	Lipitor 20 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Norvasc 5 mg daily (5)
DEVI - Dispositivo médico (medd):	stenting as outlined above (4)
PHEN - Resultados de laboratorio o test (lbtr):	BUN (2) - (blood test - Blood urea nitrogen)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report33721" (Densidad Cualitativa + Cuantitativa)
****

DISO - Localización del trastorno (fndg):	**ear pain in addition to hearing loss. (12)**
DISO - Función Patológica (patf):	**multiple sequential 70% to 80% stenosis of the (6)**
DISO - Localización del trastorno (fndg):	**negative except for hemorrhoids. (6)**
DISO - Enfermedad o síndrome (dsyn):	**Coronary artery disease status post coronary artery bypass (6)**
LIVB - Tipo de paciente (podg):	The patient also (11)
PHYS - Atributo Clínico (clna):	**admitted with chest pain (8)**
PHYS - Función Orgánica (orgf):	**His blood pressure (6)**
ANAT - Substancia del cuerpo (bdsu):	**creatinine at discharge (6)**
ANAT - Parte del cuerpo/Organo (bpoc):	placed on a prudent heart diet (5)
PROC - Procedimiento terapeútico (topp):	**stenting as outlined above. (6)**
CHEM - Elemento químico/fármaco (elii):	Transderm nitro 0.4 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Toprol-XL 150 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Zetia 10 mg at bedtime (5)
CHEM - Elemento químico/fármaco (elii):	Lipitor 20 mg daily (5)
CHEM - Elemento químico/fármaco (elii):	Norvasc 5 mg daily (5)
PHEN - Resultados de laboratorio o test (lbtr):	BUN (2) - (blood test - Blood urea nitrogen)

****

Resumen Final:

Resumen del documento "Report33721":

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **ear pain in addition to hearing loss. (12)**
- **multiple sequential 70% to 80% stenosis of the (6)**
- **negative except for hemorrhoids. (6)**
- **Coronary artery disease status post coronary artery bypass (6)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- The patient also (11)

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **admitted with chest pain (8)**
- **His blood pressure (6)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **creatinine at discharge (6)**
- placed on a prudent heart diet (5)

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **stenting as outlined above. (6)**

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)
- Transderm nitro 0.4 mg daily (5)
- Toprol-XL 150 mg daily (5)
- Zetia 10 mg at bedtime (5)
- Lipitor 20 mg daily (5)
- Norvasc 5 mg daily (5)

RESULTADOS DE LABORATORIO O TEST: (PHEN)
- BUN (2)

Report33955

- **Documento médico original con filtrado previo**

CHEST PAIN

ABDOMINAL ULTRASOUND:
PAIN.
Ultrasound scan of the liver shows normal size, configuration and echogenicity without parenchymal lesion or biliary dilatation. The gallbladder is normal. Wall thickness is normal. Common bile duct is normal measuring 3.6mm in diameter. Head, body and tail of the pancreas are normal. Negative Murphy size. No evidence of ascites. Incidental finding is a renal cyst in the midpole of the right kidney measuring 2.3cm in diameter.

NO ACUTE ABNORMALITY DEMONSTRATED. NO CHANGE SINCE THE EXAMINATION INCIDENTAL FINDING IS A 2.3cm CYST IN THE MIDPOLE OF THE RIGHT KIDNEY.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0008031 sosy CHEST PAIN
2 0 C2926613 clna CHEST PAIN
3 0 C2348813 diap ABDOMINAL ULTRASOUND
4 0 C0030193 sosy PAIN.
5 0 C0041618 diap Ultrasound scan of the liver
5 0 C0023884 bpoc Ultrasound scan of the liver
6 0 C1278929 bpoc Ultrasound scan of the liver
9 0 C0034606 diap Ultrasound scan of the liver
9 0 C0412534 diap Ultrasound scan of the liver
10 0 C0203758 diap Ultrasound scan of the liver
11 0 C1543745 clna Ultrasound scan of the liver
12 0 C0041621 npop Ultrasound scan of the liver
16 0 C0441633 diap Ultrasound scan of the liver
17 0 C1456803 npop Ultrasound scan of the liver
19 0 C1875843 medd Ultrasound scan of the liver
21 0 C2826292 fndg Ultrasound scan of the liver
29 0 C0016976 bpoc The gallbladder
30 0 C1269001 bpoc The gallbladder
38 0 C0009437 bpoc Common bile duct
45 0 C0018670 blor Head,
46 0 C1281590 bpoc Head,
50 0 C0227590 bpoc tail of the pancreas
55 0 C1513916 fndg Negative Murphy size.
59 0 C0743997 fndg Incidental finding
60 0 C0022679 dsyn a renal cyst in the midpole of the right kidney
65 0 C0442739 fndg NO CHANGE

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C2926613 clna CHEST PAIN (1)
C2826292 fndg Ultrasound scan of the liver (1)
C2348813 diap ABDOMINAL ULTRASOUND (1)
C1875843 medd Ultrasound scan of the liver (1)
C1543745 clna Ultrasound scan of the liver (1)
C1513916 fndg Negative Murphy size. (1)
C1456803 npop Ultrasound scan of the liver (1)
C1281590 bpoc Head, (1)
C1278929 bpoc Ultrasound scan of the liver (1)
C1269001 bpoc The gallbladder (1)
C0743997 fndg Incidental finding (1)
C0442739 fndg NO CHANGE (1)
---------- Corte 2º Cuartil (50 %) ----------
C0441633 diap Ultrasound scan of the liver (1)
C0412534 diap Ultrasound scan of the liver (1)
C0227590 bpoc tail of the pancreas (1)
C0203758 diap Ultrasound scan of the liver (1)
C0041621 npop Ultrasound scan of the liver (1)
C0041618 diap Ultrasound scan of the liver (1)
C0034606 diap Ultrasound scan of the liver (1)

C0030193 sosy PAIN. (1)
C0023884 bpoc Ultrasound scan of the liver (1)
C0022679 dsyn a renal cyst in the midpole of the(1)
C0018670 blor Head, (1)
C0016976 bpoc The gallbladder (1)
C0009437 bpoc Common bile duct (1)

"Report33955" (Densidad Cualitativa)
****

ANAT - Parte del cuerpo/Organo (bpoc):	Head (1)
ANAT - Parte del cuerpo/Organo (bpoc):	Ultrasound scan of the liver (1)
ANAT - Parte del cuerpo/Organo (bpoc):	The gallbladder (1)
DISO - Localización del trastorno (fndg):	Ultrasound scan of the liver (1)
DISO - Localización del trastorno (fndg):	Negative Murphy size (1)
DISO - Localización del trastorno (fndg):	Incidental finding (1)
DISO - Localización del trastorno (fndg):	NO CHANGE (1)
PHYS - Atributo Clínico (clna):	CHEST PAIN (1)
PROC - Procedimiento de diagnóstico (diap):	ABDOMINAL ULTRASOUND (1)
DEVI - Dispositivo médico (medd):	Ultrasound scan of the liver (1)
PHEN - Proceso o Fenomeno natural (npop):	Ultrasound scan of the liver (1)

****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report33955" (Densidad Cualitativa + Cuantitativa)
****

ANAT - Parte del cuerpo/Organo (bpoc):	**Head, (2)**
ANAT - Parte del cuerpo/Organo (bpoc):	**Ultrasound scan of the liver (12)**
ANAT - Parte del cuerpo/Organo (bpoc):	**The gallbladder (2)**
ANAT - Localización o región del cuerpo (blor):	**Head, (2)**
DISO - Localización del trastorno (fndg):	**Ultrasound scan of the liver (12)**
PHYS - Atributo Clínico (clna):	**Ultrasound scan of the liver (12)**
PROC - Procedimiento de diagnóstico (diap):	**Ultrasound scan of the liver (12)**
DEVI - Dispositivo médico (medd):	**Ultrasound scan of the liver (12)**
PHEN - Proceso o Fenomeno natural (npop):	**Ultrasound scan of the liver (12)**

****

Resumen Final:

Resumen del documento "Report33955":

DISPOSITIVOS MÉDICO, DISPENSADOR DE FÁRMACOS: (DEVI)
- **Ultrasound scan of the liver (12)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Head, (2)**
- **The gallbladder (2)**

Report10105

- **Documento médico original con filtrado previo**

COUGH

EMERGENCY DEPARTMENT

I interviewed and examined the patient, as well as reviewed the resident's history, physical and decision-making processes and concur with them. I have also reviewed the nursing notes and electronic health records.
Ms. is a very pleasant in 80s-year-old Caucasian female who presents to the Emergency Department with a cough and shortness of breath over the last one and half weeks. The patient does have a history of COPD and is on Combivent inhaler as well as Advair. The patient reports that she has been having coughing. She denies any significant sputum production with this. She also reports some shortness of breath with walking. She denies any chest pain.
She denies any nausea, vomiting, diarrhea, abdominal pain. She denies any UTI-type symptoms. There is no weakness, paresthesias, headache or vision changes.
Past Medical History, Medications, Allergies, are as noted in the resident's dictation.

The patient denies tobacco, alcohol or illicit drug use.

Physical exam shows a very pleasant elderly Caucasian female conscious, alert and oriented times four.
Her lung exam initially showed only fair air movement and wheezing on forced expiration. However, after her nebulizer treatment, she had clear breath sounds throughout without evidence of wheezing.

Regular rate and rhythm without murmur.
Normoactive bowel sounds. Soft, nontender. No hepatosplenomegaly or masses.

This is an elderly female who presents with the above symptoms. My impression was that they were secondary to bronchospasm. Here in the Emergency Department, the patient was given 60 mg of prednisone p.o. She had the above noted nebulized treatment with significant improvement. The patient had a chest x-ray, PA and lateral which showed no evidence of infiltrates, or other abnormalities. The patient was able to ambulate throughout the Emergency Department without difficulty. It was felt the patient could be safely discharged to home. We will put her on a Medrol Dosepak. She is to continue her other medications. She was given instruction to follow up with her primary care physician in three to five days if not better. Otherwise, she is to return to the Emergency Department if she has any worsening symptoms.

VIRAL BRONCHITIS WITH BRONCHOSPASM.

COUGH

cough, shortness of breath.
The patient is an in 80s-year-old female with a past medical history of COPD, hypertension, hypothyroidism, and GERD, who presents with a two-week history of cough and shortness of breath. The patient states that she has not had any sputum production and denies fever or chills. However, she has had this persistent cough and has had feelings of chest tightness with her coughing spells since yesterday. She denies any sick contacts or recent travel. She states she has been eating and drinking and staying well hydrated without difficulty. She denies any nausea, vomiting, or diarrhea. She has tried some Robitussin over-the-counter with minimal relief.

Hypertension, GERD, hypothyroidism, COPD, status post right total hip replacement.

1. Levothyroxine.
2. Isosorbide mononitrate.
3. Benicar HCT.
4. Caduet.
5. Lorazepam p.r.n.
6. Protonix.
7. Advair.
8. Combivent inhaler.

The patient has a 50-pack year history. She quit smoking many years ago. She denies alcohol or illicit drug use. She lives with her son. She is a widow.

As per the HPI. Otherwise all systems are negative.

Temperature is 98.5, heart rate 69, blood pressure 150/80,
respiratory rate 16, 92% on room air. : She is a Caucasian elderly female in no acute distress who is pleasant.
HEENT: Her extraocular movements are intact. Pupils are equally round and reactive to light. Her posterior pharynx is clear without erythema or exudates. She has a left-sided hearing aid in place.
NECK: Supple with no lymphadenopathy. No JVD.
LUNGS: Clear to auscultation but with deep coughing there are end expiratory wheezes. No rhonchi appreciated.
HEART: Regular rate and rhythm with normal S1 and S2. No murmurs, rubs, or gallops.
ABDOMEN: Soft, nontender, and nondistended with normoactive bowel sounds.
EXTREMITIES: No clubbing, cyanosis, or edema; 2+ pulses.
NEUROLOGIC: She is intact.
SKIN: Her skin has no rashes.
MUSCULOSKELETAL: She has no tenderness or pain.
Chest x-ray, PA and lateral, was done which shows no infiltrate.

The patient likely has a viral upper respiratory infection. She has no signs of pneumonia on chest x-ray and she appears nontoxic. She is able to eat and drink well at home and her vital signs do not show any signs of orthostasis.
She was given two nebulizer treatments while in the emergency department which did improve her aeration greatly. She was also given 60 mg prednisone to start a very short steroid taper. The patient feels much better upon leaving the emergency department and will follow up with her primary care doctor this coming Thursday for reevaluation. She was given a Medrol Dosepak to start tomorrow for a short six-day taper. She was advised to use her Advair regularly and the Combivent up to six times a day as needed for the next several days. The patient was discharged to home in stable condition with her son.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0010200 sosy COUGH
2 0 C1961131 fndg COUGH
3 0 C0562508 hcro EMERGENCY DEPARTMENT
7 0 C0332128 fndg examined
8 0 C0030705 podg the patient,
14 0 C0262926 fndg the resident's history,
16 0 C2004062 fndg the resident's history,
31 0 C1509143 fndg physical
32 0 C1184743 bpoc decision-making processes
33 0 C1522240 phpr decision-making processes
37 0 C1717476 clna the nursing notes
42 0 C0007457 popg pleasant in 80s-year-old Caucasian female
42 0 C0015780 fndg pleasant in 80s-year-old Caucasian female
43 0 C0043210 popg pleasant in 80s-year-old Caucasian female
44 0 C0086287 popg pleasant in 80s-year-old Caucasian female
47 0 C0043157 popg pleasant in 80s-year-old Caucasian female
62 0 C0562508 hcro presents to the Emergency Department
64 0 C0010200 sosy with a cough
65 0 C1961131 fndg with a cough
66 0 C0013404 sosy shortness of breath
67 0 C2707305 clna shortness of breath
70 0 C0030705 podg The patient
71 0 C0024117 dsyn a history of COPD
72 0 C1847014 dsyn a history of COPD
77 0 C0262926 fndg a history of COPD
81 0 C2004062 fndg a history of COPD
83 0 C0353615 medd is on Combivent inhaler
85 0 C0030705 podg The patient
88 0 C0010200 sosy coughing.
98 0 C0013404 sosy some shortness of breath
99 0 C2707305 clna some shortness of breath
127 0 C0030554 sosy paresthesias,
130 0 C0262926 fndg Past Medical History,
131 0 C0455458 fndg Past Medical History,
133 0 C0802604 clna Medications,
134 0 C2598133 clna Medications,
135 0 C0020517 patf Allergies,
136 0 C1314973 clna Allergies,
137 0 C1316572 clna noted in the resident's dictation.
143 0 C0030705 podg The patient
156 0 C0804816 clna Physical exam
159 0 C0001792 popg a very pleasant elderly Caucasian female conscious, alert
159 0 C0007457 popg a very pleasant elderly Caucasian female conscious, alert
159 0 C0015780 fndg a very pleasant elderly Caucasian female conscious, alert

159 0 C0239110 fndg a very pleasant elderly Caucasian female conscious, alert
160 0 C0043210 popg a very pleasant elderly Caucasian female conscious, alert
161 0 C0086287 popg a very pleasant elderly Caucasian female conscious, alert
164 0 C0043157 popg a very pleasant elderly Caucasian female conscious, alert
169 0 C1999167 fndg a very pleasant elderly Caucasian female conscious, alert
181 0 C0024109 bpoc Her lung exam initially
182 0 C1278908 bpoc Her lung exam initially
184 0 C0001868 npop only fair air movement
187 0 C0441722 phpr wheezing on forced expiration.
187 0 C0231875 sosy wheezing on forced expiration.
189 0 C0027524 medd her nebulizer treatment,
190 0 C0087111 topp her nebulizer treatment,
192 0 C1533734 topp her nebulizer treatment,
195 0 C0087111 topp her nebulizer treatment,
197 0 C1533734 topp her nebulizer treatment,
204 0 C0232693 fndg Normoactive bowel sounds.
210 0 C0001792 popg an elderly female
210 0 C0015780 fndg an elderly female
211 0 C0043210 popg an elderly female
212 0 C0086287 popg an elderly female
215 0 C1999167 fndg an elderly female
221 0 C1457887 sosy presents with the above symptoms.
223 0 C1457887 sosy presents with the above symptoms.
226 0 C0006266 dsyn secondary to bronchospasm.
227 0 C0562508 hcro Here in the Emergency Department,
228 0 C0030705 podg the patient
232 0 C0032952 horm 60 mg of prednisone p.o.
233 0 C2346927 elii 60 mg of prednisone p.o.
234 0 C1273937 fndg the above noted nebulized treatment with significant improvement.
234 0 C1549531 blor the above noted nebulized treatment with significant improvement.
235 0 C0087111 topp the above noted nebulized treatment with significant improvement.
237 0 C1533734 topp the above noted nebulized treatment with significant improvement.
249 0 C0030705 podg The patient
250 0 C0039985 diap a chest x-ray,
251 0 C1114557 clna a chest x-ray,
259 0 C1299581 fndg able to
260 0 C0562508 hcro ambulate throughout the Emergency Department
261 0 C1299586 fndg without difficulty.
264 0 C0030705 podg the patient
268 0 C1878038 drdd her on a Medrol Dosepak.
271 0 C1114750 clna her other medications.
279 0 C0589120 fndg to follow up with her primary care physician
287 0 C0562508 hcro return to the Emergency Department
289 0 C1457887 sosy any worsening symptoms.
294 0 C0276141 dsyn VIRAL BRONCHITIS WITH BRONCHOSPASM.
294 0 C0006266 dsyn VIRAL BRONCHITIS WITH BRONCHOSPASM.
295 0 C0010200 sosy COUGH
296 0 C1961131 fndg COUGH
297 0 C0010200 sosy cough,
298 0 C1961131 fndg cough,
299 0 C0013404 sosy shortness of breath.
300 0 C2707305 clna shortness of breath.
301 0 C0030705 podg The patient
302 0 C0015780 fndg an in 80s-year-old female
303 0 C0043210 popg an in 80s-year-old female
304 0 C0086287 popg an in 80s-year-old female
312 0 C0262926 fndg with a past medical history
313 0 C0455458 fndg with a past medical history
314 0 C0024117 dsyn of COPD,
315 0 C1847014 dsyn of COPD,
316 0 C0020538 dsyn hypertension,
317 0 C1963138 fndg hypertension,
318 0 C0020676 dsyn hypothyroidism,
319 0 C0017168 dsyn GERD,
320 0 C0010200 sosy presents with a two-week history of cough
321 0 C1961131 fndg presents with a two-week history of cough
326 0 C0262926 fndg presents with a two-week history of cough
330 0 C2004062 fndg presents with a two-week history of cough
368 0 C0013404 sosy shortness of breath.
369 0 C2707305 clna shortness of breath.
370 0 C0030705 podg The patient
380 0 C0562483 sosy this persistent cough
381 0 C0232292 sosy feelings of chest tightness
382 0 C0010200 sosy with her coughing spells
382 0 C0302539 fndg with her coughing spells
394 0 C0013470 orgf eating

396 0 C0684271 orgf drinking
399 0 C0720930 orch hydrated without difficulty.
399 0 C1299586 fndg hydrated without difficulty.
410 0 C0723110 orch some Robitussin
411 0 C0564405 fndg over-the-counter with minimal relief.
414 0 C0020538 dsyn Hypertension,
415 0 C1963138 fndg Hypertension,
416 0 C0017168 dsyn GERD,
417 0 C0020676 dsyn hypothyroidism,
418 0 C0024117 dsyn COPD,
419 0 C1847014 dsyn COPD,
420 0 C0040508 topp status post right total hip replacement.
425 0 C1881373 orch Levothyroxine.
429 0 C0064079 orch Isosorbide mononitrate.
433 0 C1330085 orch Benicar HCT.
435 0 C1330119 orch Caduet.
438 0 C0024002 orch Lorazepam p.r.n.
439 0 C0876139 orch Protonix.
441 0 C0353615 medd Combivent inhaler.
442 0 C0030705 podg The patient
446 0 C0262926 fndg a 50-pack year history.
448 0 C2004062 fndg a 50-pack year history.
452 0 C1881674 phpr smoking
462 0 C0557132 fndg lives with her son.
463 0 C0206275 fndg a widow.
464 0 C1510465 famg a widow.
467 0 C1704459 medd all systems
469 0 C1513916 fndg negative.
475 0 C0005823 orgf blood pressure 150/80,
476 0 C1271104 fndg blood pressure 150/80,
477 0 C1272641 fndg blood pressure 150/80,
478 0 C0231832 clna respiratory rate 16,
479 0 C0489258 clna respiratory rate 16,
502 0 C1512338 blor HEENT
503 0 C0026649 orgf Her extraocular movements
506 0 C0034121 bpoc Pupils
508 0 C0023693 npop reactive to light.
511 0 C0031354 bpoc Her posterior pharynx
512 0 C1278903 bpoc Her posterior pharynx
515 0 C0018768 medd a left-sided hearing aid in place.
521 0 C0027530 blor NECK
522 0 C1281592 bpoc NECK
525 0 C0024109 bpoc LUNGS
526 0 C0004339 diap Clear to auscultation
527 0 C0010200 sosy but with deep coughing there
528 0 C0043144 sosy end expiratory wheezes.
530 0 C0018787 bpoc HEART
531 0 C1281570 bpoc HEART
534 0 C0871269 fndg rhythm with normal S1
540 0 C1179705 blor S2.
544 0 C0000726 blor ABDOMEN
545 0 C0230168 blor ABDOMEN
546 0 C1281594 bpoc ABDOMEN
548 0 C0232693 fndg nondistended with normoactive bowel sounds.
549 0 C0015385 bpoc EXTREMITIES
550 0 C0278454 bpoc EXTREMITIES
565 0 C0444099 bdsu SKIN
566 0 C1123023 bdsy SKIN
567 0 C1278993 bdsy SKIN
573 0 C2707260 clna MUSCULOSKELETAL
577 0 C0039985 diap Chest x-ray,
578 0 C1114557 clna Chest x-ray,
584 0 C0030705 podg The patient likely
586 0 C0009443 dsyn a viral upper respiratory infection.
595 0 C1299581 fndg able to
596 0 C0013470 orgf eat
601 0 C0518766 clna her vital signs
610 0 C0027524 medd two nebulizer treatments
610 0 C0087111 topp two nebulizer treatments
612 0 C0562508 hcro while in the emergency department
620 0 C0032952 horm 60 mg prednisone to
621 0 C2346927 elii 60 mg prednisone to
626 0 C0030705 podg The patient feels
630 0 C0562508 hcro the emergency department
631 0 C0589120 fndg follow up with her primary care doctor
639 0 C1878038 drdd a Medrol Dosepak to

653 0 C0591282 orch the Combivent up to six
662 0 C0030705 podg The patient
664 0 C0012634 dsyn in stable condition
670 0 C0037683 famg with her son.

- **Distribución de Frecuencia de Conceptos (Densidad Cualitativa):**

C0030705 podg The patient likely (13)
C0010200 sosy presents with a two-week history of cough (8)
C0562508 hcro ambulate throughout the Emergency Department (7)
C0262926 fndg presents with a two-week history of cough (6)
C1961131 fndg presents with a two-week history of cough (5)
C2707305 clna some shortness of breath (4)
C2004062 fndg presents with a two-week history of cough (4)
C0087111 topp the above noted nebulized treatment with significant improvement.(4)
C0086287 popg a very pleasant elderly Caucasian female conscious, alert (4)
C0043210 popg a very pleasant elderly Caucasian female conscious, alert (4)
C0015780 fndg a very pleasant elderly Caucasian female conscious, alert (4)
C0013404 sosy some shortness of breath (4)
C1847014 dsyn a history of COPD (3)
C1533734 topp the above noted nebulized treatment with significant improvement. (3)
C1457887 sosy presents with the above symptoms. (3)
C0024117 dsyn a history of COPD (3)
C2346927 elii 60 mg of prednisone p.o. (2)
C1999167 fndg a very pleasant elderly Caucasian female conscious, alert (2)
C1963138 fndg Hypertension, (2)
C1878038 drdd her on a Medrol Dosepak. (2)
C1299586 fndg hydrated without difficulty. (2)
C1299581 fndg able to (2)
C1114557 clna a chest x-ray, (2)
C0589120 fndg to follow up with her primary care physician (2)
C0455458 fndg with a past medical history (2)
C0353615 medd is on Combivent inhaler (2)
---------- Corte 2º Cuartil (50 %) ----------
C0232693 fndg nondistended with normoactive bowel sounds. (2)
C0043157 popg a very pleasant elderly Caucasian female conscious, alert (2)
C0039985 diap a chest x-ray, (2)
C0032952 horm 60 mg of prednisone p.o. (2)
C0027524 medd two nebulizer treatments (2)
C0024109 bpoc Her lung exam initially (2)
C0020676 dsyn hypothyroidism, (2)
C0020538 dsyn hypertension, (2)
C0017168 dsyn GERD, (2)
C0013470 orgf eating (2)
C0007457 popg a very pleasant elderly Caucasian female conscious, alert (2)
C0006266 dsyn VIRAL BRONCHITIS WITH BRONCHOSPASM. (2)
C0001792 popg a very pleasant elderly Caucasian female conscious, alert (2)
--- resto (1's) ---

"Report10105" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	**The patient likely (**13)
LIVB - Grupo de población (popg):	**a very pleasant elderly Caucasian female conscious, alert (**4**)**
ANAT - Parte del cuerpo/Organo (bpoc):	**Her lung exam initially (**2)
DISO - Localización del trastorno (fndg):	**presents with a two-week history of cough (**6**)**
DISO - Localización del trastorno (fndg):	**a very pleasant elderly Caucasian female conscious, alert (**4**)**
DISO - Localización del trastorno (fndg):	**Hypertension (**2**)**
DISO - Localización del trastorno (fndg):	**hydrated without difficulty (**2**)**
DISO - Localización del trastorno (fndg):	**able to (**2**)**
DISO - Localización del trastorno (fndg):	**to follow up with her primary care physician (**2**)**
DISO - Localización del trastorno (fndg):	**with a past medical history (**2**)**
DISO - Localización del trastorno (fndg):	**nondistended with normoactive bowel sounds (**2**)**
DISO - Signo o Sintoma (sosy):	**presents with a two-week history of cough (**8**)**
DISO - Signo o Sintoma (sosy):	**some shortness of breath (**4**)**
DISO - Signo o Sintoma (sosy):	**presents with the above symptoms (**3**)**
DISO - Enfermedad o síndrome (dsyn):	**a history of COPD (**3**)**
DISO - Enfermedad o síndrome (dsyn):	**hypothyroidism (**2**)**
DISO - Enfermedad o síndrome (dsyn):	**hypertension (**2**)**
DISO - Enfermedad o síndrome (dsyn):	**GERD (2)**
DISO - Enfermedad o síndrome (dsyn):	**VIRAL BRONCHITIS WITH BRONCHOSPASM (**2**)**
PHYS - Atributo Clínico (clna):	**some shortness of breath (**4**)**

PHYS - Atributo Clínico (clna): **a chest x-ray (2)**
PHYS - Función Orgánica (orgf): **eating (2)**
PROC - Procedimiento terapeútico (topp): **the above noted nebulized treatment with significant improvement (4)**
PROC - Procedimiento de diagnóstico (diap): **a chest x-ray (2)**
CHEM - Elemento químico/fármaco (elii): **60 mg of prednisone p.o. (2)**
CHEM - Hormonas (horm): **60 mg of prednisone p.o. (2)**
DEVI - Dispositivo médico (medd): **is on Combivent inhaler (2)**
DEVI - Dispositivo médico (medd): **two nebulizer treatments (2)**
DEVI - Dispensador de fármacos (drdd): **her on a Medrol Dosepak (2)**
ORGA - Organización cuidados de la salud (hcro): **ambulate throughout the Emergency Department (7)**
****

- **Agregación de Frases (Densidad Cuantitativa):**

"Report10105" (Densidad Cualitativa + Cuantitativa)
****
LIVB - Tipo de paciente (podg): **The patient likely (13)**
LIVB - Grupo de población (popg): **a very pleasant elderly Caucasian female conscious, alert (20)**
ANAT - Parte del cuerpo/Organo (bpoc): **Her lung exam initially (2)**
DISO - Localización del trastorno (fndg): **presents with a two-week history of cough (23)**
DISO - Localización del trastorno (fndg): **a very pleasant elderly Caucasian female conscious, alert (20)**
DISO - Localización del trastorno (fndg): **hypertension (4)**
DISO - Localización del trastorno (fndg): **hydrated without difficulty (2)**
DISO - Localización del trastorno (fndg): **able to (2)**
DISO - Localización del trastorno (fndg): **to follow up with her primary care physician (2)**
DISO - Localización del trastorno (fndg): **with a past medical history (2)**
DISO - Localización del trastorno (fndg): **nondistended with normoactive bowel sounds (2)**
DISO - Signo o Sintoma (sosy): **presents with a two-week history of cough (23)**
DISO - Signo o Sintoma (sosy): **some shortness of breath (8)**
DISO - Signo o Sintoma (sosy): **presents with the above symptoms (3)**
DISO - Enfermedad o síndrome (dsyn): **a history of COPD (6)**
DISO - Enfermedad o síndrome (dsyn): **hypothyroidism (2)**
DISO - Enfermedad o síndrome (dsyn): **hypertension (4)**
DISO - Enfermedad o síndrome (dsyn): **GERD (2)**
DISO - Enfermedad o síndrome (dsyn): **VIRAL BRONCHITIS WITH BRONCHOSPASM (2)**
PHYS - Atributo Clínico (clna): **some shortness of breath (8)**
PHYS - Atributo Clínico (clna): **a chest x-ray, (4)**
PHYS - Función Orgánica (orgf): **eating (2)**
PROC - Procedimiento terapeútico (topp): **the above noted nebulized treatment with significant improvement. (7)**
PROC - Procedimiento de diagnóstico (diap): **a chest x-ray, (4)**
CHEM - Elemento químico/fármaco (elii): **60 mg of prednisone p.o. (4)**
CHEM - Hormonas (horm): **60 mg of prednisone p.o. (4)**
DEVI - Dispositivo médico (medd): **is on Combivent inhaler (2)**
DEVI - Dispositivo médico (medd): **two nebulizer treatments (2)**
DEVI - Dispensador de fármacos (drdd): **her on a Medrol Dosepak (2)**
ORGA - Organización cuidados de la salud (hcro): **ambulate throughout the Emergency Department (7)**
****

Resumen Final:

Resumen del documento "Report10105":

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **presents with a two-week history of cough (23)**
- **presents with the above symptoms (3)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **a very pleasant elderly Caucasian female conscious, alert (20)**
- **The patient likely (13)**

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **some shortness of breath (8)**
- **a chest x-ray, (4)**

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **the above noted nebulized treatment with significant improvement. (7)**
- **a chest x-ray, (4)**

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)
- **60 mg of prednisone p.o. (4)**

DISPOSITIVOS MÉDICO, DISPENSADOR DE FÁRMACOS: (DEVI)
- **is on Combivent inhaler (2)**
- **two nebulizer treatments (2)**
- **her on a Medrol Dosepak (2)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **Her lung exam initially (2)**

ORGANIZACIÓN PROFESIONAL Y SANITARIA: (ORGA)
- **ambulate throughout the Emergency Department (7)**

Report17592

- **Documento médico original con filtrado previo**

OPERATION

Operative Report

PREOPERATIVE DIAGNOSIS: End-stage osteoarthritis, left knee.
POSTOPERATIVE DIAGNOSIS: End-stage osteoarthritis, left knee.
Left total knee arthroplasty.

ANESTHESIA: Spinal.
PATHOLOGY: Yes.
CULTURES: None.
DRAINS: Drain x1.
CONDITION: Good.

Recovery room and then to the floor.
The patient is well known to me from preoperative evaluation. He is very satisfied with a previous right total knee arthroplasty and wants to proceed with the left. The patient has had continuing pain and has failed conservative measures on this side. We have discussed the risks including, but not limited to infection, neurovascular injury, fracture, bleeding, hemarthrosis, need for further surgery, need for revision surgery, need for transfusion, continued stiffness, continued pain, myocardial infarction, pulmonary embolism, stroke, and/or death. These risks have been discussed at length in the past, and there was opportunity for further questions and answers with discussion today, and he was satisfied with the process and wanted to proceed.
The patient was brought to the operating room and underwent a spinal anesthetic. A gram of Ancef was given prophylactically. A Foley catheter was placed atraumatically. The patient's left leg was prepped and draped in the usual fashion. The knee was approached through a standard longitudinal incision and carried down to a standard medial parapatellar arthrotomy. A medial joint line release was performed as necessary, and the menisci removed in 2 steps. The osteophytes were removed as necessary. A partial synovectomy was performed superiorly. The ACL was already absent. The patellofemoral ligament was released. The intercondylar notch was drilled in a 4-0 starting point, and then a intramedullary guide rod was placed to place a 6-degree 8 mm distal femoral cut. Once the jig was affixed with the drill pins, the distal femur was cut and was checked to be appropriately flush.
The femur was sized for a size 7. A size 7 cutting jig was then applied, and the anterior, posterior, anterior and posterior chamfer cuts were made. The patient then had a box cut made for the cam for the PCL device, which was performed atraumatically. The tibia was then subluxed forward, and the tibia was cut using the external tibial jig taking into account flexion, extension, internal, and external rotation and varus valgus alignment. Once this was cut, trials were inserted and the size 7 tibia, size 7 femur, and a 9-mm polyethylene all appeared to be appropriately sized. This gave full extension and flexion past 125 degrees without liftoff, and good collateral stability. At this point, the patella was everted and then osteophytes removed. The patella was then cut using a standard patellar cutting jig and sized for a 35 x 10-mm asymmetric patella. The lug holes were drilled and a trial was inserted. The patella tracked very well. The trials were removed.
The tibial trial was reinserted to prepare the fin slots using a fin punch. This was then followed by copious irrigation to cleanse the bony surfaces and dry them. This was followed by insertion of a size 7 tibia, size 7 femur, and a 35 x 10-mm polyethylene button using methylmethacrylate, removing excess cement as necessary. A trial polyethylene was inserted and the knee was held in full extension with longitudinal compression and the patella was held with a clamp. The tourniquet was released at 69 minutes. Hemostasis was obtained as necessary. Excess cement was removed as necessary. The patient's components were then checked after the cement was cured and found to have the same excellent tracking and good collateral stability with full extension and flexion as with the trials. As such, a true 9-mm polyethylene was inserted with a similar outcome using standard technique. Once hemostasis was obtained and after copious irrigation with pulsatile lavage, a separate stab wound incision was used to bring out a medium Hemovac drain followed by closure with #2 Polysorb for the quadriceps, medial retinaculum, and the distal medial tissues, followed by closure of the skin with 2-0 Polysorb and 4-0 Monocryl with Steri Strips for the skin. A sterile dressing was applied, and the patient was then taken to the recovery room in good condition having tolerated the procedure well.
In accordance with CMS guidelines for teaching hospitals, I was scrubbed for the critical components of the procedure which in this case was the entirety of the procedure up until the end of superficial closure, for the remainder of which I was available in the operating suite as well as for transportation to the recovery room.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0543467 diap OPERATION
4 0 C1318968 fndg PREOPERATIVE DIAGNOSIS
5 0 C0029408 dsyn End-stage osteoarthritis,
6 0 C0230432 bpoc left knee.
7 0 C1279572 bpoc left knee.
8 0 C1318969 fndg POSTOPERATIVE DIAGNOSIS
12 0 C0086511 topp Left total knee arthroplasty.
14 0 C0002903 topp ANESTHESIA
15 0 C0002930 bmod ANESTHESIA
16 0 C0278134 sosy ANESTHESIA
18 0 C0030664 bmod PATHOLOGY
20 0 C0677042 patf PATHOLOGY
25 0 C0013103 topp DRAINS
26 0 C0180499 medd DRAINS
29 0 C0012634 dsyn CONDITION
33 0 C0034871 hcro Recovery room
36 0 C0030705 podg The patient
39 0 C1293091 diap from preoperative evaluation.
41 0 C0086511 topp satisfied with a previous right total knee arthroplasty
43 0 C1444647 fndg wants to
47 0 C0030193 sosy continuing pain
60 0 C0038894 bmod need for further surgery,
62 0 C0543467 diap need for further surgery,
66 0 C0035110 topp need for revision surgery,
67 0 C0558347 topp need for revision surgery,
68 0 C0949626 topp need for revision surgery,
72 0 C0005841 topp need for transfusion,
74 0 C0455012 topp need for transfusion,
75 0 C1879316 topp need for transfusion,
80 0 C0427008 sosy continued stiffness,
81 0 C0030193 sosy continued pain,
82 0 C0027051 dsyn myocardial infarction,
83 0 C2926063 clna myocardial infarction,
84 0 C0034065 dsyn pulmonary embolism,
85 0 C0038454 dsyn stroke,
86 0 C0011065 orgf death.
87 0 C1306577 fndg death.
93 0 C0557061 topp answers with discussion today,
93 0 C0310367 antb answers with discussion today,
98 0 C1444647 fndg wanted
100 0 C0030705 podg The patient
101 0 C0029064 hcro brought to the operating room
103 0 C0002928 topp a spinal anesthetic.
104 0 C0700926 antb A gram of Ancef
108 0 C0179804 medd A Foley catheter
112 0 C0030705 podg The patient's left leg
112 0 C0230443 bpoc The patient's left leg
113 0 C1279606 bpoc The patient's left leg
117 0 C0180504 medd draped in the usual fashion.
118 0 C0022742 bpoc The knee
119 0 C0022745 bsoj The knee
120 0 C1283838 blor The knee
121 0 C1963703 blor The knee
122 0 C0184898 topp approached through a standard longitudinal incision
123 0 C2338258 blor approached through a standard longitudinal incision
135 0 C0699809 fndg carried down
136 0 C0185160 topp to a standard medial parapatellar arthrotomy.
137 0 C0700234 topp to a standard medial parapatellar arthrotomy.
142 0 C0022417 bsoj A medial joint line release
147 0 C1963578 topp A medial joint line release
160 0 C0392905 bdsy A medial joint line release
178 0 C1269611 bsoj A medial joint line release
196 0 C1706309 medd A medial joint line release
215 0 C0224498 bpoc the menisci
218 0 C0015302 dsyn The osteophytes
219 0 C1956089 dsyn The osteophytes
221 0 C0343132 topp A partial synovectomy
226 0 C2699517 fndg already absent.
227 0 C0023685 tisu The patellofemoral ligament
228 0 C1269080 bpoc The patellofemoral ligament
231 0 C1235660 bsoj The intercondylar notch

232 0 C0175642 medd drilled in a 4-0 starting point,
232 0 C2924612 blor drilled in a 4-0 starting point,
236 0 C0337279 topp drilled in a 4-0 starting point,
239 0 C0336581 medd a intramedullary guide rod
240 0 C0302614 medd a intramedullary guide rod
249 0 C0015811 bpoc a 6-degree 8 mm distal femoral cut.
249 0 C0000925 inpo a 6-degree 8 mm distal femoral cut.
250 0 C1883724 phpr a 6-degree 8 mm distal femoral cut.
259 0 C0175718 medd affixed with the drill pins,
260 0 C1168556 medd affixed with the drill pins,
261 0 C1704584 medd affixed with the drill pins,
262 0 C0448194 bpoc the distal femur
263 0 C0000925 inpo cut
264 0 C1883724 phpr cut
267 0 C0016382 sosy appropriately flush.
269 0 C0015811 bpoc The femur
270 0 C1279112 bpoc The femur
272 0 C0152060 topp A size 7 cutting jig
277 0 C0000925 inpo posterior chamfer cuts
279 0 C0030705 podg The patient then
280 0 C0000925 inpo a box cut
281 0 C1883724 phpr a box cut
282 0 C1533124 medd a box cut
285 0 C0025080 medd for the PCL device,
288 0 C0040184 bpoc The tibia
289 0 C1279118 bpoc The tibia
291 0 C0040184 bpoc the tibia
292 0 C1279118 bpoc the tibia
293 0 C0000925 inpo cut
294 0 C1883724 phpr cut
296 0 C0040184 bpoc the external tibial jig taking into account flexion,
296 0 C0231452 ortf the external tibial jig taking into account flexion,
297 0 C0426070 sosy the external tibial jig taking into account flexion,
298 0 C2741673 clna the external tibial jig taking into account flexion,
310 0 C0000925 inpo cut,
311 0 C1883724 phpr cut,
314 0 C0040184 bpoc the size 7 tibia,
315 0 C1279118 bpoc the size 7 tibia,
316 0 C0015811 bpoc size 7 femur,
317 0 C1279112 bpoc size 7 femur,
329 0 C0231452 ortf flexion past 125 degrees without liftoff,
331 0 C0426070 sosy flexion past 125 degrees without liftoff,
339 0 C0030647 bpoc the patella
341 0 C1279123 bpoc the patella
342 0 C0015302 dsyn then osteophytes
343 0 C1956089 dsyn then osteophytes
345 0 C0030647 bpoc The patella
347 0 C1279123 bpoc The patella
349 0 C0000925 inpo cut
350 0 C1883724 phpr cut
352 0 C0030647 bpoc a standard patellar cutting jig
352 0 C0152060 topp a standard patellar cutting jig
358 0 C0030647 bpoc sized for a 35 x 10-mm asymmetric patella.
360 0 C1279123 bpoc sized for a 35 x 10-mm asymmetric patella.
361 0 C0544726 fndg The lug holes
362 0 C0175642 medd drilled
364 0 C0337279 topp drilled
367 0 C0030647 bpoc The patella
369 0 C1279123 bpoc The patella
376 0 C0040184 bpoc The tibial trial
379 0 C0230113 bpoc the fin slots
381 0 C0230113 bpoc a fin punch.
381 0 C0182555 medd a fin punch.
383 0 C0022100 topp followed by copious irrigation to
387 0 C0825429 blor the bony surfaces
388 0 C0011682 npop dry
391 0 C0021107 topp followed by insertion of a size 7 tibia,
391 0 C0040184 bpoc followed by insertion of a size 7 tibia,
392 0 C1279118 bpoc followed by insertion of a size 7 tibia,
399 0 C0015811 bpoc size 7 femur,
400 0 C1279112 bpoc size 7 femur,
408 0 C0011343 bpoc excess cement as necessary.
415 0 C0022742 bpoc the knee
416 0 C0022745 bsoj the knee
417 0 C1283838 blor the knee
418 0 C1963703 blor the knee

424 0 C0565514 topp with longitudinal compression
425 0 C0728907 npop with longitudinal compression
427 0 C0030647 bpoc the patella
429 0 C1279123 bpoc the patella
430 0 C0175721 medd held with a clamp.
433 0 C0175721 medd held with a clamp.
436 0 C0040519 medd The tourniquet
443 0 C0019116 ortf Hemostasis
444 0 C0740166 topp Hemostasis
446 0 C0011343 bpoc Excess cement
451 0 C0030705 podg The patient's components
453 0 C0011343 bpoc checked after the cement
458 0 C0037088 sosy found
459 0 C0243095 fndg found
460 0 C2825141 fndg found
465 0 C1275670 bpoc good collateral stability with full extension
473 0 C0231452 ortf flexion
474 0 C0426070 sosy flexion
483 0 C0019116 ortf Once hemostasis
484 0 C0740166 topp Once hemostasis
486 0 C0022100 topp and after copious irrigation
489 0 C0022100 topp with pulsatile lavage,
491 0 C0043255 inpo a separate stab wound incision
491 0 C0184898 topp a separate stab wound incision
492 0 C2338258 blor a separate stab wound incision
493 0 C1273517 fndg used
502 0 C0185003 topp followed by closure
504 0 C0071681 orch with #2 Polysorb
507 0 C0224440 bpoc for the quadriceps,
508 0 C0224996 bpoc medial retinaculum,
509 0 C0040300 tisu the distal medial tissues,
510 0 C0185003 topp followed by closure of the skin
510 0 C0444099 bdsu followed by closure of the skin
511 0 C1123023 bdsy followed by closure of the skin
512 0 C1278993 bdsy followed by closure of the skin
515 0 C1278993 bdsy followed by closure of the skin
516 0 C0071681 orch with 2-0 Polysorb
520 0 C1321564 medd Strips for the skin.
520 0 C0444099 bdsu Strips for the skin.
521 0 C1123023 bdsy Strips for the skin.
522 0 C1278993 bdsy Strips for the skin.
523 0 C0013119 medd A sterile dressing
524 0 C0278286 topp A sterile dressing
525 0 C0518459 fndg A sterile dressing
527 0 C0678108 patf A sterile dressing
532 0 C0030705 podg the patient
534 0 C0034871 hcro taken to the recovery room
535 0 C0012634 dsyn in good condition
538 0 C0184661 diap the procedure well.
540 0 C1948041 topp the procedure well.
547 0 C0020027 hcro for teaching hospitals,
549 0 C0184661 diap scrubbed for the critical components of the procedure
550 0 C1948041 topp scrubbed for the critical components of the procedure
553 0 C1706255 medd which in this case
555 0 C0184661 diap the entirety of the procedure
556 0 C1948041 topp the entirety of the procedure
561 0 C0185003 topp the end of superficial closure,
571 0 C0034871 hcro to the recovery room.

- **Distribución de Densidad de Conceptos (Densidad Cualitativa):**

C0000925 inpo a 6-degree 8 mm distal femoral cut. (7)
C1883724 phpr a 6-degree 8 mm distal femoral cut. (6)
C0040184 bpoc the external tibial jig taking into account flexion, (6)
C0030705 podg The patient's left leg (6)
C0030647 bpoc a standard patellar cutting jig (6)
C1279123 bpoc sized for a 35 x 10-mm asymmetric patella. (5)
C1279118 bpoc followed by insertion of a size 7 tibia, (4)
C0015811 bpoc a 6-degree 8 mm distal femoral cut. (4)
C1948041 topp scrubbed for the critical components of the procedure (3)
C1279112 bpoc size 7 femur, (3)
C1278993 bdsy followed by closure of the skin (3)
C0426070 sosy the external tibial jig taking into account flexion, (3)

C0231452 ortf flexion past 125 degrees without liftoff, (3)
C0185003 topp followed by closure of the skin (2)
C0184661 diap scrubbed for the critical components of the procedure (2)
C0034871 hcro taken to the recovery room (2)
C0022100 topp followed by copious irrigation to (2)
C0011343 bpoc excess cement as necessary. (2)
C2338258 blor approached through a standard longitudinal incision (2)
C1963703 blor The knee (2)
C1956089 dsyn The osteophytes (2)
C1283838 blor The knee (2)
C1123023 bdsy followed by closure of the skin (2)
C0740166 topp Once hemostasis (2)
C0543467 diap need for further surgery, (2)
C0444099 bdsu Strips for the skin. (2)
C0337279 topp drilled in a 4-0 starting point, (2)
C0230113 bpoc the fin slots (2)
C0184898 topp approached through a standard longitudinal incision (2)
C0175721 medd held with a clamp. (2)
C0175642 medd drilled in a 4-0 starting point, (2)
---------- Corte 2º Cuartil (50 %) ----------
C0152060 topp a standard patellar cutting jig (2)
C0086511 topp satisfied with a previous right total knee arthroplasty (2)
C0071681 orch with 2-0 Polysorb (2)
C0030193 sosy continuing pain (2)
C0022745 bsoj The knee (2)
C0022742 bpoc The knee (2)
C0019116 ortf Once hemostasis (2)
C0015302 dsyn The osteophytes (2)
C0012634 dsyn in good condition (2)
Resto(1's)

"Report17592" (Densidad Cualitativa)
****

LIVB - Tipo de paciente (podg):	**The patient's left leg (6)**
ANAT - Parte del cuerpo/Organo (bpoc):	**the external tibial jig taking into account flexion (6)**
ANAT - Parte del cuerpo/Organo (bpoc):	**a standard patellar cutting jig (6)**
ANAT - Parte del cuerpo/Organo (bpoc):	**sized for a 35 x 10-mm asymmetric patella. (5)**
ANAT - Parte del cuerpo/Organo (bpoc):	**followed by insertion of a size 7 tibia (4)**
ANAT - Parte del cuerpo/Organo (bpoc):	**a 6-degree 8 mm distal femoral cut. (4)**
ANAT - Parte del cuerpo/Organo (bpoc):	**size 7 femur (3)**
ANAT - Parte del cuerpo/Organo (bpoc):	**excess cement as necessary (2)**
ANAT - Parte del cuerpo/Organo (bpoc):	**the fin slots (2)**
ANAT - Parte del cuerpo/Organo (bpoc):	**The knee (2)**
ANAT - Substancia del cuerpo (bdsu):	**Strips for the skin (2)**
ANAT - Localización o región del cuerpo (blor):	**approached through a standard longitudinal incision (2)**
ANAT - Localización o región del cuerpo (blor):	**The knee (2)**
ANAT - Espacio o unión del cuerpo (bsoj):	**The knee (2)**
ANAT - Sistema corporal (bdsy):	**followed by closure of the skin (3)**
DISO - Signo o Sintoma (sosy):	**the external tibial jig taking into account flexion (3)**
DISO - Signo o Sintoma (sosy):	**continuing pain (2)**
DISO - Enfermedad o síndrome (dsyn):	**The osteophytes (2)**
DISO - Enfermedad o síndrome (dsyn):	**in good condition (2)**
DISO - Lesión o envenenamiento (inpo)	**a 6-degree 8 mm distal femoral cut (7)**
PHYS - Órgano o Tejido (ortf):	**flexion past 125 degrees without liftoff (3)**
PHYS - Órgano o Tejido (ortf):	**Once hemostasis (2)**
PROC - Procedimiento terapeútico (topp):	**scrubbed for the critical components of the procedure (3)**
PROC - Procedimiento terapeútico (topp):	**followed by closure of the skin (2)**
PROC - Procedimiento terapeútico (topp):	**followed by copious irrigation to (2)**
PROC - Procedimiento terapeútico (topp):	**Once hemostasis (2)**
PROC - Procedimiento terapeútico (topp):	**drilled in a 4-0 starting point (2)**
PROC - Procedimiento terapeútico (topp):	**approached through a standard longitudinal incision (2)**
PROC - Procedimiento terapeútico (topp):	**a standard patellar cutting jig (2)**
PROC - Procedimiento terapeútico (topp):	**satisfied with a previous right total knee arthroplasty (2)**
PROC - Procedimiento de diagnóstico (diap):	**scrubbed for the critical components of the procedure (2)**
PROC - Procedimiento de diagnóstico (diap):	**need for further surgery (2)**
CHEM - Química Orgánica/fármaco (orch):	**with 2-0 Polysorb (2)**
DEVI - Dispositivo médico (medd):	**held with a clamp (2)**
DEVI - Dispositivo médico (medd):	**drilled in a 4-0 starting point (2)**
PHEN - Fenomeno o proceso (phpr):	**a 6-degree 8 mm distal femoral cut (6)**
ORGA - Organización cuidados de la salud (hcro):	**taken to the recovery room (2)**

****

- **Repetición de Frases (Densidad Cuantitativa):**

"Report17592" (Densidad Cualitativa + Cuantitativa)
****

LIVB - Tipo de paciente (podg):	**The patient's left leg (6)**
ANAT - Parte del cuerpo/Organo (bpoc):	**the external tibial jig taking into account flexion, (9)**
ANAT - Parte del cuerpo/Organo (bpoc):	**a standard patellar cutting jig (8)**
ANAT - Parte del cuerpo/Organo (bpoc):	**sized for a 35 x 10-mm asymmetric patella. (5)**
ANAT - Parte del cuerpo/Organo (bpoc):	**followed by insertion of a size 7 tibia (4)**
ANAT - Parte del cuerpo/Organo (bpoc):	**a 6-degree 8 mm distal femoral cut. (17)**
ANAT - Parte del cuerpo/Organo (bpoc):	**size 7 femur (3)**
ANAT - Parte del cuerpo/Organo (bpoc):	**excess cement as necessary (2)**
ANAT - Parte del cuerpo/Organo (bpoc):	**the fin slots (2)**
ANAT - Parte del cuerpo/Organo (bpoc):	**The knee (8)**
ANAT - Substancia del cuerpo (bdsu):	**Strips for the skin (2)**
ANAT - Localización o región del cuerpo (blor):	**approached through a standard longitudinal incision (4)**
ANAT - Localización o región del cuerpo (blor):	**The knee (4-8)**
ANAT - Espacio o unión del cuerpo (bsoj):	**The knee (4-8)**
ANAT - Sistema corporal (bdsy):	**followed by closure of the skin (7)**
DISO - Signo o Sintoma (sosy):	**the external tibial jig taking into account flexion, (9)**
DISO - Signo o Sintoma (sosy):	**continuing pain (2)**
DISO - Enfermedad o síndrome (dsyn):	**The osteophytes (4)**
DISO - Enfermedad o síndrome (dsyn):	**in good condition (2)**
DISO - Lesión o envenenamiento (inpo)	**a 6-degree 8 mm distal femoral cut. (17)**
PHYS - Órgano o Tejido (ortf):	**flexion past 125 degrees without liftoff (3)**
PHYS - Órgano o Tejido (ortf):	**Once hemostasis (2-4)**
PROC - Procedimiento terapeútico (topp):	**scrubbed for the critical components of the procedure (5)**
PROC - Procedimiento terapeútico (topp):	**followed by closure of the skin (7)**
PROC - Procedimiento terapeútico (topp):	**followed by copious irrigation to (2)**
PROC - Procedimiento terapeútico (topp):	**Once hemostasis (2-4)**
PROC - Procedimiento terapeútico (topp):	**drilled in a 4-0 starting point (4)**
PROC - Procedimiento terapeútico (topp):	**approached through a standard longitudinal incision (4)**
PROC - Procedimiento terapeútico (topp):	**a standard patellar cutting jig (6-8)**
PROC - Procedimiento terapeútico (topp):	**satisfied with a previous right total knee arthroplasty (2)**
PROC - Procedimiento de diagnóstico (diap):	**scrubbed for the critical components of the procedure (5)**
PROC - Procedimiento de diagnóstico (diap):	**need for further surgery (2)**
CHEM - Química Orgánica/fármaco (orch):	**with 2-0 Polysorb (2)**
DEVI - Dispositivo médico (medd):	**held with a clamp (2)**
DEVI - Dispositivo médico (medd):	**drilled in a 4-0 starting point (4)**
PHEN - Fenomeno o proceso (phpr):	**a 6-degree 8 mm distal femoral cut. (17)**
ORGA - Organización cuidados de la salud (hcro):	**taken to the recovery room (2)**

****

Resumen Final:

Resumen del documento "Report17592":

RESULTADOS DE LABORATORIO O TEST: (PHEN)
- **a 6-degree 8 mm distal femoral cut. (17)**

ENFERMEDAD, SÍNDROME, TRASTORNO O PATOLOGÍA: (DISO)
- **the external tibial jig taking into account flexion, (9)**

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **the external tibial jig taking into account flexion, (9)**
- **The knee (8)**
- **followed by closure of the skin (7)**
- **sized for a 35 x 10-mm asymmetric patella. (5)**

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **a standard patellar cutting jig (8)**
- **scrubbed for the critical components of the procedure (5)**

GRUPO POBLACIÓN, GÉNERO, RAZA: (LIVB)
- **The patient's left leg (6)**

FUNCIONES ORGÁNICAS, MENTALES Y ATRIBUTOS MÉDICOS : (PHYS)
- **Once hemostasis (4)**
- **flexion past 125 degrees without liftoff (3)**

DISPOSITIVOS MÉDICO, DISPENSADOR DE FÁRMACOS: (DEVI)
- **drilled in a 4-0 starting point (4)**
- **held with a clamp (2)**

FÁRMACOS, QUIMICA ORGÁNICA E INORGÁNICA : (CHEM)
- **with 2-0 Polysorb (2)**

ORGANIZACIÓN PROFESIONAL Y SANITARIA: (ORGA)
- **taken to the recovery room (2)**

Report17721

- **Documento médico original con filtrado previo**

KNEE
RAD

XRAY KNEE 1 OR 2 VIEWS LEFT
Postop.

Two views left knee show total knee arthroplasty near anatomically aligned without periprostatic fracture or osteolysis. Effusion,
soft tissue air, drain, splint and dressing are present.

LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY ALIGNED WITHOUT
PERIPROSTHETIC FRACTURE. IMMEDIATE POSTOPERATIVE APPEARANCE.

- **Representación basada en conceptos, con eliminación de CUIs repetidos en cada frase y eliminación de tipos semánticos no relevantes.**

1 0 C0022742 bpoc KNEE
2 0 C0022745 bsoj KNEE
3 0 C1283838 blor KNEE
4 0 C1963703 blor KNEE
6 0 C0203273 diap XRAY KNEE 1
15 0 C0230432 bpoc left knee
16 0 C1279572 bpoc left knee
18 0 C0086511 topp total knee arthroplasty near anatomically
23 0 C0013687 patf Effusion,
25 0 C2317432 bdsu Effusion,
26 0 C0225317 tisu soft tissue air,
28 0 C0013103 topp drain,
29 0 C0180499 medd drain,
30 0 C0038009 medd splint
31 0 C0013119 medd dressing
32 0 C0278286 topp dressing
33 0 C0518459 fndg dressing
37 0 C0086511 topp LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY
42 0 C0241311 fndg IMMEDIATE POSTOPERATIVE APPEARANCE.

- **Distribución de Densidad de Conceptos (Densidad Cualitativa):**

C0086511 topp LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY (2)
C2317432 bdsu Effusion, (1)
C1963703 blor KNEE (1)
C1283838 blor KNEE (1)
C1279572 bpoc left knee (1)
C0518459 fndg dressing (1)
C0278286 topp dressing (1)
C0241311 fndg IMMEDIATE POSTOPERATIVE APPEARANCE. (1)
---------- Corte 2º Cuartil (50 %) ----------
C0230432 bpoc left knee (1)
C0225317 tisu soft tissue air, (1)
C0203273 diap XRAY KNEE 1 (1)
C0180499 medd drain, (1)
C0038009 medd splint (1)
C0022745 bsoj KNEE (1)
C0022742 bpoc KNEE (1)
C0013687 patf Effusion, (1)
C0013119 medd dressing (1)

"Report17721" (Densidad Cualitativa)
****
ANAT - Parte del cuerpo/Organo (bpoc): **left knee (2)**
ANAT - Parte del cuerpo/Organo (bpoc): **KNEE (3)**
ANAT - Substancia del cuerpo (bdsu): **Effusion, (2)**
PROC - Procedimiento terapeútico (topp): **LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY (2)**
PROC - Procedimiento terapeútico (topp): **dressing (3)**
****

- **Repetición de Frases (Densidad Cuantitativa):**

"Report17721" (Densidad Cualitativa + Cuantitativa)
****
ANAT - Parte del cuerpo/Organo (bpoc): **KNEE (3)**
ANAT - Parte del cuerpo/Organo (bpoc): **left knee (2)**
ANAT - Substancia del cuerpo (bdsu): **Effusion, (2) (derrame)**
PROC - Procedimiento terapeútico (topp): **dressing (3)**
PROC - Procedimiento terapeútico (topp): **LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY (2)**
****

Resumen Final:

Resumen del documento "Report17721":

LOCALIZACIÓN, ÓRGANO, TEJIDO O SUSTANCIA CORPORAL: (ANAT)
- **KNEE (3)**
- **left knee (2)**
- **Effusion, (2)**

PROCEDIMIENTOS TERAPÉUTICOS O DIAGNÓSTICOS: (PROC)
- **dressing (3)**
- **LEFT TOTAL KNEE ARTHROPLASTY NEAR ANATOMICALLY (2)**