Departamento de Ingeniería del Software e Inteligencia Artificial

FACULTAD DE INFORMÁTICA

UNIVERSIDAD COMPLUTENSE DE MADRID

GENERACIÓN AUTOMÁTICA DE RESÚMENES CON APOYO EN ONTOLOGÍAS APLICADA AL DOMINIO BIOMÉDICO

LAURA PLAZA MORALES

Madrid, Junio 2008

GENERACIÓN AUTOMÁTICA DE RESÚMENES CON APOYO EN ONTOLOGÍAS APLICADA AL DOMINIO BIOMÉDICO

LAURA PLAZA MORALES

Tesis propuesta como cumplimiento parcial
de los requisitos para el título de:

MÁSTER EN INVESTIGACIÓN INFORMÁTICA

Dirigida por :

PABLO GERVÁS GÓMEZ-NAVARRO

ALBERTO DÍAZ ESTEBAN

A mi madre, Mercedes

A mis hermanas, Natalia y María

Agradecimientos

Esta página es, sin duda alguna, la que más satisfacción me produce escribir. Con ella culmina un año de trabajo, y supone el placer del deber cumplido y la esperanza de haber aportado algo al conocimiento científico. Pero sobretodo, me brinda la oportunidad de dar las gracias a todas las personas sin las que esta tesis no habría sido posible, y a quienes, de uno u otro modo, han estado a mi lado durante su realización.

Ante todo, quiero dar las gracias a mis directores de tesis, el Dr. D. Pablo Gervás Gómez-Navarro y el Dr. D. Alberto Díaz Esteban, del Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA) de la Universidad Complutense de Madrid, por haberme guiado en la realización de este trabajo. Sus comentarios y sugerencias han sido inestimables. A Pablo, por haberme acogido en su grupo de investigación, NIL (Natural Interaction Based on Languaje), y por haber confiado en mí desde el principio. A Alberto, por el interés que ha demostrado por mi trabajo.

Por supuesto, también quiero dar las gracias al resto de componentes de NIL, por ofrecerme su ayuda y apoyo en todo momento.

A mi madre, a quien dedico esta tesis, gracias por animarme a perseguir mis sueños y por enseñarme a luchar por ellos. Sin ella, no me habría decidido a hacer este máster y a volver a la universidad. A mis hermanas, gracias por su paciencia y comprensión. Este trabajo nos ha obligado a estar menos tiempo juntas.

Finalmente, gracias a mis amigos y amigas por confiar en mi. En los momentos de desmotivación, gracias por animarme a continuar.

Resumen

En esta memoria de tesis se propone una arquitectura para la generación de resúmenes informativos monodocumento en un dominio específico: la biomedicina. La utilidad de estos resúmenes es indudable, en un campo en el que los profesionales han de estar continuamente al corriente de los nuevos avances científicos, pero a la vez necesitan economizar el tiempo que dedican a su formación. A lo largo de la exposición, se presenta un método de extracción de oraciones, basado en la teoría de redes complejas, que realiza un mapeo del texto a los conceptos de la ontología UMLS, y representa el documento y las oraciones como grafos. La selección de las oraciones se realiza a partir del grado de conexión de sus conceptos en el grafo del documento, utilizando para ello un algoritmo de agrupamiento basado en la conectividad. Se desarrolla un sistema que implementa el método propuesto y se muestran los resultados empíricos de la aplicación de distintas heurísticas para la selección de las oraciones del resumen. Se realiza una evaluación formal del sistema y se compara con otros que resuelven tareas similares. Los resultados de esta evaluación demuestran que la propuesta es útil para la creación de resúmenes muy similares en contenido a los creados por humanos. Finalmente, se identifican algunos problemas y líneas de trabajo futuras.

Abstract

In this thesis, a new approach to biomedical text summarization is presented. In recent years, the amount of online information has increased explosively. But as time is precious, efficient access to data has become necessary. This is especially crucial for physicians and biomedical researchers, since they have to consult constantly up-to-date and heterogeneous information according to their needs. In order to tackle this overload of information, text summarization can undoubtedly play a role.

We introduce an ontology-based extractive method for summarization. It is based on mapping the text to concepts in the ontology and representing the document and its sentences as graphs. To assess the importance of the sentences in the document, we compute the centrality of their concepts in the document graph. We have applied our approach to summarize scientific biomedical literature, taking advantages from free resources as UMLS. Empirical results and conclusions are presented. We also evaluate generated summaries using existing metrics and confirm that our methodology is promising. Finally, pending problems and future work are identified.

Tabla de Contenidos

Capítulo 1

Introducción

1. Motivación

En la era en la que vivimos, la creación, distribución y manipulación de la información forman parte fundamental de las actividades culturales y económicas. La cantidad de documentos electrónicos accesible desde cualquier lugar y cualquier dispositivo, crece de manera exponencial, pero el tiempo sigue siendo un recurso valioso y limitado. Gracias a los avances tecnológicos conseguidos a lo largo de las últimas décadas, el almacenamiento y el acceso a ellos ya no suponen un problema, y los esfuerzos actuales se dirigen a la investigación en sistemas de búsqueda y recuperación cada vez más eficaces.

Hoy día la documentación es en su mayoría digital. Las revistas científicas, los periódicos, las informaciones de las empresas e incluso los libros y documentos de archivos son accedidos cada vez más exclusivamente a través de la web. Es por ello que la inmensa mayoría de la investigación está dirigida a este medio, caracterizado por su continuo dinamismo y crecimiento, y por la heterogeneidad de la naturaleza de la información que contiene (texto, imágenes, vídeo, sonido, etc.).

Internet recuerda a la célebre, *Biblioteca de Babel*, de Borges. En el relato se especula sobre un universo compuesto de una biblioteca de todos los libros posibles, cuyos libros están arbitrariamente ordenados, o sin orden, y que preexiste al hombre y es infinita. En la web, la sobrecarga de información es tal

que su localización, organización y comprensión se ven muy limitadas. En este contexto, la generación automática de resúmenes, ya sean informativos o meramente indicativos, puede ser de gran utilidad.

Por otra parte, los recursos digitales en el campo de la medicina son muchos y muy variados, además de constituir la principal fuente de información tanto durante la formación como durante el ejercicio de la profesión. Por citar un ejemplo, MEDLINE, la mayor base de datos de bibliografía biomédica, dispone de más de 16 millones de artículos, y más de 10.000 nuevos se añaden diariamente. Por este motivo, muchos trabajos recientes exploran el uso de técnicas de procesamiento de lenguaje natural aplicadas al dominio biomédico, en busca de mecanismos que faciliten la búsqueda, comprensión y utilización de esta ingente cantidad de información.

El creciente interés por el desarrollo de sistemas de apoyo al acceso y tratamiento de este tipo de información, junto al importante presupuesto que tanto entidades empresariales como gubernamentales han destinado a éste propósito, ha contribuido a la formación de una amplia comunidad de investigadores y al nacimiento de una importante cantidad de proyectos. En España, el número de investigaciones en este campo se ha multiplicado en los últimos años. Por ejemplo, los proyectos SINAMED e ISIS, financiados por el Ministerio de Ciencia e Innovación y el Ministerio de Industria, exploran nuevas técnicas para la "mejora en el acceso a la información biomédica mediante la integración de generación de resúmenes, categorización automática de textos y ontologías".

Durante los últimos años, y como respuesta a la situación descrita, se han incrementado notablemente los recursos lingüísticos disponibles para el tratamiento información. Diccionarios, tesauros, bases de datos léxicas y grandes bases de conocimiento biomédico, muchos de ellos de disponibilidad pública, facilitan la construcción de sistemas de procesamiento de lenguaje y les confieren mayores posibilidades y garantías de éxito.

Aunque lejos de conseguir resultados excepcionales, actualmente es posible encontrar sistemas de generación de resúmenes completamente operativos, casi todos ellos en el ámbito de los artículos periodísticos en la Web, como el sistema NewsBlaster, de la Universidad de Columbia, o el

sistema SUMMARIST, de la Universidad de Southern California, donde además los documentos a resumir pueden estar redactados en distintos idiomas.

Todo lo expuesto justifica sin duda la realización de esta tesis, en la que se propone una arquitectura para la generación automática de resúmenes de documentos del campo de la medicina clínica. Para ello, se pretende adoptar un enfoque fundamentalmente semántico y lingüísticamente motivado, que haga un uso intensivo de conocimiento del dominio y de los recursos disponibles, y que recurra a técnicas de generación de lenguaje natural para la producción de los textos de los resúmenes. El trabajo que se presenta en esta memoria se centra en la fase de extracción de oraciones, aunque el objetivo a medio plazo será completar las fases de análisis y generación para reescribir el resumen final utilizando abstracción.

2. Descripción del problema

Según Sparck-Jones (1999), un resumen consiste en la transformación de un texto fuente, a través de la reducción de su contenido, ya sea por selección o por generalización de lo que es importante. La elaboración de un resumen debe abordarse teniendo en mente las características del texto a resumir, el propósito con el que se realiza el resumen y las propiedades, en forma y contenido, que se desea que tenga el resumen producido. Es decir, el contexto condiciona tanto el proceso de generación como el resumen resultante, y es posible identificar tres clases de *factores de contexto* que denomina, respectivamente, *factores de entrada*, *factores de propósito* y *factores de salida.*

Los *factores de entrada* se derivan de las características del texto a resumir, y se subdividen a su vez en *forma, especificidad* y *multiplicidad de la fuente.*

- La **forma** del documento se define en función de su *estructura, escala, medio* y *género.* La estructura hace referencia tanto a la organización explícita y generalmente marcada en el texto (diferentes secciones o apartados cubriendo el objetivo, método, datos, etc.) como a la organización que se encuentra en el discurso.

La escala se refiere al tamaño del resumen a producir e influye tanto en el factor de condensación o compresión como en la transformación de contenido necesaria. El medio hace referencia al tipo de lenguaje utilizado (telegráfico, prosa, periodístico, etc.). Por último, el género se refiere al estilo literario del documento (descriptivo, narrativo, etc.)

- La **especificidad** alude al nivel de especialización del texto, y en función de ella, un texto puede considerarse como *ordinario*, *especializado* o *restringido*.
- La **multiplicidad de la fuente** hace referencia al número de documentos que intervienen en la elaboración del resumen y al grado en que estos se relacionan entre sí.

Los *factores de propósito*, si bien son los más importantes, también son los más frecuentemente ignorados y se derivan del objetivo para el que se elabora el resumen. Sparck-Jones distingue dentro de este grupo entre *situación, audiencia* y *función.*

- La **situación** pretende discernir entre la generación de resúmenes en un contexto conocido a priori y su generación en un contexto variable o indeterminado.
- La **audiencia** hace referencia a la existencia o no de un público objetivo prototípico que comparte conocimiento, habilidades de lenguaje, formación, etc.
- La **función** a la que va destinado el resumen permite distinguir entre aquellos que se utilizan para ayudar al usuario decidir si el texto es de su interés, sustituir al documento original o pre visualizar un texto que se va a leer.

Por último, los *factores de salida* son aquellos que determinan las propiedades que ha de cumplir el texto generado como resumen. Se subdividen en *material o extensión, formato* y *estilo.*

- El **material o extensión** determina el grado en que el resumen ha de capturar el contenido presente en la fuente. Puede cubrir todo el contenido del texto original o estar diseñado para cubrir únicamente

algunos tipos de información. Estos últimos se denominan *resúmenes parciales*.

- En relación al **formato,** el resumen puede presentarse como un texto continuo o como una sucesión de apartados o secciones, reflejando así la estructura del documento.
- Finalmente, en función del **estilo**, un resumen puede ser *informativo*, *indicativo*, *crítico* o *agregativo*.

Partiendo del estudio de los factores anteriores, se han propuesto muchas taxonomías, que se distinguen unas de otras en las características del resumen tomadas en consideración. A continuación se enumeran algunas de las más aceptadas.

- Tradicionalmente, se ha distinguido entre *resúmenes indicativos*, que proporcionan al usuario una función de referencia para seleccionar documentos para una lectura más profunda, y *resúmenes informativos*, que cubren toda la información esencial de los textos de entrada, actuando como sustitutos de éstos. A veces, se habla también de *resúmenes críticos*, que evalúan el tema o contenido del texto de entrada, expresando el punto de vista de la persona que realiza el resumen.
- Otra clasificación frecuente es aquella que distingue entre *resúmenes adaptados al usuario* y *resúmenes genéricos,* en función de si tienen en cuenta las preferencias y /o características del usuario a la hora de seleccionar los contenidos del resumen y la forma de presentación. Mientras que los primeros se concentran en los temas que son de interés para el lector, los segundos reflejan el punto de vista del autor.
- Atendiendo a si el contenido del documento a resumir versa o no sobre una temática especializada, se puede distinguir entre resúmenes *generalistas* o *especializados.*
- En función del número de documentos que intervienen en la generación del resumen cabe hablar de resúmenes *monodocumento* y *multidocumento*.

- Los resúmenes pueden ser *monolinguës,* si procesan un texto escrito en un solo idioma, o *multilingües,* si el texto original está escrito en diferentes idiomas.
- Por último, dependiendo de la naturaleza del proceso de generación del resumen, conviene distinguir entre aquellos que se componen íntegramente por material (palabras, oraciones, etc.) reutilizado del texto de entrada, y aquellos que incluyen contenidos que no están presentes, al menos de manera explícita, en el documento original.

Para Sparck-Jones (1999), la idea de generar resúmenes genéricos y totalmente independientes del contexto es un *ignis fatuus*, y la investigación debe centrarse en cómo se deben analizar los factores anteriores para la interpretación del texto y la generación del resumen. Por ello, restringir el problema a un dominio concreto, la biomedicina, y un tipo de documentos específico, el artículo científico, sin duda reduce la complejidad del proceso y redunda en una mayor calidad de los resúmenes generados, al permitir establecer generalizaciones más precisas en cuanto a la estructura, el estilo y el contenido informativo.

Los textos médicos se caracterizan por su volumen y heterogeneidad. Pueden versar sobre medicina clínica, bioinformática, medicina pública, etc.; y presentarse en forma de artículo científico, registro médico, informes, imágenes de rayos X o vídeos. A la hora de generar automáticamente un resumen, es necesario considerar las particularidades del tipo de documento. Por ejemplo, mientras que los artículos científicos están compuestos principalmente por texto, los registros médicos electrónicos suelen contener información estructurada e imágenes. Pero aun restringiendo el ámbito de estudio al artículo científico, dependiendo de la publicación para la que haya sido redactado, su estructura, estilo y contenido pueden variar significativamente. No obstante, es posible identificar una serie de características comunes a todos ellos. En primer lugar, todos presentan un *Abstract* que resume el contenido del texto y que, habiendo sido redactado por el propio autor, puede resultar de utilidad para la evaluación del resumen generado automáticamente. El primer apartado casi siempre se trata de la *Introducción*, y consta de unas pocas líneas en las que se explica la estructura y el contenido del resto del documento. Le siguen varios apartados en los que

se desarrolla el problema, y se presentan los métodos y técnicas utilizados, los grupos de pacientes tratados, etc., y que constituyen el grueso del documento. Finalmente, uno o varios apartados de *Resultados y Conclusiones* resaltan las observaciones efectuadas con el método empleado y exponen la interpretación de los resultados y las opiniones del autor. Frecuentemente, estos artículos vienen acompañados de tablas y gráficos.

Una clasificación de alto nivel de los sistemas de generación de resúmenes es la que distingue entre aquellos que utilizan técnicas de extracción y aquellos que utilizan técnicas de abstracción. Aunque típicamente los humanos realizan resúmenes mediante abstracción, la mayor parte de la investigación hoy día sigue siendo en extracción, y en general, los resultados alcanzados son comparativamente mejores.

Los sistemas basados en extracción de oraciones realizan un análisis superficial del texto, y no van más allá del nivel sintáctico. Este enfoque presenta algunos problemas importantes, derivados de la no consideración de la estructura semántica del documento y de las relaciones existentes entre los términos que lo componen (sinonimia, hiperonimia, homonimia, coocurrencias o asociaciones semánticas). Para ilustrar alguno de estos problemas, consideremos las siguientes oraciones (Yoo et al., 2007).

1. *Cerebrovascular disorders during pregnancy results from any of three major mechanisms: arterial infarction, hemorrhage, or venous thrombosis*

2. *Central nervous system diseases during gestation results from any of three major mechanisms: arterial infarction, hemorrhage, or venous thrombosis*

Puesto que ambas secuencias contienen términos diferentes, la dificultad radica en determinar que ambas oraciones presentan una semántica común.

El método que se presenta trata de solventar este problema (Plaza et al., 2008). Para ello, se ha adoptado un enfoque basado en la representación del documento en forma de grafo, utilizando los conceptos de UMLS asociados a sus términos, extendidos con sus correspondientes hiperónimos y relaciones asociativas. A diferencia otros trabajos (Yoo et al., 2007; Erkan y Radev,

2004), que se centran en la construcción de clusters de oraciones para determinar los temas comunes en múltiples documentos, y en la identificación de las oraciones centrales de cada cluster, en este trabajo el algoritmo de agrupamiento es aplicado a la identificación de conjuntos de conceptos estrechamente relacionados, que delimitan los distintos subtemas que se tratan dentro de un texto, y cuya presencia en las oraciones del documento determinará su grado de relevancia.

El enfoque presenta dos características que lo hacen especialmente interesante. En primer lugar, se puede extender fácilmente para considerar múltiples fuentes a la hora de elaborar el resumen, aunque como se verá al estudiar el estado del arte, habría que solventar los problemas típicos de la generación de resúmenes multidocumento. En segundo lugar, el método desarrollado se puede utilizar directamente para generar resúmenes en otros dominios distintos a la biomedicina, siempre que se disponga de una ontología adecuada que formalice el conocimiento del dominio en cuestión.

3. Objetivos

El objetivo fundamental perseguido con la elaboración de este trabajo es proponer una arquitectura para la generación de resúmenes de documentos biomédicos, utilizando técnicas de extracción y haciendo un uso intensivo del conocimiento del dominio.

Además de este objetivo general, se persiguen los siguientes objetivos específicos:

- La integración de recursos léxicos y semánticos, como SNOMED o UMLS, en el proceso de generación de resumen.
- El diseño de un método e implementación de un sistema de extracción de oraciones para la generación automática de resúmenes de artículos biomédicos, basado en la representación del documento y de sus oraciones como grafos de conceptos y en el cálculo de conectividades.

- El estudio de la aplicabilidad de la teoría de grafos y de redes complejas en tareas de procesamiento de lenguaje natural, en general; y en generación automática de resúmenes, en particular.
- La evaluación de la calidad de los resúmenes generados con el método propuesto, y la definición del alcance y las limitaciones del mismo.

4. Estructura del documento

El contenido del presente documento ha sido estructurado tal y como se expone a continuación:

- A lo largo del **Capítulo 1: Introducción,** se presenta someramente el problema abordado y el contexto en el que se encuadra, a la vez que se establecen brevemente los principales objetivos perseguidos por el proyecto y la estructura y los contenidos de la documentación asociada al mismo.
- El **Capítulo 2: Generación Automática de Resúmenes** exponen en mayor detalle los principios de la generación automática de resúmenes, realizando una clasificación de los métodos y técnicas más utilizadas, e incidiendo en el estado actual de la cuestión.
- El **Capítulo 3: Recursos utilizados** introduce los distintos corpus y ontologías evaluados para su utilización en este trabajo, a la vez que se exponen los criterios de decisión y se presenta la selección final.
- En el **Capítulo 4: Herramientas software utilizadas** se especifican de forma detallada las distintas herramientas especializadas empleadas para el desarrollo del proyecto**.**
- En el **Capítulo 5: Método propuesto para la Generación Automática de Resúmenes** se describen las distintas etapas del algoritmo desarrollado para la resolución de la tarea, a la vez que se muestran los resultados de su aplicación sobre un artículo concreto.

- El **Capítulo 6: Evaluación** describe el proceso realizado para determinar la efectividad del algoritmo y la calidad de los resúmenes.
- El **Capítulo 7: Conclusiones y Trabajo Futuro** presenta las reflexiones sobre los resultados obtenidos; así como el grado de consecución de los objetivos marcados. Asimismo, se incluyen una serie de propuestas o posibles líneas de trabajo futuro.
- Completan el documento los **Anexos** y la **Bibliografía**.

Capítulo 2

Estado del Arte

El presente capítulo tiene como objetivo presentar al lector la tarea de Generación Automática de Resúmenes. Para ello, se realiza una revisión de los trabajos más relevantes en generación monodocumento y multidocumento, analizando la evolución de las técnicas desde los orígenes de esta disciplina, y destacando las fortalezas y debilidades de cada una de ellas.

1. Orígenes de la Generación Automática de Resúmenes

Las primeras investigaciones en generación automática de resúmenes datan de finales de la década de los 50 y comienzos de la década de los 60, de la mano de Luhn y Edmundson, dos autores de influencia decisiva en el desarrollo posterior de la disciplina. Como veremos en más detalle a lo largo del capítulo, los primeros trabajos se basaron en la extracción de oraciones del documento original para construir el resumen, si bien han sido diversas las aproximaciones adoptadas para el cálculo de la relevancia de las oraciones. Algunos enfoques utilizan la frecuencia en el documento de ciertas *palabras clave* (Luhn, 1958; Edmundson, 1969; Kupiec *et al.*, 1995; Teufel y Moens, 1997), otros evalúan la presencia de ciertas *expresiones o palabras indicadoras* (Edmundson, 1969), otros tienen en cuenta la posición de la oración en el documento (Brandow et al., 1995; Lin y Hovy, 1997), etc. Durante esta primera etapa, las limitaciones

impuestas por los recursos hardware disponibles y por la rigidez de los lenguajes de programación existentes frenaron el avance de las investigaciones;

y durante las dos décadas posteriores fueron pocas las aportaciones de cierta relevancia.

La mejora de los computadores, el aumento de la información en la web y los avances en ingeniería lingüística y procesamiento de lenguaje natural despertaron en la década de los 90 el interés por investigar en procesamiento automático de textos. Desde entonces, la dedicación de la comunidad científica a la generación automática de resúmenes ha ido en aumento y hoy son muchos los investigadores de renombre dedicados a ella. Sin embargo, a pesar de haber cosechado resultados alentadores, algunos problemas siguen sin ser resueltos.

Durante la última década, han aparecido nuevos enfoques que utilizan técnicas de abstracción, profundizando en la estructura del discurso y haciendo uso de complejas técnicas lingüísticas (Marcu, 1997,1999, 2000, 2001; Teufel y Moens, 2002; Alonso, 2005). No obstante, la mayor parte de la investigación hoy día continúa siendo en extracción.

2. Etapas en la Generación Automática de Resúmenes

En paralelo al rápido crecimiento de la investigación en el área, surgen un número importante de propuestas en cuanto a etapas en el proceso de elaboración del resumen se refiere. A continuación se detallan las más significativas.

- Hovy (2001) propone la división más popular, señalando tres fases en la realización de la tarea: durante la *identificación del tópico*, se identifica el tema central del documento origen, a través de la selección de las unidades (palabras, oraciones o párrafos) más importantes; posteriormente, en la *fase de interpretación*, se lleva a cabo un análisis y comprensión del texto más profundos, mediante el uso de sofisticadas herramientas y recursos lingüísticos, obteniéndose una representación intermedia de la información contenida en la fuente; finalmente, durante la *etapa de generación*,

el resultado de la etapa anterior se traduce al texto final del resumen, en un formato y lenguaje adecuado para su lectura por parte del usuario.

- Siguiendo a Endres-Niggemeyer (1998), una persona realiza un resumen en tres fases: *exploración del documento, identificación de contenidos relevantes o themes* y *construcción del resumen.*
- Según Sparck-Jones (1999), las tres fases en la elaboración del resumen son la fase de *análisis,* la fase de *transformación* y la fase de *síntesis.*
- Paice (1981) afirma que, cuando se acomete la tarea, se realizan las siguientes operaciones de condensación de la información: *selección, agregación* y *generalización*, operaciones que se pueden realizar a nivel de palabra, frase, proposición, oración o discurso.
- Finalmente, Mani (2001) divide el proceso de generación de un resumen en tres fases: *análisis, transformación o refinamiento* y *síntesis.* Durante la fase de *análisis,* se analiza el texto de entrada y se construye una representación interna del mismo. Durante la fase de *transformación,* se transforma la representación interna a una representación del resumen, y sólo es aplicable a procesos que generan resúmenes mediante abstracción. Por último, durante la fase de *síntesis,* la representación del resumen se traduce a lenguaje natural. (Figura 1)

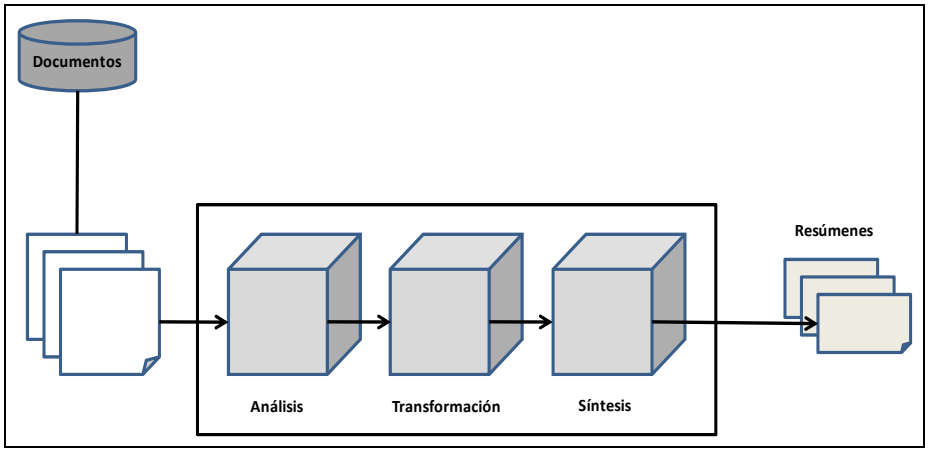

Figura 1 Arquitectura de un sistema de generación de resúmenes (Mani, 2001)

3. *Técnicas de Generación Automática de Resúmenes*

A pesar del amplio abanico de técnicas estudiadas en la generación de resúmenes, existe común acuerdo a la hora de clasificarlas en una de dos posibles categorías: *técnicas de extracción o enfoques superficiales* y *técnicas de abstracción o enfoques profundos* (Mani, 2001).

Las **técnicas de extracción** realizan un análisis superficial del texto fuente, y no van más allá del nivel sintáctico. Durante la fase de análisis, se limitan a la extracción de segmentos clave del texto mediante la utilización de criterios estadísticos, posicionales o palabras clave; y durante la fase de síntesis, se dedican a eliminar la incoherencia y la redundancia, e incluso a resolver referencias anafóricas. Se trata de un enfoque muy independiente del dominio, y en esta característica radica su principal ventaja. El inconveniente es que los resúmenes generados pueden resultar inconexos y de baja calidad en cuanto a la relevancia del contenido se refiere.

Las **técnicas de abstracción** realizan durante la fase de análisis, al menos a nivel de oración, una representación semántica del texto fuente, a partir de la interpretación del texto para identificar conceptos genéricos y relaciones entre conceptos, generalmente haciendo uso de alguna plantilla o esquema que de alguna forma marca la información que se considera importante de acuerdo con el contexto particular en el que se genera el resumen. La fase de síntesis implica el uso de generación de lenguaje natural a nivel semántico o discursivo. Su principal inconveniente es que son muy intensivas en conocimiento, por lo que sólo son aplicables a dominios muy específicos, aunque se pueden obtener resúmenes de mayor calidad que con las técnicas de extracción.

En función de los factores que intervienen en la construcción del resumen, revisados en el apartado 2 del primer capítulo, el paradigma más adecuado puede variar. A continuación, realiza un repaso de las técnicas concretas de mayor aceptación, según la clasificación presentada en (Mani, 2001), encuadrándolas en uno de los dos grupos anteriores y analizando sus ventajas e inconvenientes. Para cada una de ellas se citarán los trabajos más relevantes y los autores que más han contribuido a su desarrollo. Finalmente, y

a modo de conclusión, se realizará un análisis comparativo y se establecerán las que, según nuestra opinión, han de ser las directrices para el trabajo futuro.

3.1. Técnicas de Extracción

Extracción es el proceso de identificar la información importante en el texto. Los enfoques clásicos abordan la generación de resúmenes desde esta perspectiva, y aún hoy día, gran parte del trabajo continúa siendo en esta línea.

En sus inicios, la mayor parte de la investigación en extracción hacía uso de técnicas superficiales para identificar los segmentos relevantes en la fuente. El tamaño de estos segmentos o unidades puede variar de unas técnicas a otras. Aunque la mayoría trabajan a nivel de oraciones, algunas utilizan unidades más pequeñas, como sintagmas nominales o proposiciones, mientras que otras utilizan unidades mayores, como párrafos.

Sin embargo, la arquitectura general varía poco o nada de unos enfoques a otros. Hahn y Mani (2000) identifican dos fases en la generación de un extracto: *análisis* y *síntesis*. La mayor parte del trabajo se realiza durante la fase de análisis, si bien es bastante superficial. Típicamente, el texto de entrada se escanea, calculando para cada unidad (frase, oración o párrafo) un peso o puntuación indicativa de su importancia. Para ello, se computan un conjunto de características para cada oración, se normalizan y se suman. Durante la fase de síntesis, se extraen las frases mejor puntuadas y se construye el resumen mediante una simple concatenación de las mismas. Adicionalmente, algunos trabajos incluyen eliminación de redundancia y algún otro procesamiento para mejorar la coherencia.

La principal diferencia entre unas técnicas y otras radica en las características utilizadas por la función de peso. A groso modo, las métricas empleadas se pueden clasificar en *estadísticas,* que calculan la importancia de una oración en función de la frecuencia de aparición de ciertos términos considerados relevantes; *posicionales,* que tienen en cuenta la posición que ocupa la oración en el documento, y *lingüísticas,* que buscan ciertas expresiones o palabras indicativas.

Seguidamente se muestran algunas de las heurísticas más utilizadas para la selección de oraciones relevantes en el texto (Paice, 1990).

- **Frecuencia de palabras**

Se trata de un método muy simple e intuitivo que consiste en buscar ocurrencias de palabras y frases que se refieran al tema o temas centrales del documento. El método más utilizado toma el texto completo del documento, elimina las palabras comunes utilizando una *lista de parada (stop list)* y ordena las restantes palabras en una lista de frecuencias. Todas las palabras con una frecuencia menor a una preestablecida son eliminadas de la lista. Para cada oración del documento, se anotan las ocurrencias de sus palabras y, utilizando sus respectivas frecuencias, se aplica uno de los muchos métodos existentes para calcular una puntuación para la frase.

Los sistemas propuestos en (Luhn, 1958; Edmundson, 1969; Kupiec et. al, 1995; Hovy y Lin, 1999; Teufel y Moens, 1997), entre otros, utilizan diferentes medidas para el cálculo de las frecuencias de los términos: *tf* (*frecuencia del término*) y *tf.idf* (*frecuencia del término* multiplicada por la *inversa de la frecuencia del documento*).

Con respecto a la forma de puntuar la oración, algunos enfoques puntúan la aparición de grupos de palabras en la frase, con el objetivo de determinar oraciones importantes a partir de conceptos significativos (Luhn, 1958; Tombros y Sanderson, 1998), mientras que otros tienen en cuenta las apariciones de palabras individuales en la oración (Edmundson, 1969; Kupiec et al., 1995; Teufel y Moens, 1997).

Luhn (1958), precursor de la generación automática de resúmenes, describe una técnica simple y orientada a resúmenes de género específico, en la que utiliza las frecuencias de los términos para determinar las oraciones a extraer. Primero, realiza el filtrado del texto, eliminado pronombres, preposiciones y artículos. Después normaliza los términos agregando aquellos que son ortográficamente similares, contando el número de caracteres en los que difieren. Si la cuenta es inferior a seis, los dos términos se consideran de forma conjunta. Se anotan las frecuencias de estos términos en el documento, y

se eliminan los de baja frecuencia. Las oraciones se puntúan en función de la importancia de los términos que contiene. Cada oración se divide en segmentos separados por términos significativos que distan entre sí en no más de otros cuatro términos significativos. Cada segmento se puntúa como el cuadrado del número de términos significativos que contiene, dividido por el total de sus términos. La puntuación del mejor de los segmentos se toma como puntuación de la frase. Como posible mejora, plantea la posibilidad de extender el algoritmo premiando las palabras específicas del dominio.

- **Palabras clave en el título:**

Partiendo de la hipótesis de que los títulos, subtítulos y encabezados contienen los conceptos principales del documento, se extraen de ellos palabras claves, utilizando también una lista de parada para eliminar palabras comunes.

El primer trabajo en el que se aplica esta heurística es (Edmundson, 1969). En él se asignan pesos a las palabras significativas del título, subtítulo y encabezados, y se calcula la puntuación final de la frase como la suma de los pesos de los términos que incluye. En (Teufel y Moens, 1997) se puntúa cada frase dividiendo el número de palabras que también forman parte del título entre el total de términos de la frase.

- **Criterios posicionales**

Baxendale (1958) fue el primero en observar que, dentro de un párrafo, la primera oración es la que contiene información más relacionada con el tema del párrafo. Partiendo de esta idea, posteriores trabajos han estudiado la estructura de los textos para determinar distintas posiciones de las oraciones que indican, con una alta probabilidad, la presencia de información relevante (Kupiec et al, 1995; Teufel y Moens, 1997), como pueden ser oraciones bajo los encabezados de secciones como “Conclusión”, frases que ocupan los primeros y últimos párrafos del documento, etc. Sin embargo, las posiciones relevantes dependen en gran medida del tipo de documentos considerados.

- **Palabras clave**

Algunas palabras o sintagmas dentro de una oración, aunque no sean en sí mismas palabras clave, dan una indicación de si la oración trata con información clave. Los trabajos de (Edmundson, 1969; Rush et al., 1971) son los ejemplos más antiguos de esta técnica. Edmundson (1969) utiliza un corpus de entrenamiento para extraer las palabras significativas, clasificándolas en palabras *bonus* y palabras *stigma.* Las palabras bonus, muy frecuentes en el corpus, indicaban contenido importante del texto, mientras que las palabras stigma indicaban contenido irrelevante. Rush (1971) utiliza un método muy similar en el que considera también expresiones cortas. Kupiec (1995) también utilizan grupos de palabras en lugar de palabras individuales ("en conclusión").

- **Frases indicadoras**

Es un enfoque muy similar al anterior, en el que se localizan expresiones que generalmente acompañan afirmaciones explícitas sobre el texto. Paice (1981) intenta localizar en el texto *plantillas* del estilo "En el presente artículo se describe un método…". El principal problema es que, de nuevo, los términos considerados son muy dependientes del tipo de documento concreto. Además, mientras que algunos documentos, especialmente los discursivos, pueden contener algunas frases indicadoras, otros pueden no contener ninguna.

Merece especial atención el trabajo de Edmundson, ya que define el marco en el que se desarrolla la mayor parte del trabajo posterior en extracción. En su experimentación, utilizó cuatro métodos de extracción diferentes: posición, palabras indicadoras, palabras clave en el título, y frecuencia. Los mejores resultados se obtuvieron combinando los tres primeros métodos, utilizando una función lineal con los coeficientes apropiados para obtener la puntuación total. Sin embargo, la mejor característica individual resultó ser la posición de las oraciones en el texto. Estos resultados se deben sin duda al tipo de documentos que utilizó, artículos científicos, pero no se pueden generalizar a cualquier tipo de documentos. No obstante, en general, de las características

estudiadas, la posición y las palabras clave parecen ser las más efectivas en la tarea de extracción, aunque otras características de propósito específico también pueden ser muy útiles en contextos acotados.

Los enfoques actuales utilizan técnicas más sofisticadas para decidir qué oraciones extraer. En los últimos años, el uso de algoritmos de aprendizaje automático para determinar el conjunto de atributos que mejor se comportan en la extracción de oraciones ha alcanzado una cierta popularidad. Para ello, se necesita disponer de un corpus de textos junto con sus resúmenes generados de forma manual, que permitan aprender automáticamente los pesos de las distintas métricas. Evidentemente, esta técnica es muy dependiente del corpus.

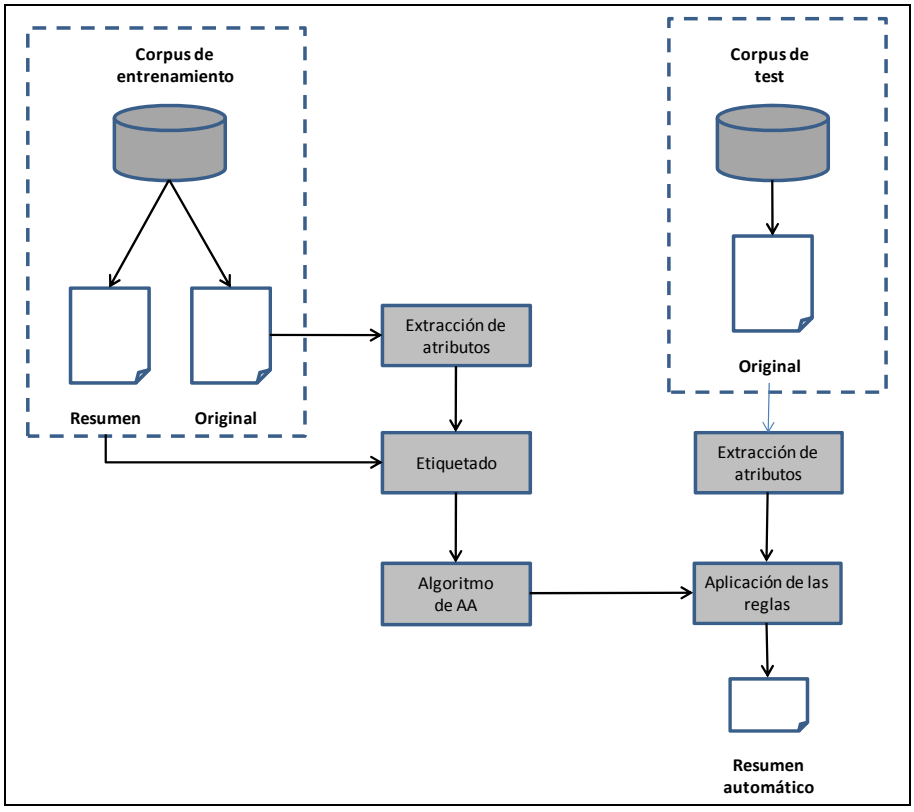

Figura 2 Arquitectura de un sistema de extracción con técnicas de Aprendizaje Automático

La aplicación de aprendizaje automático a la generación de resúmenes se aborda por primera vez en (Kupiec et al., 1995), donde se utiliza un clasificador bayesiano para determinar la combinación óptima de las métricas a considerar. Las características utilizadas en los experimentos son la longitud de

las oraciones, las frases clave, la posición de las oraciones y la presencia de palabras de alta frecuencia y nombres propios. Lin y Hovy (1997) introducen un nuevo método que, entrenando sobre un corpus de documentos, identifica la localización de la información relevante, generando una lista de posiciones que contienen las palabras más estrechamente relacionadas con el tema. Chuang y Yang (2000) utilizan 23 atributos para representar las oraciones, y aplican sobre ellos distintos algoritmos de aprendizaje. De estos 23 atributos, los primeros cinco se denominan *atributos no estructurales* (frecuencias, similitud con el título, etc.). El segundo grupo, formado por atributos dependientes del dominio y el lenguaje, se denominan *relaciones retóricas*. Neto y colegas (2002) emplean 13 atributos, de los cuales cuatro son independientes del dominio (centroide de las oraciones, longitud y posición, similitud con el título y presencia de nombres propios), mientras que el resto dependen de conocimiento externo. Zhou, Ticrea y Hovy (2005) construyen un sistema para la generación de resúmenes de biografías, en el que utilizan categorización de oraciones y otras ideas procedentes de la recuperación de información. En una primera fase, las oraciones son clasificadas en una de diez categorías en función del tipo de información que aportan ("bio", "fama", "nacionalidad", etc.) En la segunda fase, se clasifican en una de dos posibles categorías, dependiendo de si deben o no aparecer en el resumen. Aone (1999) describe el sistema *DimSum*, que utiliza técnicas estadísticas de análisis de corpus, junto con otras tecnologías robustas de procesamiento de lenguaje natural para extraer conocimiento. Aunque sigue el paradigma de la extracción de oraciones, contempla un conjunto de características lingüísticas que, una vez generadas, son combinadas automáticamente, utilizando un modelo bayesiano, de acuerdo con los requerimientos del usuario. En la extracción de las diferentes características lingüísticas, utiliza tanto estadísticas sobre el texto (frecuencia de los términos) como estadísticas sobre un corpus (inversa de la frecuencia del documento). En la adquisición de conocimiento de dominio, emplea un gran corpus y añade este conocimiento a las características para el resumen, de tres formas: calculando los valores *idf* para unas palabras claves seleccionadas, derivando colocaciones estadísticamente, y creando un índice de asociación de palabras. Utiliza dos mecanismos de cohesión: referencia y cohesión léxica a través de las reiteraciones de aliases de nombres, sinónimos y variaciones morfológicas. A la hora de contar las ocurrencias de los términos, agrega también las ocurrencias de sus sinónimos, considerándolas como

ocurrencias de un mismo concepto. A la hora de contar los nombres propios, maneja también sus aliases, tratando con referencias a nombres en lugar de con menciones.

Las técnicas de extracción presentan la ventaja de su relativa sencillez y bajo coste, ya que apenas hacen uso de conocimiento lingüístico. Además, son independientes del dominio. Sin embargo, se debe observar que no son apropiadas para todos los tipos de resúmenes. En primer lugar, si se trata de resumir textos muy extensos, el ratio de comprensión que se necesita es muy elevado, y no puede ser alcanzado con éxito sin utilizar abstracción. En segundo lugar, cuando se realizan resúmenes multidocumento, es necesario resaltar tanto las similitudes como las diferencias entre los documentos, tarea imposible utilizando extracción. Por último, los humanos generan abstractos, no extractos, y para ello emplean técnicas de generalización y especialización. Además, el resumen resultante puede resultar incoherente. Mani (2001) señala tres posibles problemas:

- **Anáforas pendientes:** si en el resumen generado se incluye una anáfora pero no la entidad referenciada, el resumen puede ser ilegible.
- **Huecos:** si del texto original, donde se supone que todas las frases están conectadas entre sí, se eliminan algunas de ellas, el resultado puede ser un texto incoherente.
- **Entornos estructurados:** si en lugar de texto, se trata de resumir listas, tablas, etc. la supresión de información debe realizarse más cuidadosamente para no producir un resumen sin sentido.

Sin embargo, distintos estudios empíricos justifican el uso del método de extracción. Kupiec (1995) realiza una evaluación con la que demuestra que aproximadamente el 80% de las frases incluidas en resúmenes manuales aparecen tal cual o con pequeñas modificaciones en el texto original. Por otra parte, en (Morris et al., 1992) se lleva a cabo un experimento en el que se solicita a un conjunto de jueces que respondan a varias preguntas con cinco posibles respuestas, basándose en resúmenes automáticos generados por extracción de oraciones y en resúmenes manuales, no encontrándose

diferencias significativas en los resultados obtenidos con uno y otro tipo de resúmenes.

Las deficiencias identificadas pueden mejorarse si, una vez construida lo que llamaremos una primera versión del resumen, éste se somete a un proceso de revisión, de la misma manera que los humanos revisamos nuestros resúmenes para mejorar su coherencia, fluidez y concisión. Durante el proceso de revisión se puede, por ejemplo, compactar oraciones excesivamente largas utilizando técnicas de eliminación; se pueden realizar algunas sustituciones léxicas, o incluso aplicar algunas operaciones de generalización y abstracción.

A continuación, se reproduce una tabla tomada de (Mani, 2001) en la que, para un conjunto de problemas de coherencia observados empíricamente, se proponen posibles soluciones (Tabla 1).

Problema	Solución	Conocimiento Necesario
Falta de conjunciones entre oraciones	Añadir o eliminar conjunciones	Estructura del discurso
Falta de participios adverbiales	Añadir o eliminar participios adverbiales	Estructura sintáctica
Oraciones de sintaxis compleja	Dividir oraciones en otras más simples	Estructura sintáctica
Repetición, por ejemplo, de nombres propios	Pronominalizar, omitir expresiones o añadir demostrativos	Estructura sintáctica
Falta de información, por ejemplo, anáforas pendientes o nombres propios sin descripciones adecuadas	Reemplazar las anáforas por sus antecedentes, eliminar anáforas o añadir información descriptiva	Resolución de anáforas y elipsis Extracción de información

Tabla 1 Problemas de coherencia (Nanba y Okumura, 2000)

Para solucionar estos problemas de coherencia y desinformación, se proponen tres tipos de operaciones de revisión. Sin embargo, a pesar de que con el uso de estas técnicas se puede mejorar la calidad de los extractos, ninguna de las soluciones propuestas incluye semántica, ni realiza ningún tipo de abstracción o generalización.

3.2. Técnicas basadas en la Estructura del Discurso

Los enfoques recientes hacen uso cada vez más de un sofisticado análisis del lenguaje natural para identificar el contenido relevante en el documento, y para ello analizan las relaciones entre palabras o la estructura del discurso. Numerosos estudios a cerca del comportamiento de los profesionales en generación de resúmenes indican que, sin lugar a dudas, a la hora de enfrentarse a la tarea crean un modelo mental de lo que esperan que sea la estructura del documento, en función del género del que se trate (tema o *theme*).

El primer grupo de técnicas estudiado realiza un **análisis de la cohesión** del documento. Halliday y Hasan, (1996) definen la cohesión textual en términos de las relaciones entre palabras, sentidos de palabras o expresiones referidas, que determinan cómo de estrechamente conectado está el texto. Distinguen entre *cohesión gramatical*, refiriéndose a ciertas relaciones lingüísticas como la anáfora, la elipsis y la conjunción; y *cohesión léxica*, refiriéndose a relaciones como la reiteración, la sinonimia y la homonimia, pudiéndose combinar entre sí ambos tipos de relaciones. Halliday (1978), Mann y Thompson (1988) y Van Dijk (1988) coinciden en que la coherencia textual representa la estructura general o superestructura de un texto, visto como un conjunto de oraciones, y en términos de las relaciones de alto nivel que se establecen entre sus proposiciones u oraciones. La distinción de ambas propiedades resulta de gran utilidad a la hora de dividir el espacio de técnicas de generación de resúmenes que utilizan información sobre el discurso. El tipo de estructura discursiva que emerge de la cohesión tiene que ver con patrones de significación, mientras que la coherencia nos lleva a hablar del concepto de *theme,* y tiene que ver con patrones de razonamiento presentes en el texto. Algunas investigaciones utilizan el grado de conectividad léxica entre los distintos fragmentos y el resto del texto. La conectividad se puede medir como el número de palabras que comparte, sinónimos o anáforas. El trabajo de Hearst (1997) compara bloques de texto adyacentes, en función del solapamiento de vocabulario, para identificar las fronteras temáticas.

Otras aproximaciones de gran interés utilizan la noción de *cadenas léxicas*. Morris y Hirst (1991) las definen como una secuencia de palabras interrelacionadas que abarcan un tópico del texto. En (Barzilay y Elhadad, 1999), se presenta una solución que, sin hacer uso de interpretación semántica compleja, produce un resumen identificando en el texto fuente secuencias de términos agrupados mediante relaciones de cohesión textual: repetición, sinonimia, hiperonimia, antonimia y holonimia. Para establecer las relaciones, utiliza el tesauro *WordNet*, tratando el problema de la polisemia mediante la creación de cadenas alternativas para los distintos posibles significados y la elección de la mejor cadena en función del número de relaciones y sus pesos en la cadena. Los nodos de la cadena pueden ser nombres simples o compuestos. Las cadenas se construyen inicialmente para segmentos individuales de texto, y luego se combinan entre sí aquellas que comparten un mismo término con un mismo sentido en WordNet. Otros enfoques (Cardie y Buckley, 1997) identifican pasajes que contienen ciertas palabras que se sabe que están correlacionadas con el tema central del texto fuente.

La aproximación más aceptada para la representación de la cohesión textual son los llamados *grafos de cohesión.* Como ya hemos comentado, dentro de un documento, las palabras y oraciones se encuentran conectadas entre sí por medio de distintos tipos de relaciones, indicadoras de la cohesión entre sus elementos textuales. Estas relaciones entre los distintos elementos de un texto se pueden representar en una estructura de grafo, en el que los vértices son las oraciones y los arcos representan las relaciones significativas entre ellas. Skorokhod'ko (1972) propone un método de extracción de oraciones que incluye la construcción de una estructura semántica para el documento utilizando un grafo de este tipo, en el que los arcos definen relaciones de repetición, hiponimia, sinonimia o referencias a palabras relevantes. La idea subyacente es que las oraciones más significativas son aquellas que están relacionadas con un mayor número de otras oraciones y son las primeras candidatas a la extracción. Mani (2001) presenta esta misma idea con el nombre de *Suposición de la Conexión de un grafo (Graph Connectivity Assumption).* En (Salton et al., 1997), las unidades consideradas como nodos del grafo son párrafos en lugar de oraciones, y las relaciones indican la similitud de palabras entre párrafos. En (Mani y Bloedorn, 1999), se propone un modelo de grafo más sofisticado. Los nodos son palabras, mientras que los

arcos establecen relaciones de proximidad, repetición, sinonimia, hiponimia y referencias a entidades. La selección de una oración como candidata al resumen se hace en función del peso de sus términos. Para ello, se actúa recorriendo el grafo y asignando los pesos a los distintos nodos en función de las relaciones que se atraviesan hasta llegar a él.

Por último, cabe hablar de las técnicas que realizan un **análisis de la coherencia**, para identificar las macro relaciones que implícitamente existen en todo texto. Han sido muchas las teorías propuestas para el análisis de la estructura argumentativa: la *Rethorical Structure Theory* (Mann y Thompson, 1988), las *Gramáticas Discursivas* (Longacre, 1979), las *Macroestructuras* (Van Dijk, 1988) o las *Relaciones de Coherencia* (Hobbs, 1985). En nuestra exposición, no obstante, sólo comentaremos la teoría de Mann y Thompson, por ser la de mayor difusión y aplicación en la generación de resúmenes.

La *Rethorical Structure Theory (RST)* ha gozado, desde que fuera propuesta en la década de los ochenta, de una gran aceptación académica, y ha sido aplicada a la resolución de muchas tareas dentro de la computación lingüística, en concreto, a la generación de resúmenes. Proporciona un análisis de la argumentación de los textos, dirigiendo la organización del discurso a través de las relaciones que se establecen entre las distintas partes del texto. Una de las aportaciones más interesantes de la teoría es la definición del concepto de *relación retórica*, para referirse a un tipo de relación que se establece entre dos segmentos de texto a los que denomina *núcleo* y *satélite*. El núcleo contiene información que es central en el documento, mientras que el satélite completa al núcleo. Las relaciones en la RST se definen en términos de cuatro campos: restricciones sobre el núcleo, restricciones sobre el satélite, restricciones en la combinación del núcleo y el satélite y efecto sobre el texto. Marcu (1999, 2000) aplica esta idea a la generación de resúmenes, en concreto, durante la fase de análisis, para construir un árbol representando la estructura retórica del texto, y posteriormente utiliza este árbol para calcular la relevancia de los términos que actúan como nodos del árbol y que permiten la composición de resúmenes a distintos niveles de detalle.

La siguiente tabla sintetiza las fortalezas y debilidades de las dos aproximaciones a nivel de discurso.

Método	Características	Fortalezas	Debilidades
Generación basada en Cohesión	Uso de grafos de relaciones entre los elementos textuales Definición de relaciones a distintos niveles Análisis a nivel léxico	Muy general e intuitivo La cohesión mejora la legibilidad del texto Permite nivelar la topología del grafo para determinar la relevancia	La idea es difícil de precisar Requiere mecanismos complejos de desambiguación y resolución de anáforas Se centra en la selección para generar extractos
Generación basada en Coherencia	Uso de árboles para representar la estructura retórica del texto El análisis discursivo se basa en pruebas léxicas (oraciones indicativas) y análisis sintáctico Uso de relaciones núcleo-satélite como criterio para determinar la relevancia	Muy general e intuitivo La importancia de las unidades discursivos elementales puede calcularse en función de la profundidad en el árbol La coherencia de los textos puede mejorar en función de la estructura retórica	Demasiado detallado En la práctica, los profesionales en generación de resúmenes no construyen una estructura detallada del discurso No garantiza la coherencia de los textos Anotar la estructura retórica de un texto continúa siendo un desafío

Tabla 2 Comparación entre aproximaciones a nivel de discurso (Mani, 2001)

3.3. Técnicas de Abstracción

Citando a Radev (2002), *extracción* es el proceso de identificar contenido importante en el texto, *abstracción* es el proceso de reformularlo en otros términos, *fusión* es el proceso de combinar porciones extraídas y *compresión* es el proceso de ignorar información irrelevante. En base a esta definición, clasifica dentro de la abstracción cualquier enfoque que no utilice extracción.

Dentro de las investigaciones que utilizan técnicas de abstracción se pueden distinguir dos líneas bien diferenciadas: aquellas que utilizan extracción de información y las que utilizan compresión. De hecho, aunque los primeros trabajos se incluían fundamentalmente en el primer grupo, los avances en las técnicas de generación de lenguaje están desviando progresivamente el interés hacia el segundo grupo.

Los **enfoques basados en extracción de información** recorren el texto buscando un conjunto de información predefinida para incluir en el resumen. Este método, a pesar de producir resúmenes muy legibles y de alta calidad, tiene validez únicamente en dominios muy restringidos.

Una técnica muy popular consiste en la utilización de *plantillas*, que recogen la información que se considera relevante para una categoría predefinida de textos. El documento fuente se analiza para extraer la información necesaria para rellenar los campos de la plantilla, y dicha información se puede utilizar en una fase posterior para generar el resumen en lenguaje natural. Un ejemplo del uso de plantillas para la generación de resúmenes es el sistema FRUMP (DeJong, 1982), aplicado sobre artículos periodísticos de cincuenta dominios diferentes, definiendo una plantilla o *guión* distinto para cada posible tipo de artículo. La principal debilidad de este enfoque es su dificultad para extenderlo a nuevas situaciones. Radev y McKeown (1998) utilizan plantillas para obtener resúmenes multidocumento de artículos periodísticos sobre ataques terroristas. Será estudiado con más detalle en el apartado dedicado a la generación de resúmenes multidocumento.

Algunos enfoques actuales combinan la extracción de plantillas con técnicas de análisis estadístico. Paice y Jones (1993) parten de la convicción de que es imposible realizar un resumen de calidad de manera totalmente independiente del contexto. En su sistema, combinan técnicas de indexado con técnicas de abstracción, y utilizan estructuras semánticas para organizar el contenido de documentos de investigación en el dominio específico de los cultivos agrícolas. Este tipo de documentos se caracterizan por ser muy estructurados y en ellos se pueden observar una organización, estilismo y semántica relativamente constante. Un trabajo similar se presenta en (Rau et al., 1989) y su sistema SCISOR, en el que los conceptos involucrados establecen interrelaciones más complejas para configurar una red de conceptos.

Trabajos posteriores utilizan métodos más sofisticados para rellenar las plantillas, en conjunción con análisis estadístico y aprendizaje automático de patrones a partir de corpus anotados (Mikheev, 1998; Riloff y Jones, 1999).

La utilización de plantillas, sin embargo, no permite ningún grado de interpretación de la información extraída. Una alternativa es el uso de *jerarquías de conceptos.* Si se dispone de una base de conocimiento del dominio, la generación del resumen se puede abordar como un proceso de abstracción sobre la base de conocimiento. Hanh y Reimer (1999) consideran la realización de resúmenes como una transformación basada en operador, que toma la salida de un analizador de lenguaje natural y crea estructuras abstractas de conocimiento. Su sistema TOPIC utiliza los conceptos de relevancia y generalización para crear una estructura jerárquica denominada *grafo del texto.* En este grafo, los nodos hoja corresponden a conceptos más específicos, y conforme se asciende en el árbol, se generaliza sobre los mismos. Los autores ilustran el funcionamiento de TOPIC en el dominio de los informes legales y tecnológicos en alemán. Ofrece un amplio rango de parámetros que se pueden configurar para generar resúmenes a distintos niveles de detalle. Debido al elevado coste que supone disponer de una base de conocimiento configurada *ad hoc* para el dominio de los textos, algunos enfoques hacen uso de tesauros de propósito general y uso público como WordNet. Hovy y Lin (1999) lo utilizan, entre otras cosas, para realizar la generalización. En su sistema SUMMARIST, el recuento de conceptos se realiza de manera que, cuando una palabra aparece en el texto, tanto ella como todos sus conceptos asociados en WordNet reciben la correspondiente puntuación, de manera que los pesos se propagan a través de WordNet. La puntuación de un concepto se computa como la suma de su frecuencia y el peso de todos sus hijos en la jerarquía. Sin embargo, la ausencia de conocimiento específico de dominios particulares limita la capacidad de generalización.

Una aproximación radicalmente distinta a la vista hasta el momento consiste en utilizar las características de los eventos para producir resúmenes. Según Alterman y Bookman (1990) existen semántica y relaciones temporales entre los eventos de una historia que deben respetarse para mantener la coherencia de la historia. Para ello, construyen, manualmente y para cada dominio, el llamado *grafo de conectividad de eventos,* desarrollado por

Alterman (1985), que consiste en un grafo dirigido cuyos nodos son eventos y sus arcos, relaciones de precedencia o pertenencia entre ellos. Nótese que, a pesar de las similitudes con los *grafos de cohesión*, existen varias diferencias. Por un lado, los nodos son eventos en lugar de conceptos; en segundo lugar, las relaciones que se establecen son semánticas, en lugar de léxicas o sintácticas.

Los **enfoques basados en compresión** abordan el problema desde el punto de vista de la generación de lenguaje, e incluyen operaciones de selección, agregación y generalización realizadas a la hora de reescribir el resumen. Mani (2001) introduce una aproximación al problema a la que denomina *reescritura de texto*. Partiendo de una representación semántica de las oraciones, en términos de expresiones lógicas sobre conjuntos de términos, los términos son individual o colectivamente seleccionados, agregados o generalizados para producir abstractos. El sistema *SUSY* (Fum et al., 1985) es un buen ejemplo de esta técnica. Aplicado al dominio de los artículos técnicos sobre sistemas informáticos, utiliza una pequeña base de conocimiento con unos treinta conceptos. Cada oración es representada mediante una lista de términos lógicos y su importancia se calcula en función de un conjunto de reglas de relevancia. Un concepto se considera de alta relevancia si el número de referencias en la representación semántica de todas las oraciones supera un umbral preestablecido. Es lo que se denomina *Suposición de Conceptos Altamente Referenciados (Highly Referenced Concept Assumption*). Witbrock y Mittal (1999) extraen del documento original un conjunto de palabras que luego ordenan en oraciones utilizando un modelo de lenguaje basado en bigramas. Algunos trabajos, aunque utilizan técnicas de extracción para localizar oraciones relevantes en la fuente, aplican posteriormente un proceso de reducción y regeneración para reescribir el resumen (Jing y McKeown, 1999; Knight y Marcu, 2000). McKeown et al. (1995) ilustran el uso de ciertas expresiones lingüísticas para empaquetar el texto de forma que se consiga comunicar la mayor información en el menor espacio posible. Para ello, aplican distintas operaciones de eliminación, como el borrado de repeticiones, y otras operaciones de agregación. Aplican estas ideas a dos dominios distintos: *STREAK*, que genera resúmenes de partidos de baloncesto y *PLANDOC*, que resume la actividad planificada de una red. Marcu y Knight (2000) utilizan compresión a nivel de oración para conseguir capturar información y gramaticalidad. Para ello, ensayan con dos modelos, uno

determinista y otro probabilístico que, actuando sobre pares <*texto, resumen*>, permiten determinar qué es importante en una oración y cómo comunicar la información utilizando sólo unas pocas palabras.

Sin embargo, la construcción de un abstracto implica un uso más intensivo en conocimiento, dada la restricción de que el material de la fuente no debe mencionarse explícitamente en el resumen. Los enfoques más recientes avanzan en esta dirección y hacen uso de recursos externos como ontologías y bases de conocimiento, y son capaces de realizar inferencias.

4. Generación de Resúmenes basada en Grafos

El propósito de este apartado es realizar un repaso de los métodos y aplicaciones de generación de resúmenes que, al igual que el enfoque que se propone en este trabajo, utilizan algoritmos basados en grafos para representar la estructura de los documentos y elaborar el resumen.

4.1. Teoría de Grafos en Procesamiento de Lenguaje Natural

Las técnicas de modelado de texto basado en grafos se asientan sobre los principios de las *redes complejas,* un área de investigación que surge de la intersección entre la teoría de grafos y la estadística, y que ha sido objeto de mucha atención en los últimos años. Aunque sus orígenes se remontan al siglo XVIII, con el trabajo de Leonhard Euler y la solución del problema de los puentes de Köningsberg, es el trabajo de Erdös y Rényi (1959), dos siglos después, el que establece las bases de esta teoría tal y como se concibe en la actualidad. Según estos autores, la formación de las redes reales se puede explicar mediante la llamada *teoría aleatoria de grafos,* según la cuál, las redes en el mundo real presentan una distribución aleatoria, y en ellas, casi todos los nodos tienen un número similar de conexiones. Aunque la validez de esta teoría fue aceptada durante varias décadas, trabajos más recientes han demostrado que muchas de las redes reales no son aleatorias. Las denominadas *redes del mundo pequeño (small-world networks*) modelan un tipo especial de

redes en las que es posible alcanzar cualquier nodo a través de un número relativamente pequeño de otros nodos, y tienen una alta tendencia a formar grupos locales de nodos interconectados (Watts y Strogatz, 1998). Barabási y Albert (1999) introducen otro tipo de redes complejas, denominadas *redes libres de escala (scale-free networks),* en las que la distribución de las conexiones no es uniforme, sino que presentan un pequeño número de nodos (denominados *hubs)* con un gran número de conexiones, mientras que el resto de nodos están muy poco conectados. Este tipo de redes, por ser el que se utiliza en nuestra investigación, será tratado en más detalle en el apartado dedicado a la explicación del método propuesto. La investigación actual en redes complejas incorpora ideas de la mecánica estadística, y trata de caracterizar las redes en términos de su estructura y dinamismo.

Durante los últimos años, se han multiplicado los trabajos en los que los principios de la teoría de grafos se aplican a resolver problemas de procesamiento de lenguaje natural. Investigaciones recientes demuestran que, a través de la representación del texto como un grafo, se pueden alcanzar soluciones eficientes para una amplia variedad de tareas, tan diversas como la desambiguación semántica, la extracción de palabras clave, la categorización de textos, la construcción de tesauros, la recuperación de pasajes, la extracción de información, la generación de resúmenes o la clasificación de sentimientos. En (Lin, 1998; Lee, 1999), el problema de la desambiguación semántica se aborda utilizando información derivada de diccionarios y redes semánticas para construir grafos, en los que la similitud se establece entre pares de conceptos o entre los conceptos y el contexto que los rodea. Antequeira y colegas (2007) demuestran que las redes complejas que representan los documentos pueden capturar ciertas características del estilo del autor, de manera que pueden utilizarse en tareas de identificación de autoría. Jannink y Wiederhold (1999) utilizan diccionarios *online* para derivar una estructura de grafo que luego utilizan para generar tesauros automáticamente, consiguiendo resultados comparables a otros tesauros construidos manualmente, como WordNet. En (Otterbacher et al., 2005), se utiliza un algoritmo de aprendizaje semi supervisado para la recuperación de pasajes. La idea es propagar información

desde nodos etiquetados hasta nodos sin etiquetar, a través de las conexiones del grafo.

4.2. Teoría de Grafos en Generación Automática de Resúmenes

En el apartado dedicado a las técnicas de generación de resúmenes se han revisado distintos enfoques que utilizan grafos para representar las unidades lingüísticas del documento, ya sea para asegurar la coherencia del resumen o para analizar su cohesión, y que se clasifican en lo que hemos denominado técnicas a nivel del discurso. Dentro de las aproximaciones de generación basada en la cohesión, se han estudiado trabajos que utilizan *cadenas léxicas* y *grafos de cohesión;* mientras que dentro de las aproximaciones basadas en la coherencia se ha presentado, por su importancia, la *Teoría de la Estructura Retórica.* Tanto las relaciones de coherencia como las de cohesión pueden utilizarse para determinar la relevancia de las oraciones para su selección.

En este apartado, nos centraremos en el estudio de métodos de extracción que utilizan el concepto de *centralidad (centrality)* para capturar los términos o conceptos "centrales" en un documento o cluster de documentos. Típicamente, estos enfoques representan el texto como una red compleja. En ella, los nodos representan cada una de las unidades textuales en las que se divida el texto, que dependiendo de la aplicación, pueden variar desde palabras u oraciones hasta párrafos o incluso documentos. Por su parte, las aristas representan algún tipo de relación entre estas unidades, que pueden ser de naturaleza semántica o léxica.

Muchos de los métodos para el cálculo de la centralidad están basados en *PageRank,* el algoritmo desarrollado por la Universidad de Stanford y utilizado por Google para calcular la relevancia de los documentos o páginas web indexados por el motor de búsqueda. PageRank establece un mecanismo de voto democrático, en el que los enlaces de las páginas se utilizan como indicadores del valor de una página concreta, de modo que un enlace de una página *A* a una página *B* es interpretado como un voto para la página *B.* El algoritmo básico se puede extender para considerar el *prestigio* de las páginas

que emiten el voto, de manera que los votos de las páginas consideradas importantes valen más que los de otras páginas de poca importancia.

Radev y Erkov (2004) presentan *LexRank*, uno de los métodos más aceptados para calcular la centralidad en un grafo, aplicado a generación automática de resúmenes multidocumento. LexRank utiliza caminos aleatorios y vectores propios (*eigenvectors*) para estimar la importancia relativa de las oraciones, y construye un grafo para el cluster de documentos en el que existe un vértice por cada oración del mismo. Para determinar los enlaces entre los vértices, las oraciones se representan con sus vectores de frecuencias (*tf*idf*), y se calcula la similitud semántica entre ellos utilizando la métrica del coseno, calculando una matriz de similitudes como la de la Figura 4. Aquellos pares de oraciones que presenten una similitud superior a un determinado umbral, se enlazan entre sí en el grafo (Figura 3). Partiendo de la hipótesis de que las oraciones que son similares a muchas otras en el cluster son las más "centrales" (*salient)* en relación al tema (*theme),* la extracción de oraciones relevantes consiste en identificar las oraciones más centrales en el cluster que proporcionan la información necesaria y suficiente en relación con el tema principal. En el artículo se investigan distintas definiciones de centralidad léxica en múltiples documentos:

- **Centralidad basada en el grado (degree centrality),** que define la centralidad de la oración como el grado del correspondiente nodo en el grafo de similitud, de manera que cada arista se considera un voto para el nodo al que se encuentra conectada.
- **Centralidad basada en vectores propios (eigenvector centrality),** que pondera cada voto por la importancia o prestigio del nodo que lo emite.

El resumen final se construye combinando las *n* oraciones más centrales, donde *n* variará en función de la longitud del resumen deseado.

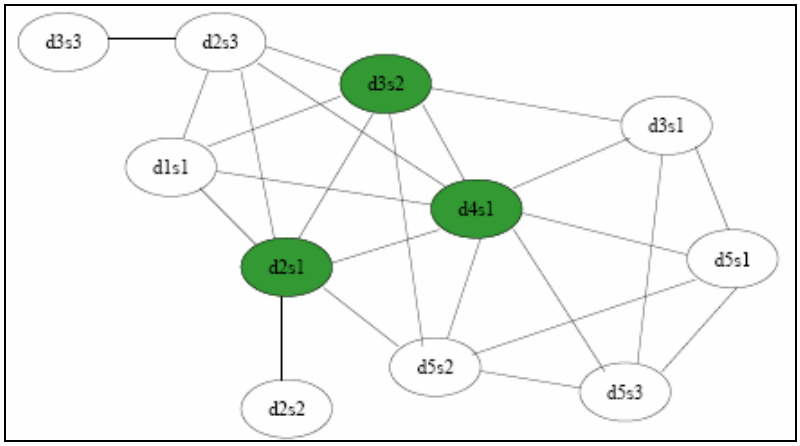

Figura 3 Grafo de similitud (Radev et al., 2004)

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

Figura 4 Matriz de similitudes entre oraciones (Radev et al., 2004)

LexRank se ha utilizado como parte del sistema MEAD, para la generación de resúmenes multidocumento, combinado con otras características como la posición y la longitud de las oraciones. Para abordar el problema de la redundancia, MEAD aplica una etapa de post-procesamiento en la que elimina las oraciones que son muy similares a otras ya seleccionadas, utilizando el método de *Subsunción de Información entre Oraciones (Cross-Sentence Information Subsumption, CSIS).* Los resultados de este sistema se encuentran

entre los tres mejores de las tareas en las que participaron en las conferencias DUC 2003 y DUC 2004.

Un algoritmo similar es *TextRank*, utilizado para la generación de resúmenes monodocumento, aunque también ha sido aplicado a otras tareas de procesamiento de lenguaje natural, como a la extracción de palabras clave. TextRank básicamente ejecuta PageRank sobre un grafo diseñado para la tarea particular que se desea abordar. Los vértices del grafo pueden ser distintas unidades textuales (oraciones o palabras, dependiendo de la tarea), mientras que las aristas miden la similitud léxica o semántica entre las unidades textuales. A diferencia de PageRank, los enlaces no son dirigidos, y pueden tener un peso para reflejar el grado de similitud. El grafo se utiliza para construir una matriz estocástica, combinada con un factor de amortiguamiento, y el ranking de los vértices se obtiene encontrando los vectores propios correspondientes al valor propio 1 (i.e., la distribución estacionaria de los caminos aleatorios en el grafo). Al igual que LexRank, cuando TextRank se aplica a la generación de resúmenes, los nodos del grafo representan a las oraciones. Sin embargo, TextRank utiliza una medida de la similitud entre dos oraciones basada en el número de palabras que tienen en común, normalizada por su longitud.

En cualquier caso, ambos algoritmos presentan resultados prometedores sin necesidad de realizar un análisis en profundidad del texto.

Otro trabajo interesante se presenta en (Yoo et al., 2007). En él se propone un método para la agrupación de documentos con contenidos similares, concebido como paso previo a la generación de resúmenes multidocumento en el dominio biomédico. Para ello, crea una representación en forma de grafo de cada documento de la colección, donde los nodos representan los descriptores de MeSH asociados a los términos, y los enlaces representan las relaciones de hiperonimia y co-ocurrencia entre ellos. Una vez construidos los grafos de cada documento, se fusionan en un único grafo que representa a toda la colección, y que se comporta como una red libre de escala.

A continuación, localiza en el grafo del corpus los nodos que concentran un mayor número de enlaces, y utilizan el algoritmo de clustering *SFGC (Scale-free Graph Clustering)* para agruparlos en conjuntos de conceptos estrechamente conectados entre sí. El resto de nodos se asignan al grupo con el que presentan una mayor conectividad, de manera que cada uno de los grupos representa un *theme* o tópico en el corpus. Finalmente, se utiliza un mecanismo de voto democrático para asignar los documentos a los distintos clusters.

5. Generación de Resúmenes Multidocumento

La gran cantidad de información disponible, fundamentalmente a través de internet, ha propiciado una situación en la que a menudo nos encontramos con decenas e incluso cientos de documentos cubriendo un mismo asunto, y en los que la mayor parte de la información comunicada coincide. Nadie puede cuestionar la utilidad de la información. Sin embargo, en ocasiones, tal sobrecarga puede resultar contraproducente, simplemente porque no se dispone del tiempo necesario para procesarla.

Supongamos que estamos interesados en recopilar información a cerca de la enfermedad leucemia. Acudimos a internet y tecleamos en algún buscador la palabra "leucemia". Al instante disponemos de cientos de documentos, en muy distintos formatos, en los que, de algún modo, se informa sobre esta enfermedad. Probablemente, la mayoría de los documentos recuperados sean irrelevantes y decidamos restringir la búsqueda, por ejemplo, a aquellos documentos en los que se estudia la leucemia en relación a sus síntomas. Pero incluso aunque consigamos, con la combinación perfecta de parámetros para la búsqueda, aislar los documentos que efectivamente contienen información de nuestro interés, el número será tan elevado que una lectura de todos ellos resultará prohibitiva en tiempo y esfuerzo. En este contexto, será de gran utilidad disponer de un resumen en el que se identifique lo que es común a todos ellos y lo que difiere de unos a otros. La consideración de este problema ha resultado en la extensión de las investigaciones en generación de resúmenes monodocumento a la generación de resúmenes de colecciones de documentos relacionados temáticamente. El haber sido motivado por la World Wide Web,

tiene como consecuencia que sea precisamente en esta área en la que se encuadran la mayoría de las investigaciones; concretamente, en el dominio de los artículos periodísticos, que no exigen conocimiento alguno del dominio.

La tarea de generar resúmenes a partir de múltiples fuentes es mucho más compleja que la generación de resúmenes de un solo documento, y plantea retos adicionales. En primer lugar, la selección de documentos que comparten una relación semántica y que contribuirá a la redacción de un mismo resumen debe realizarse cuidadosamente, para evitar mezclar en un mismo resumen informaciones inconexas. En segundo lugar, el hecho de contener los documentos información común puede dar lugar a resúmenes redundantes. Por ello, la detección y eliminación de redundancias es uno de los principales problemas a los que se enfrenta la generación automática de resúmenes multidocumento. Tercero, es igual de importante reconocer las diferencias entre documentos, que pueden deberse a la consideración de información adicional o al planteamiento de la misma bajo distintos puntos de vista. Por último, se debe asegurar la coherencia del resumen, teniendo en cuenta que las diferentes porciones de información provienen de diferentes fuentes. Además de estos problemas que podríamos considerar principales, hay otros aspectos a considerar:

- Los documentos pueden ser de tamaño muy dispar.
- El ratio de comprensión ha de ser mucho mayor para dar lugar a un resumen razonable.
- El uso de diagramas u otros tipos de visualizaciones puede ser de gran utilidad, dado que el resumen tendrá que condensar mucha más información.
- Se requiere de fusión de información cruzada entre documentos, que puede implicar eliminación, agregación y generalización aplicada sobre todo el conjunto, y no sobre documentos individuales.
- Pueden aparecer problemas de incoherencia, e incluso puede darse el caso de que dos documentos aporten información contradictoria.

Como paso previo a la construcción de un sistema generador de resúmenes multidocumento, es necesaria la identificación de subconjuntos de documentos relativos a un mismo tema. Cuanto mayor sea la similitud entre

ellos, más sencilla será la identificación de la información compartida y por consiguiente, la elaboración del resumen.

Al igual que ocurriera en la generación de resúmenes monodocumento, los sistemas multidocumento se pueden clasificar en función del uso de conocimiento lingüístico que hacen. Además, pueden utilizar distintas unidades textuales y a distinto nivel para la comparación entre los documentos.

- A **nivel morfológico**, oraciones, párrafos y documentos se pueden comparar utilizando bolsas de palabras para medir el solapamiento de vocabulario entre ellos, clusters de palabras relacionadas, o incluso en base a la presencia de nombres propios comunes en distintos documentos.
- Alternativamente, las proposiciones y oraciones se pueden comparar a **nivel sintáctico** entre documentos, de manera que podemos comprobar si la estructura sintáctica de una oración está relacionada con la estructura de otra, o si una es un parafraseado sintáctico de la otra.
- Finalmente, los documentos también se pueden comparar a **nivel semántico**, para ver si tratan los mismos temas. Esta tarea se suele abordar mediante técnicas de extracción de plantillas.

Las **aproximaciones a nivel morfológico** trabajan identificando elementos en los textos y estableciendo una relación entre ellos a través de la comparación del vocabulario que los componen. Esta comparación se puede realizar a nivel de palabra o de grupos de palabras, y puede hacer uso de distintas medidas de similitud como son el *dice coefficient,* el *jaccard coefficient,* el *inclusion coefficient* y el *cosine similarity coefficient*. A continuación, seleccionan un subconjunto de los elementos más relacionados, y los agrupan en función de la matriz de similitud construida. La principal ventaja de este enfoque es su robustez, pero presenta el inconveniente de generar resúmenes altamente redundantes. En (Salton et al., 1997) se localizan pasajes similares en función del solapamiento de vocabulario y se extraen

utilizando grafos de conectividad, en los que los nodos son vectores de documentos y los arcos, relaciones de similitud entre ellos. Como medida de esta similitud se utiliza la métrica del coseno. Aunque captura la información común en los documentos, no detecta las diferencias o particularidades de cada uno de ellos. En (Ando et al., 2000) se aplican principios similares a los de *la Indexación de Semántica Latente (LSI, Latent Semantic Indexing)* de Deerwester, para representar los distintos elementos textuales (documentos, frases y términos) en un espacio semántico que refleja similitudes entre términos basadas en su aparición en contextos comunes.

Por lo general, las **aproximaciones a nivel sintáctico** utilizan conocimiento sintáctico al relacionar la información a través de los documentos y para determinar la equivalencia informativa. Barzilay et al. (1999) consideran que existe equivalencia informativa entre dos elementos si los dos se refieren al mismo objeto y realizan la misma acción o se describen del mismo modo. Además, utilizan generación de lenguaje (basada en sintaxis) para reescribir el resumen.

Por último, las **aproximaciones a nivel semántico** básicamente identifican elementos en el texto y los relacionan de modo que salgan a la luz las similitudes y diferencias semánticas. Generalmente, utilizan agregación y generalización para producir descripciones más concisas. Además, la etapa de síntesis suele incluir algo de generación de lenguaje natural.

Los primeros enfoques semánticos proponen el uso de técnicas de extracción de información para la identificación de similitudes y diferencias entre documentos. Al igual que en el caso de un solo documento, se pueden considerar un tipo de técnicas de abstracción, que realizan un análisis muy dependiente del dominio y capturan únicamente ciertos tipos de información predefinida (Mckeown y Radev, 1995). En trabajos posteriores, los mismos autores combinan las técnicas de extracción con métodos de reescritura de texto para conseguir resúmenes de mayor calidad. Las diferencias importantes entre documentos, como contradicciones, actualizaciones, etc. se identifican a través de un conjunto de reglas de discurso (Mckeown y Radev, 1998). Otras investigaciones posteriores mejoran las técnicas de extracción de información utilizadas tradicionalmente e incluyen fórmulas de contraste adicionales (White y Cardie, 2002).

Los trabajos en generación de resúmenes multidocumento, con independencia del enfoque que utilicen, adolecen en su mayoría de una serie de problemas:

- En primer lugar, no tratan la **redundancia**. Una primera aproximación, muy simple, para mejorar este problema es el uso de sinónimos, aunque en general se necesitan medidas más complejas. Un enfoque común consiste en medir la similitud entre pares de oraciones y utilizar clustering para identificar *themes* de información común (McKeown et al., 1999; Radev, Jing y Budzikowska, 2000; Marcu y Gerber, 2001). Otros sistemas miden la similitud entre los pasajes candidatos y aquellos que ya han sido seleccionados, incluyéndolos únicamente en el caso de contener suficiente información nueva. Una medida muy popular es la *Relevancia Marginal Máxima (MMR, Maximum Marginal Relevance),* utilizada en los trabajos de Carbonell, Geng y Goldstein (1997) y Carbonell y Goldstein (1998). En el contexto multidocumento, y de un resumen orientado al usuario, trata de evitar el problema de cubrir solamente la información común. Los textos se ordenan en términos de relevancia para la consulta, y el usuario, mediante el ajuste de los parámetros de ordenación puede establecer el grado de diversidad que desea incluir en el resultado. De esta manera, se está controlando también la redundancia, si la diversidad requerida se establece a su valor máximo. El problema es que en función de este valor, los resúmenes que se obtienen pueden ser muy diferentes, y el usuario tendrá que aprender a controlar la parametrización en función de sus necesidades. Goldstein (2000) extiende este concepto al de *Relevancia Marginal Máxima Multidocumento (MMR-MD, Maximum Marginal Relevance-MultiDocument).* En (Radev et al., 2000) se introduce la noción de *Subsunción de Información entre Frases (CSIS, Cross-Sentence Informational Subsumption)*, que permite distinguir entre inclusión y equivalencia informativa, y se muestra muy eficaz a la hora de tratar la redundancia. Mani (2001) identifica una serie de relaciones entre documentos que caracterizan la redundancia entre ellos: *equivalencia semántica*, *equivalencia informativa*, *igualdad literal* e

inclusión informativa. En este contexto, gana importancia el criterio de selección basado en el concepto de cobertura, en el sentido de que una frase de un documento que contiene a varias de otros es preferible a estas. Por último, Mani, Gates y Bloedorn (1999) describen el uso de compresión y reglas de reformulación.

- Otro problema importante a resolver es el de la **captura de diferencias** entre documentos. Los trabajos existentes identifican bien las similitudes entre documentos, pero no tan bien las diferencias. En (Radev, 2000), se propone la *Teoría de la Estructura Inter-Documento, (CST, Cross-Document Structure Theory)*, que identifica un total de veinticuatro relaciones entre documentos o pasajes de documentos, algunas de las cuales reflejan las diferencias además de las similitudes.
- Precisamente estas diferencias llevan al problema de la **calidad de las fuentes informativas**. El hecho de que en una fuente se reporte un acontecimiento que es significativamente opuesto o diferente al contenido de otras fuentes puede significar que se trata de información errónea o poco importante, y que por tanto, no debe formar parte del sumario; o por el contrario, puede constituir un punto de vista diferente de un mismo suceso, pero igualmente interesante. Incluso dentro de un mismo documento puede ocurrir que se contrasten distintas visiones.
- Por último, otro aspecto a solucionar es el de la **coherencia**. Cuando se trata con resúmenes de múltiples documentos, asegurar la coherencia es difícil, ya que requiere cierta comprensión del contenido de cada pasaje y conocimiento de la estructura del discurso. La mayoría de los sistemas actuales simplemente se limitan a disponer los pasajes u oraciones en orden temporal y respetando el orden en el que aparecen en el documento original. Barzilay, Elhadad y McKeown (2001) combinan restricciones temporales y cohesivas para ordenar las oraciones. Radev y Luo (2002) utilizan una revisión basada en el discurso para mejorar la coherencia del resumen.

En la Figura 5 se muestra una arquitectura genérica para un sistema de generación de resúmenes de múltiples documentos. En ella se distinguen las siguientes etapas:

- **Agrupamiento** de documentos por su proximidad semántica, de manera que se identifiquen los grupos de textos que contribuirán a producir un mismo resumen
- **Análisis superficial para la selección de pasajes** relevantes, utilizando los conceptos de la ontología presentes en el texto.
- **Análisis detallado para la interpretación** de los pasajes seleccionados, y la representación de los mismos en la ontología.
- **Detección de pasajes comunes y de diferencias entre documentos.**
- **Eliminación de redundancias**
- **Generación del resumen,** utilizando técnicas de procesamiento de lenguaje natural.

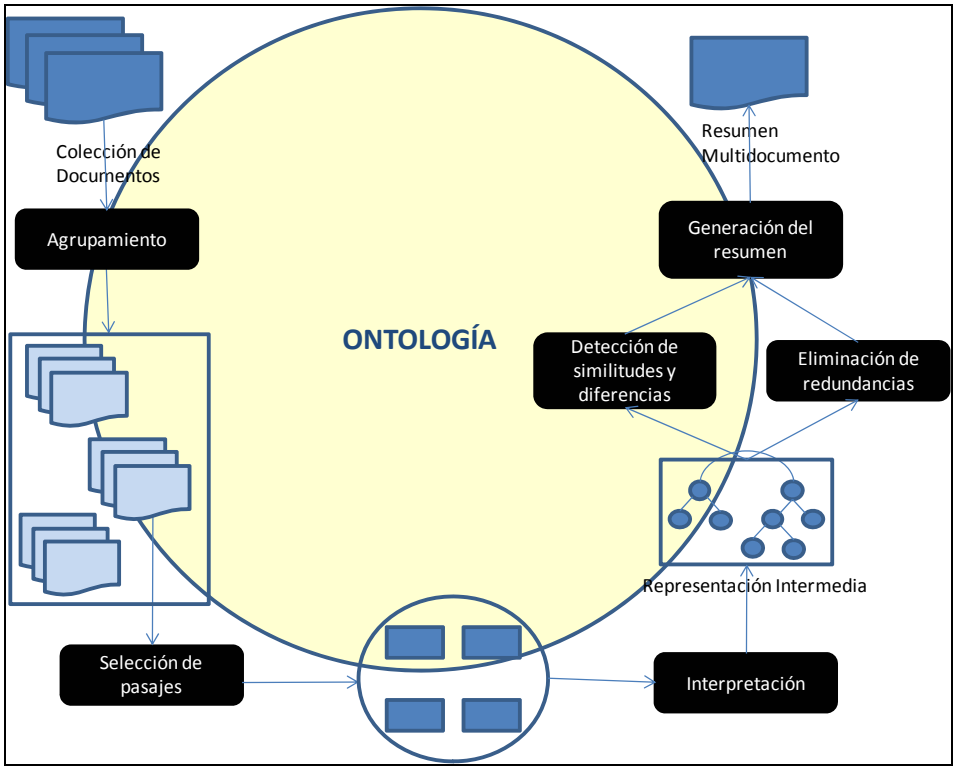

Figura 5 Propuesta de arquitectura para generación de resúmenes de múltiples documentos

En todas las etapas, a su vez, se puede hacer uso de conocimiento del dominio, formalizado en una ontología, así como tener en cuenta la estructura y particularidades de los documentos a procesar. No obstante, es importante observar que no todos los trabajos en generación multidocumento respetan esta arquitectura. Dependiendo del enfoque, muchas de las fases propuestas pueden no acometerse, o incluso se puede modificar el orden en el que se ejecutan. Por ejemplo, los enfoques a nivel morfológico raramente contemplan la interpretación de las oraciones o pasajes seleccionados, ni utilizan generación de lenguaje natural para reescribir el resumen. Los enfoques a nivel morfológico y sintáctico raramente consiguen identificar las diferencias entre documentos, mientras que muchos trabajos en cualquiera de los enfoques no resuelven el problema de la redundancia.

Quizás uno de los sistemas de generación automática de resúmenes multidocumento más populares sea *NewsBlaster,* desarrollado por la Universidad de Columbia para generar resúmenes de noticias sobre un mismo evento. NewsBlaster se compone de seis módulos generales, que se encargan de rastrear un conjunto de sitios web de noticas diarias, filtrarlas para eliminar la información que no son noticias (como por ejemplo, anuncios publicitarios), agrupar las noticias relativas a un mismo evento, realizar un resumen de cada grupo de noticias, clasificarlas en un conjunto predefinido de categorías y generar una página web en la que se muestra los resúmenes generados, junto con enlaces a los artículos originales (Figura 6).

Desde la perspectiva de la generación de resúmenes multidocumento, los módulos más relevantes son el de *clustering* y el de *summarization,* por lo que será en ellos en los que centremos nuestra discusión.

Para realizar el agrupamiento de las noticias sobre un mismo evento, NewsBlaster utiliza el algoritmo de clustering no jerárquico *groupwise-average.* En cuanto a las características de los documentos que se consideran para guiar el agrupamiento, además de los tradicionales vectores de palabras, se utilizan dos características lingüísticas de los textos: los sintagmas nominales y los nombres propios. Los sintagmas nominales se identifican con la herramienta *LinkIt*, desarrollada por ellos mismos; mientras que para

identificar los nombres propios, se utiliza *Nominator* de IBM. Para el cálculo de la similitud entre los vectores correspondientes a dos documentos, se calcula la inversa de la frecuencia de los términos (*tf*idf*).

En cuanto a la generación del resumen se refiere, los clusters construidos de la etapa anterior se dirigen, en función de sus características, a uno de los siguientes sistemas: MULTIGEN, en caso de que los artículos en el cluster de entrada estén referidos a un mismo suceso; o DEM, en caso contrario (por ejemplo, todos se refieren a la guerra de Irak pero reportan sucesos distintos).

Multigen Summarizer presenta a su vez dos componentes: un componente de análisis y un componente de generación. El *componente de análisis,* divide los artículos de entrada en párrafos y crea conjuntos de párrafos similares a los que denomina *themes*, utilizando aprendizaje automático sobre características lingüísticas para obtener medidas de la similitud entre párrafos, que luego son enviadas a un algoritmo de clustering. De nuevo, hace uso de otras herramientas lingüísticas disponibles, como LinkIt y WordNet. El *componente de generación* analiza la estructura gramatical de cada *theme* para determinar qué oraciones se repiten lo suficiente para ser incluidas en el resumen. Posteriormente, analiza cada *theme* y utiliza el sistema FUF/SURGE para generar el resumen final.

DEM Summarizer (Dissimilarity Engine for Multidocument Summarization) procede asignando una medida de relevancia a cada una de las frases de los artículos, de manera que sólo las más significativas formarán parte del resumen. Para ello, utiliza tres medidas principales:

1) *Lead values,* o palabras que aparecen mucho más en los titulares que en el texto completo. Se toma como hipótesis que las frases que contengan muchos *lead values* serán muy significativas.
2) *Verb specificity.* Partiendo de la idea de que existen verbos muy ligados a un sujeto concreto, se considera que las frases que los contienen aportan mucha información.
3) *Concepts:* En lugar de trabajar sobre sustantivos o nombres concretos, se trabaja sobre conceptos, utilizándose WordNet para construir lo que denominan *Concept Sets*, consiguiendo además resolver las referencias satisfactoriamente.

Junto a estas, se utilizan otras heurísticas como premiar las oraciones del inicio del artículo y penalizar las del final, premiar los documentos más recientes, penalizar las oraciones muy cortas o muy largas y aquellas que presentan pronombres al inicio, etc.

Una vez se tienen las frases relevantes, se ordenan y se elimina el solapamiento entre ellas. Para resolver el problema de las entidades referenciadas, se utiliza NOMINATOR, de manera que la primera vez que se referencia a un sujeto se utiliza su nombre completo y/o descripción, y para las siguientes, se sigue la regla de "el más corto y más común".

Figura 6 Sitio web de *Columbia Newsblaster*

6. *Generación de Resúmenes Adaptada al Usuario*

Estudios empíricos han demostrado que, ante la generación manual de resúmenes de un mismo documento por parte de personas con conocimiento previo, áreas de interés y necesidades de información diferentes, la información seleccionada como relevante difiere significativamente de una persona a otra (Paice y Jones, 1993). Esta consideración nos lleva al estudio de técnicas que permitan elaborar resúmenes teniendo en cuenta las características del lector o grupo de lectores a quien va dirigido.

Distintos trabajos han utilizado la generación de resúmenes personalizados en conjunción con técnicas de recuperación de información y de navegación para mejorar el proceso de localización de documentos relevantes a una consulta (Maña, *et al*., 1999). Los sistemas de acceso a la información tienen como principal objetivo ayudar a los usuarios a localizar documentos que satisfagan sus necesidades (Baeza-Yates y Ribeiro-Neto, 1999). Estas necesidades pueden ser a corto plazo o a largo plazo, en función de su persistencia en el tiempo. Mientras que los sistemas de navegación y recuperación generalmente hacen uso de necesidades a corto plazo, los sistemas de filtrado utilizan necesidades a largo plazo para proveer información de acuerdo al perfil del usuario.

Para generar un resumen adaptativo se necesita disponer de un modelo del usuario, entendido como una representación de sus intereses y preferencias. Dependiendo del contexto donde sea aplicado, la complejidad de la representación puede variar. En Maña (1999) se utiliza un conjunto de palabras claves para describir al usuario, mientras que en Díaz, Gervás y García (2005), se hace uso de una representación compleja de acuerdo a varios sistemas de referencia.

Si bien los primeros trabajos en generación automática de resúmenes ya apuntaban a la posibilidad de utilizar un proceso adaptativo (Luhn, 1958; Edmundson, 1969), la mayor parte del trabajo surge a partir de la década de los 90, como consecuencia de los buenos resultados alcanzados en recuperación de información, donde los resúmenes personalizados adquieren gran relevancia. En (Carbonell et al., 1997) los resúmenes generados se ajustan fielmente a la consulta del usuario. Para ello, utilizan la denominada *Relevancia Marginal Máxima (MMR, Maximal Marginal Relevance)*, que permite ordenar las oraciones en función de su similitud con los términos introducidos en la consulta y evitar la inclusión de oraciones redundantes. En (Sanderson, 1998) se selecciona el pasaje más relevante para la consulta, utilizando la técnica del *Análisis del Contexto Local* (*LCA, Local Context Analysis*). Esta técnica permite extender la consulta original con las palabras más frecuentes del contexto en el que las palabras de la consulta aparecen en el primer documento recuperado. En (Maña et al., 1999; Amini y Gallinari, 2002) la expansión de la consulta se realiza utilizando WordNet.

Finalmente, también es posible utilizar técnicas de aprendizaje para confeccionar resúmenes personalizados. Un trabajo en esta dirección es (Lin, 1999), y en él, las características que se exploran son el número de palabras de la consulta y el número de palabras frecuentes que aparecen en la oración. Las palabras frecuentes se extraen de entre los *n* primeros documentos recuperados. En (Mani y Bloedorn, 1998), para determinar los intereses del usuario, se le pide que elija, entre una colección artículos, los diez más acordes a sus intereses. A continuación, se calcula el centroide de estos documentos, que caracteriza el tema de interés del usuario, y se expande con otras palabras con las que mantiene relaciones de cohesión en el corpus. Las características utilizadas en el proceso de aprendizaje son el número de palabras clave del tema que aparecen en la oración y el número de palabras clave del tema por el número de palabras significativas en la oración.

7. Evaluación de Resúmenes Automáticos

Ya desde los primeros trabajos en generación automática de resúmenes, la comunidad investigadora ha sido consciente de la necesidad de evaluar la calidad de los resúmenes generados. Sin duda alguna, se trata de una tarea compleja y controvertida, y a pesar del esfuerzo dedicado, aún no se ha alcanzado un acuerdo acerca de cuáles deberían ser las prácticas a seguir.

La evaluación de resúmenes generados automáticamente requiere, como en la evaluación de cualquier otro producto, la construcción de conjuntos de datos estándares, y la definición consensuada de diferentes métricas. Aunque lo deseable sería contar con procedimientos que permitieran realizar la tarea de manera automática, la evaluación presenta una serie de problemas que la convierten en una tarea difícil de acometer (Mani, 2001):

- En primer lugar, generalmente es necesario acudir a jueces humanos para realizar la evaluación, lo cual resulta tedioso y costoso en tiempo y recursos. Además, frecuentemente las opiniones de estos jueces son contradictorias o difieren significativamente.

- En segundo lugar, cualquier afirmación sobre la calidad de un resumen está sujeta a la apreciación subjetiva de la persona encargada de realizar la evaluación.
- En tercer lugar, la evaluación de un resumen no debe restringirse a la valoración de su correcta redacción o legibilidad, sino que también es importante considerar el grado en que satisface las necesidades de información del usuario, que dependerá fundamentalmente del uso para el que vaya a ser destinado.
- Finalmente, deberá evaluarse el resumen en relación a la tasa de compresión respecto del texto original, lo que implica disponer de distintos resúmenes generados de manera manual para realizar la comparación.

7.1. Clasificación

La clasificación más aceptada de los métodos de evaluación fue propuesta por Sparck-Jones (1996), y distingue entre *métodos directos o intrínsecos* y *métodos indirectos o extrínsecos.*

Los **métodos intrínsecos** se basan en el análisis directo del resumen producido para juzgar aspectos como su cohesión, coherencia, y cobertura u omisión de los temas principales del texto de entrada. Mientras que para evaluar el grado de cobertura generalmente se compara el resumen con el texto original, para evaluar la corrección gramatical del texto, únicamente se analiza el resumen generado. En el primer caso, lo habitual es recurrir a resúmenes realizados manualmente por expertos humanos para compararlos con los generados de manera automática, lo que conlleva, como ya hemos adelantado, un elevado coste, más aún teniendo en cuenta que la evaluación de un generador ha de realizarse sobre una colección de textos suficientemente extensa para que los resultados sean estadísticamente significativos. En el segundo caso, también es frecuente solicitar a jueces humanos una valoración de los resúmenes, durante la cual pueden producirse fuertes discrepancias entre los distintos jueces que invaliden la evaluación.

Los **métodos extrínsecos** estudian el resumen en el contexto de la tarea para la que ha sido generado, y para ser utilizados en lugar del documento original. Evalúan, por ejemplo, si permiten clasificar correctamente un documento o responder a ciertas preguntas sobre el texto de entrada. A la hora de realizar este tipo de evaluación, de nuevo se puede acudir a jueces humanos que califiquen la adecuación del resumen a la tarea o, siempre que el problema particular lo permita, medir automáticamente su desempeño en la resolución de la misma. Este tipo de evaluación resulta muy adecuada en problemas de categorización y recuperación de información, en los que los experimentos realizados con distintos sistemas han demostrado que los resúmenes pueden llegar a producir resultados muy similares a los obtenidos con los textos completos, con el consiguiente ahorro en tiempo y recursos que esto conlleva.

7.2. Métodos de evaluación

Especialmente en la última década, han sido numerosas las propuestas que han perseguido el consenso entre las distintas aproximaciones a la evaluación de resúmenes, tanto en relación a métodos intrínsecos como extrínsecos.

Una primera aproximación utiliza las métricas tradicionales de recuperación de información para medir la calidad de los resúmenes automáticos, en comparación con otros redactados manualmente. No obstante, estas técnicas presentan el inconveniente de que pueden proporcionar resultados distintos para resúmenes que contengan la misma información, por lo que en la actualidad prácticamente no se utilizan. Considérese la tabla de la Figura 7, en la que se representan las posibles combinaciones respecto a la coincidencia de oraciones entre el resumen generado automáticamente y el redactado de manera manual.

	Automático	
Manual	+	-
+	TP	FN
-	FP	TN

TP → True Positive
FP → False Positive
TN → True Negative
FN→ False Negative

Figura 7 Posibles combinaciones de resultados en la evaluación de resúmenes

Si una oración se selecciona en ambos resúmenes o en ninguno de ellos, estaremos ante un *verdadero positivo (true positive)* o un *verdadero negativo (true negative).* Por el contrario, si únicamente se selecciona en el resumen manual o en el automático, estaremos, respectivamente, ante un *falso negativo (false negative)* o un *falso positivo (false positive).* A partir de esta clasificación, podemos definir las siguientes métricas clásicas de recuperación de información:

- La **precisión** (*precision*) mide el número de frases coincidentes en ambos resúmenes en relación al número total de oraciones presentes en el resumen automático.

$$precision = \frac{TP}{TP + FP}$$

Ecuación 1

- La **cobertura** (*recall*) mide la tasa de frases del resumen de referencia presentes en el resumen generado automáticamente.

$$cobertura = \frac{TP}{TP + FN}$$

Ecuación 2

- La **medida–F** (*F-Score*), una combinación de las medidas anteriores que representa la intersección entre las oraciones implicadas en la precisión y la cobertura, normalizada por la suma de ambas.

$$medida - F = \frac{2 \times precision \times cobertura}{precision + cobertura}$$

Ecuación 3

Otro método de evaluación es el denominado *Índice de Utilidad* (Radev et al., 2000). Partiendo de la hipótesis de que no todas las oraciones seleccionadas para formar parte del resumen tiene la misma relevancia, se asigna a cada una de ellas un grado de pertenencia al resumen. En primer lugar, se pide a los jueces que puntúen las oraciones del texto fuente entre 1 y 10, y los distintos resúmenes de referencia se construyen seleccionando las oraciones con una mayor valoración. El índice de utilidad se calcula dividendo la suma de valoraciones de las frases seleccionadas por el sistema entre la suma de valoraciones de las frases del resumen de referencia.

En (Donaway et al., 2000) se propone como métrica la denominada *similitud de contenidos*, que consiste en el cálculo de la distancia entre las representaciones de los resúmenes automático y manual utilizando la medida del coseno.

La herramienta MEADeval[1], permite calcular los índices de precisión, cobertura, utilidad y similitud anteriormente descritos.

La primera iniciativa independiente y de gran escala fue *SUMMAC*[2] (*The TIPSTER Text Summarization Evaluation Conference)*, llevada a cabo por el gobierno de los Estados Unidos en el año 1998. SUMMAC definía dos tareas principales de evaluación extrínseca, basadas en actividades realizadas por el gobierno americano, y una tarea de evaluación intrínseca. Para la primera tarea, de recuperación *ad hoc*, los sistemas recibían un documento y un tema, y el objetivo era producir un resumen indicativo centrado en el tema especificado. El analista debía clasificar el documento en una de dos categorías, dependiendo de si trataba o no sobre el tema en cuestión. Los resultados de esta tarea muestran que la utilización de resúmenes consigue una efectividad (en términos de la métrica F1), similar a la que se alcanza con los textos completos, pero con un ahorro de tiempo superior al 40%. En la segunda tarea, de categorización, el sistema debía producir resúmenes genéricos, que posteriormente eran clasificados por un analista en una de las *n* categorías propuestas o en ninguna de ellas. El objetivo perseguido era determinar si un

[1] http://www.summarization.com/mead/

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/

resumen genérico podía presentar suficiente información como para reducir el tiempo empleado por los analistas en su clasificación sin incurrir en un mayor error. De nuevo, se consiguen resultados similares a los obtenidos con los textos originales, y un ahorro de tiempo en torno al 40%. Por último, la tercera tarea, de evaluación intrínseca, consistía en generar resúmenes sobre un tema específico y evaluar el grado en que una serie de preguntas sobre dicho tema encontraban respuesta en el resumen generado, introduciendo para ello métodos automáticos de evaluación. Todas las tareas se realizan tanto para resúmenes de longitud fija como para resúmenes de longitud variable. Los resultados demostraron la utilidad de los resúmenes automáticos en tareas de recuperación de información y categorización, pero a la vez pusieron de manifiesto la necesidad de disponer de métodos automáticos, ya que fue necesaria la colaboración de 51 analistas para realizar la evaluación.

La serie de conferencias *DUC* (*Document Understanding Conferences*)[3] se inician en el año 2000, promovidas por la *NIST (National Institute of Standards and Technology),* y desde entonces se han celebrado cada año. Surgen como una iniciativa para el desarrollo de un marco común para la evaluación (y consecuente mejora) de los sistemas de resumen automático, y desde hace ya algún tiempo se ha convertido en el principal foro de evaluación de este tipo de sistemas.

En todas las ediciones DUC se presentan diversas tareas que incluyen, entre otras, la obtención de resúmenes genéricos y específicos a partir de un único documento o de conjuntos de textos relativos a un tema común, y para distintos porcentajes de compresión. La organización prepara los conjuntos de documentos, que suelen consistir en artículos periodísticos en lengua inglesa, y elabora los correspondientes resúmenes modelo para la posterior evaluación de los resultados.

Hasta la edición de 2004, los textos eran evaluados de manera manual por jueces humanos que debían comparar los resultados de los distintos

[3] Document Understanding Conference: http://duc.nist.gov/

sistemas con los modelos disponibles, valorando tanto la calidad de la redacción como el contenido de los resúmenes. Para limitar en lo posible el grado de subjetividad asociado a la evaluación, se utilizaba la herramienta *SEE (Summary Evaluation Environment)*[4]. El método de evaluación subyacente consiste en dividir los textos a comparar en "unidades de discurso", que el revisor deberá asociar a unidades del modelo e indicar si los contenidos de la unidad en el resumen coinciden, total o parcialmente, con aquellos de la unidad en el modelo. El revisor también puede indicar la calidad gramatical de cada unidad y, por último, evaluar de manera global la coherencia, gramática y organización del resumen automático. Finalmente, en lugar de utilizar el *recall* como medida de la coincidencia entre el resumen y el modelo, se utiliza la cobertura *(coverage),* definida de la siguiente manera (Ecuación 4):

$$C = \frac{\text{MU que coinciden} \times \text{E}}{\text{Total MU modelo}}$$

Ecuación 4

Donde MU representa a una unidad de discurso, y E indica el ratio de completitud o de coincidencia entre las unidades, y puede variar de 1 a 0: ¾ para *most,* ½ para *some,* ¼ para *hardly any* y 0 para *none* (Lin y Hovy, 2002). La Figura 8 muestra el sistema software desarrollado para realizar la evaluación según la metodología SEE.

Desde la edición de 2004, sin embargo, la evaluación se realiza de modo automático. En DUC 2004 comienza a utilizarse un sistema denominado ROUGE[5] *(Recall-Oriented Understudy for Gisting Evaluation).* ROUGE (Lin, 2004) evalúa la cobertura de contenidos mediante la comparación del resumen con otro u otros considerados ideales, y mediante el cálculo de las unidades coincidentes entre ambos tipos de resúmenes. Esta herramienta permite el cálculo de distintas medidas, siendo las más destacadas ROUGE-N, ROUGE-L y ROUGE-W. La primera se basa en el número de n-gramas de palabras que coinciden entre un resumen candidato y uno o más modelos, por lo que se pueden calcular las medidas ROUGE-1, -2, -3, etc., no siendo habitual emplear

[4] *SEE* (*Summary Evaluation Environment*): http://www.isi.edu/~cyl/SEE/

[5] ROUGE package: http://haydn.isi.edu/ROUGE/index.html

más allá de los 4-gramas. ROUGE-L emplea la longitud de las secuencias más largas que son comunes en el candidato y en el modelo. ROUGE-W es una versión ponderada de ROUGE-L que además de la longitud de la secuencia valora la ausencia de "huecos" en la misma. A pesar de las críticas recibidas, Lin demostró, utilizando las evaluaciones de las ediciones anteriores, que las medidas obtenidas con ROUGE muestran una elevada correlación con las arrojadas por los jueces humanos, y que además, es posible aplicar la metodología de manera completamente automática.

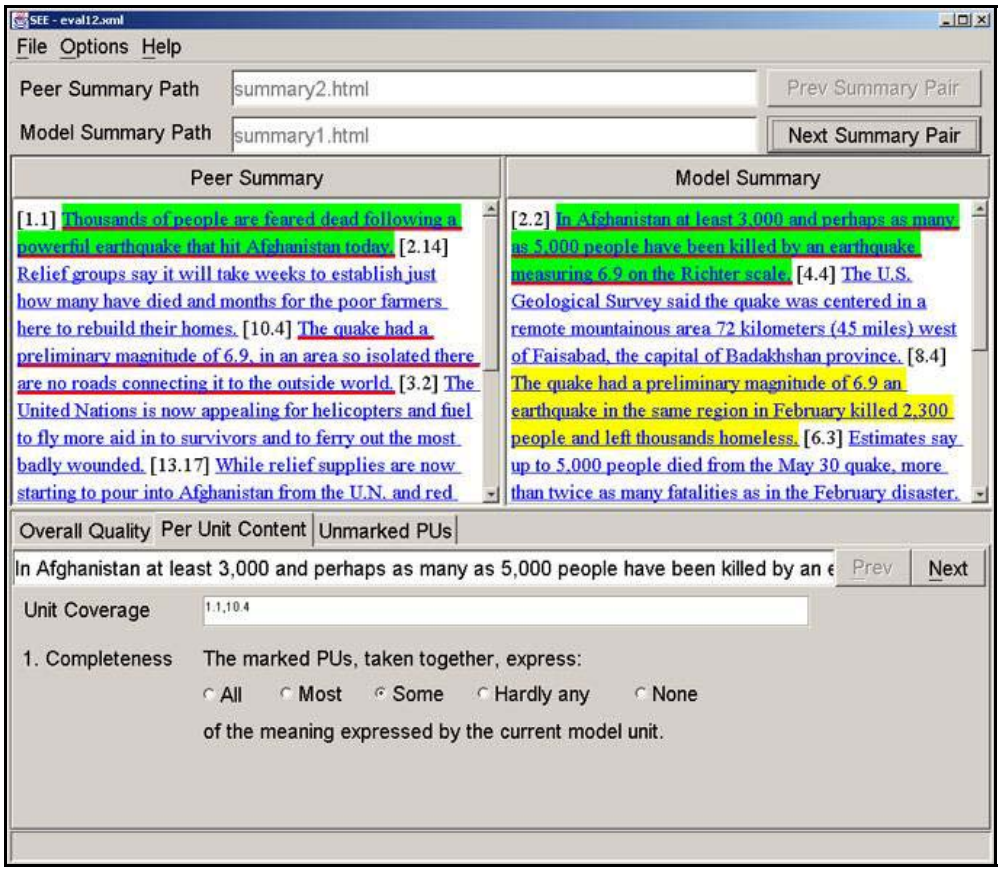

Figura 8 Interfaz de usuario de SEE

En DUC 2005, comienza a utilizarse el entorno de evaluación *The BE package*[6], presentado por Hovy, Lin y Zhou (2005), y basado en la

[6] BE Package: http://www.isi.edu/~cyl/BE/

identificación de unidades semánticas mínimas denominadas *Elementos Básicos (BE, Basic Elements)*. El método propuesto toma como entrada el resumen que se desea evaluar y un conjunto de resúmenes de referencia elaborados manualmente. Seguidamente, divide estos resúmenes en una lista de BE de referencia, fusiona los que son semánticamente equivalentes y asigna una puntuación a cada uno de ellos. A continuación, divide el resumen a evaluar en una lista de BE, los compara con los de referencia y, asigna a los que coinciden la puntuación correspondiente, calculando finalmente una puntuación global para el resumen. Este método ha mostrado una alta correlación tanto con los juicios realizados por los revisores humanos como con las medidas obtenidas con ROUGE.

Opcionalmente, en esta misma edición, los investigadores tenían la posibilidad de someter sus resúmenes a una evaluación manual según el método *Pyramid* (Passonneau et al., 2005). Desarrollado por la Universidad de Columbia, se basa en la observación de que los humanos al realizar un resumen de un texto no siempre seleccionan los mismos elementos relevantes. Para aplicar esta métrica, los resúmenes generados automáticamente se fragmentan en unidades informativas denominadas *SCU (Summarization Content Units)* y se identifican segmentos similares entre los resúmenes, asignando diferentes pesos a cada segmento de información según el número de resúmenes modelo en los que aparece. Se construye una pirámide de SCU de altura *n*, para cada uno de los *n* resúmenes de referencia considerados. A cada SCU de una capa T_i se le asigna un peso W_i, por lo que las SCU de mayor importancia se sitúan en la cúspide de la pirámide. De este modo, el mejor resumen será aquel que contenga más SCU de los niveles superiores.

Con respecto a la legibilidad de los resúmenes, en DUC 2005 se aplicaron los siguientes criterios lingüísticos: errores gramaticales, redundancia, errores de referencia, estructura y coherencia. La calidad de un resumen con respecto a cada uno de estos criterios recibe una puntuación en una escala de cinco posibles valores: *very poor, poor, aceptable, good* and *very good.* Sin embargo, estas medidas no tienen en cuenta la tarea para la que ha sido realizado el resumen, ni garantizan la adecuación de la información contenida en el mismo.

7.3. Corpus para evaluación de resúmenes

A continuación se citan algunos corpus que pueden ser utilizados para la evaluación de resúmenes mediante técnicas intrínsecas.

- **Computation and Language (cmp-lg)**7. Consiste en una colección de 183 documentos desarrollada como parte del proyecto TIPSTER SUMMAC, marcados en xml y muy adecuados para la evaluación de tareas de recuperación de información, extracción de información y generación de resúmenes. Los documentos son artículos científicos que han aparecido en las conferencias de la *Association for Computacional Linguistics (ACL)*, cada uno de los cuales se presenta acompañado de un resumen confeccionado por el propio autor. (Teufel y Moens, 1997).
- Las colecciones utilizadas en las **conferencias DUC**[8]celebradas entre los años 2001 y 2007 se componen de los documentos originales, fundamentalmente artículos periodísticos de las colecciones TREC y TIPSTER, junto con los resúmenes construidos manualmente y los generados automáticamente por los sistemas presentados a concurso.
- El **RST Corpora**[9], consistente en una selección de 385 artículos del Wall Street Journal procedentes del *LDC Treebank*, anotadas con la estructura discursiva siguiendo las directrices de la *Rethorical Structure Theory (RST)*. Además, el corpus incluye los resúmenes (extractos y abstractos) generados manualmente para los distintos artículos.
- **SummBank**[10], pensado fundamentalmente para la evaluación de resúmenes multidocumento, incluye 40 clusters de 10 artículos periodísticos en inglés y chino, junto con los correspondientes

[7] http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html

[8] http://www-nlpir.nist.gov/projects/duc/data.html

[9] http://www.isi.edu/~marcu/discourse/

[10] http://www.summarization.com/summbank/

resúmenes multidocumento redactados manualmente y otros generados por distintos sistemas automáticos.

- **MEAD**[11]. También especialmente diseñado para la evaluación de resúmenes de múltiples documentos, está formado por 6 grupos de textos que contienen, cada uno de ellos, entre 2 y 10 artículos periodísticos que versan sobre el mismo tema (Radev *et al.*, 2000).

[11] http://www.summarization.com/mead/

Capítulo 3

Recursos Utilizados

El propósito de este capítulo es presentar los recursos lingüísticos y ontológicos utilizados como apoyo en la generación automática de resúmenes. Por ello, en primer lugar, se analizan distintas ontologías del dominio biomédico; en concreto, SNOMED, MeSH y UMLS, y se justifica la elección de esta última. Seguidamente, se estudian diversos corpus de documentos biomédicos, y finalmente, se selecciona uno de ellos para nuestro trabajo.

1. Ontologías y Recursos Lingüísticos

El término *ontología* se define el diccionario de la lengua española como la *parte de la metafísica que trata del ser en general y de sus propiedades trascendentales.* Derivado de este significado filosófico, y con un sentido mucho más pragmático, una ontología se entiende como *una especificación formal y explícita de un conocimiento común y compartido de un dominio, que puede ser comunicado entre expertos y sistemas* (Gruber ,1993). Más concretamente, Weigand (1997) define una ontología como *una base de datos que describe los conceptos del mundo o de algún dominio, sus propiedades y las relaciones que se establecen entre conceptos.*

Según Steve et al. (1998), existen tres tipos de ontologías, a los que hay que sumar las ontologías creadas para una tarea específica, como por ejemplo, el diagnóstico de una enfermedad

- **Ontologías de un dominio**, que representan el conocimiento especializado de un determinado campo, como la medicina.
- **Ontologías genéricas**, en las que se representan conceptos generales.
- **Ontologías representacionales o meta-ontologías**, en las que se conceptualizan los formalismos de representación del conocimiento.

A pesar de la amplia diversidad de ontologías disponibles, existe común acuerdo en los siguientes aspectos (Chandrasekaran, 1999).

- En el mundo existen *objetos.*
- Los objetos tienen *propiedades* o *atributos* que pueden tomar *valores.*
- Los objetos pueden *relacionarse* de distintas maneras unos con otros.
- Las propiedades y relaciones pueden *cambiar* en el tiempo.
- Hay *eventos* que ocurren en distintos instantes de tiempo.
- Los objetos participan en procesos que ocurren en el tiempo.
- El mundo y sus objetos pueden estar en distintos *estados.*
- Los eventos pueden *causar* otros efectos o estados.
- Los objetos pueden tener *partes.*

La elección de la forma de describir los conceptos y el formalismo de representación dependen de la aplicación concreta para la que sea desarrollada y del dominio. Aunque las ontologías generalmente aparecen como un árbol o taxonomía de conceptos, dos ontologías pueden diferir en el análisis de los conceptos generales y en la identificación de sus subconceptos. Para ilustrar esta situación, considérese la especificación del concepto *Thing* realizada por cuatro ontologías distintas: WordNet, CYC, GUM y Sowa's (Chandrasekaran, 1999), presentada en la Figura 9.

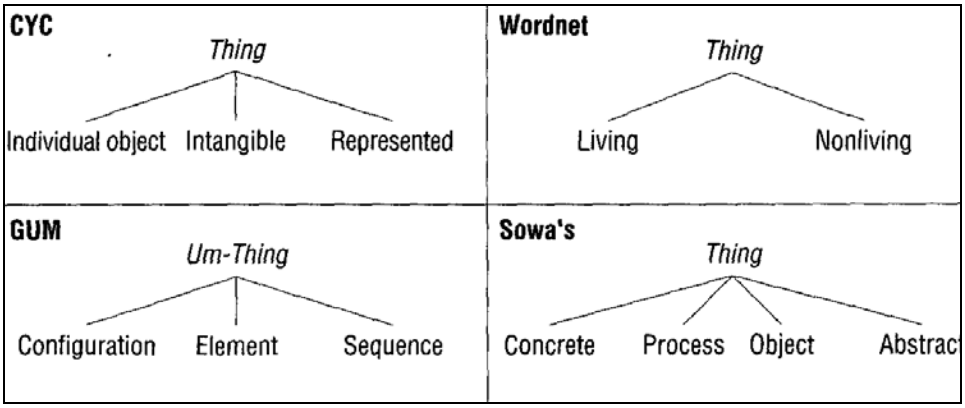

Figura 9 Representaciones de un mismo concepto en distintas ontologías

Según Gómez-Pérez (2004), una ontología se compone de los siguientes elementos:

- **Clases.** Son las ideas a formalizar y representan los conceptos en el sentido más amplio. Las clases en una ontología se suelen organizar en taxonomías a las que se les pueden aplicar los mecanismos de herencia.
- **Relaciones.** Representan las interacciones entre clases, es decir, entre conceptos. Las relaciones más habituales son binarias; del tipo "subclase de" o "conectado con".
- **Funciones.** Son casos especiales de relaciones donde se identifican elementos mediante el cálculo de una función.
- **Instancias.** Se usan para representar elementos o individuos en una ontología.
- **Axiomas.** Sirven para modelar sentencias que son siempre ciertas. Normalmente se usan para representar conocimiento que no puede ser formalmente definido por los componentes descritos anteriormente. También se usan para verificar la consistencia de la propia ontología.

1.3. Ontologías y Terminologías Biomédicas

Mientras que las **ontologías biomédicas** proveen un marco organizacional de los conceptos involucrados en entidades y procesos biológicos, en un sistema de relaciones jerárquicas y asociativas que permite razonar sobre el conocimiento del dominio, las **terminologías biomédicas** promueven una manera estándar de nombrar los conceptos del dominio (Bodenreider et al., 2003).

Sin lugar a dudas, los recursos más utilizados para recuperación de información, en el dominio biomédico, son SNOMED, UMLS y MeSH; y es por ello que los hemos escogido como posibles candidatos a utilizar en el estudio que nos ocupa.

SNOMED-CT

SNOMED–CT[12] son las siglas de *Systematized Nomenclature of Medicine Clinical Terms*, una extensa terminología médica desarrollada por el *College of American Pathologists (CAP),* y mantenida por *The International Health Terminology Standards Development Organisation* (*IHTSDO*).

Disponible en inglés, español y alemán, proporciona un lenguaje común que facilita la indexación, el almacenamiento, la recuperación y la agregación de datos médicos. Los componentes básicos de SNOMED son:

- **Conceptos**: representan una unidad mínima de significado.
- **Jerarquías**: compuestas por categorías de primer nivel, que a su vez se descomponen en sub categorías.
- **Relaciones**: que enlazan conceptos entre sí. Existen dos tipos de relaciones: de tipo "es un", que conectan conceptos en una jerarquía; y las relaciones de atributos, que enlazan conceptos entre jerarquías.

[12] SNOMED International. SNOMED-CT. URL: http://www.snomed.org/snomedct

- **Descripciones**: términos o nombres asociados a un concepto, que posibilitan una mayor flexibilidad en la expresión de los conceptos médicos.

La versión actual en español recoge más de 310.000 conceptos, 974.204 descripciones y 923.000 relaciones semánticas. Los conceptos se organizan en las jerarquías fundamentales, tal y como se observa a continuación (Figura 10). Entre paréntesis se muestra un ejemplo para cada uno de los conceptos.

Hallazgo clínico

Hallazgo (edema del brazo)

Enfermedad (neumonía)

Procedimiento / intervención (biopsia de pulmón)

Entidad observable (estadio tumoral)

Estructura corporal (estructura de glándula tiroides)

Estructura morfológicamente anormal(granuloma)

Organismo (Mycobacterium tuberculosis)

Sustancia (ácido gástrico)

Producto farmacéutico/biológico(tamoxifeno)

Espécimen (espécimen de orina)

Calificador (derecho)

Elemento de registro (certificado de defunción)

Objeto físico (aguja de sutura)

Fuerza física (fricción)

Evento (inundación)

Medio ambiente/localización geográfica (unidad de cuidados intensivos)

Contexto social (donante de órganos)

Situación con contexto explícito (sin náuseas)

Estadificación y escalas (índice de Barthel)

Concepto de enlace

Aserción de enlace (tiene etiología)

Atributo (localización del hallazgo)

Concepto especial (concepto inactivo)

Figura 10 Jerarquía de conceptos en SNOMED

Unified Medical Language System

El UMLS[13] (*Unified Medical Language System*), desarrollado por la *National Library of Medicine (NLM)* de los Estados Unidos, es un sistema que garantiza referencias cruzadas entre más de treinta vocabularios y clasificaciones, incluyendo la CIE, *MESH, CPT, COSTAR, DSM IV, READ 3.1 y SNOMED.* UMLS presenta tres fuentes de conocimiento: el *Meta-tesauro*, el *Léxico Especializado* y la *Red Semántica*.

El **Meta-tesauro** es una base de datos multilingüe y multipropósito que contiene información sobre conceptos biomédicos y relacionados con la salud, incluyendo sus diferentes nombres y sus relaciones. Está construido a partir de las versiones electrónicas de diferentes tesauros, clasificaciones y listas de términos controlados utilizados en el cuidado de pacientes, en la elaboración de estadísticas sobre salud, en el indexado y la catalogación de literatura biomédica y en la investigación clínica. El meta-tesauro está organizado por conceptos o significado. Su propósito es enlazar nombres alternativos y vistas de un mismo concepto, así como identificar relaciones útiles entre diferentes conceptos. Todos ellos están asignados al menos a un tipo de la red semántica. Muchas de las palabras y términos que aparecen en el meta-tesauro también aparecen en el léxico especializado.

El **Léxico Especializado**, en lengua inglesa, contiene en su versión actual unos 108.000 informes léxicos y más de 186.000 cadenas de términos. Cada entrada presenta información sintáctica, morfológica y ortográfica. La información léxica incluye la categoría sintáctica, la variación de la inflexión (singular o plural para los sustantivos, conjugación de los verbos, comparativo y superlativo para los adjetivos y adverbios), y posible patrones de complementación (objetos y otros argumentos que pueden acompañar a los verbos, nombre y adjetivos). El léxico distingue entre once categorías sintácticas: verbos, nombres, adjetivos, adverbios, auxiliares, modales, pronombres, preposiciones, conjunciones y determinantes. Los patrones básicos de la oración se determinan por el número y la naturaleza de los complementos que rigen los verbos. Se reconocen cinco tipos generales de

[13] Unified Medical Language System (UMLS). URL: http://www.nlm.nih.gov/research/umls

complementación: intransitiva, transitiva, ditransitiva, de enlace y transitiva-compleja. Las entradas verbales contemplan las formas del verbo, si son regulares o irregulares. En cuanto a los sustantivos, se recogen patrones de pluralización y de nominalización. A continuación, y a modo de ejemplo, se muestra la información asociada a la entrada *anaesthetic* en el léxico especializado:

```
{base=anaesthetic
spelling_variant=anesthetic
entry=E0008769
cat=noun
variants=reg
entry=E0008770
variants=inv
position=attrib(3)}
```

La **Red Semántica** presenta 132 tipos semánticos, y garantiza una categorización consistente de todos los conceptos representados en el meta-tesauro. Los 53 enlaces entre los tipos semánticos establecen la estructura de la red y representan las relaciones más importantes en el dominio biomédico. Se puede decir, por lo tanto, que los tipos semánticos son los nodos en la red y las relaciones entre ellos son los enlaces. El enlace principal es el "es un", que establece la jerarquía entre los tipos de la red. Existe otro grupo de relaciones, agrupadas en cinco categorías principales: *physically_related_to, spatially_related_to, temporally_related_to, functionally_related_to* y *conceptually_related_to.* Además, los tipos semánticos se clasifican en seis agrupaciones básicas: *organismos, estructuras anatómicas, funciones biológicas, productos químicos, eventos, objetos físicos y conceptos o ideas,* y permiten la categorización semántica de un amplio abanico de terminología en múltiples dominios de especialidad.

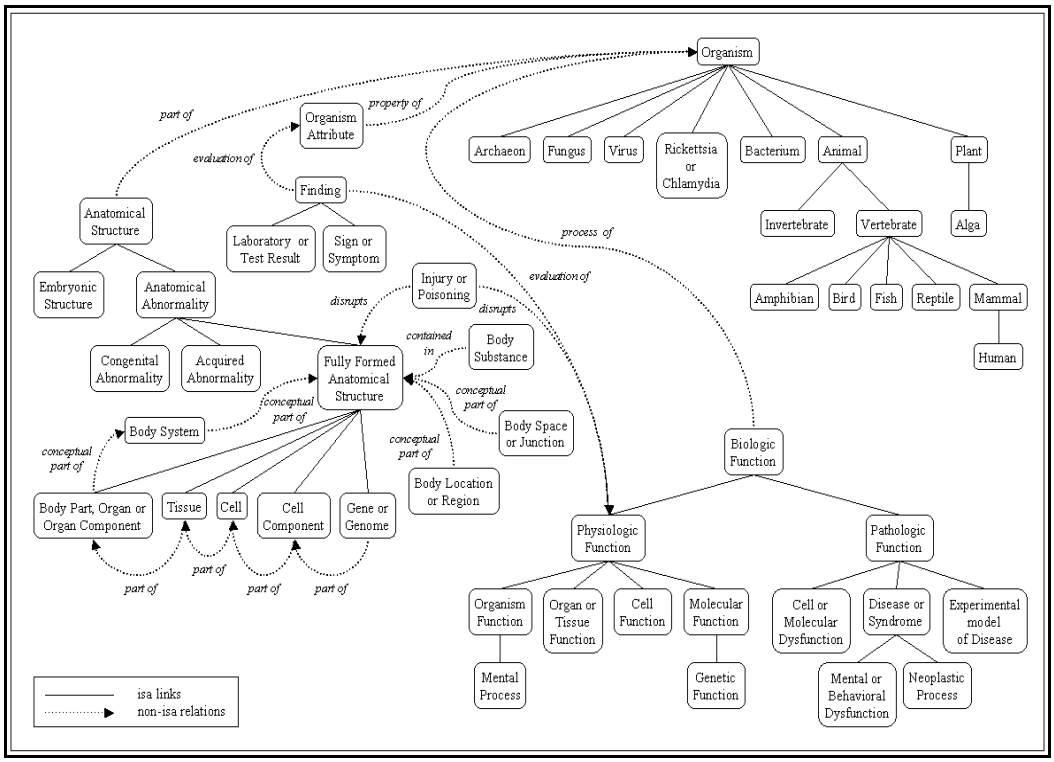

Figura 11 Red asociada al tipo semántico *Organism*

UMLS proporciona varios mecanismos de acceso a los datos del meta-tesauro, el léxico y la red semántica:

- A través de aplicaciones java, utilizando un API que permite la conexión de los programas de usuario al UMLSKS. En la implementación actual, la conexión se puede realizar vía *RMI (Java Remote Method Invocation)* o *Web Services*; no obstante, la organización pretende eliminar paulatinamente el acceso a través de RMI.
- Cargando los archivos relacionales de UMLS en una base de datos local y accediendo a ellos mediante consultas SQL. En este caso puede resultar particularmente interesante el uso de la herramienta *MetamorphoSys,* que permite a los usuarios adaptar el meta-tesauro a las necesidades particulares de su aplicación. Puesto que ha sido utilizada en este trabajo, se analizará en detalle en el apartado 3 del capítulo de Herramientas software utilizadas.

- A través de la interfaz web al UMLS Knowledge Source Server, desarrollada especialmente para la visualización y la navegación a través de los datos de UMLS (Figura 12).

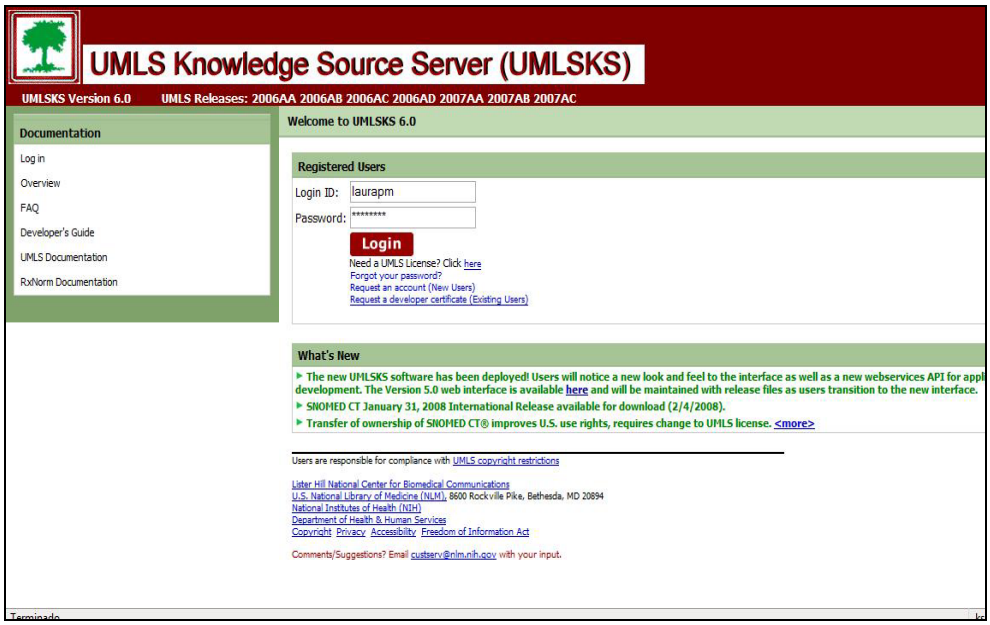

Figura 12 Interfaz Web del UMLSKS

Medical Subject Headings

Los Medical Subject Headings[14] (MeSH) son un tesauro desarrollado por la *National Library of Medicine* de los Estados Unidos, consistente en un conjunto de términos, denominados descriptores (*descriptors)*, dispuestos en una estructura jerárquica que permite la búsqueda a varios niveles de especificidad.

Los descriptores se organizan de dos modos distintos: alfabéticamente y mediante una estructura jerárquica de once niveles. En el primer nivel de la jerarquía se encuentran descriptores muy amplios, como *anatomía* o *desórdenes mentales*, mientras que conforme se desciende en la jerarquía, los descriptores se concretan, de manera que en el último nivel se encuentran

[14] Medical Subject Headings (MeSH). URL: http://www.nlm.nih.gov/mesh/

conceptos como *tobillo.* En la versión 2008 de MeSH, se cuentan 24.767 descriptores, además de 172.000 conceptos suplementarios (*Supplementary Concept Records*) recogidos en un tesauro separado. Hay también más de 97.000 términos de ayudan para localizar el descriptor más apropiado.

Los árboles de descriptores no constituyen una clasificación exhaustiva de las materias, sino que están diseñados para la búsqueda en la base de datos MEDLINE; además de como guía para las personas encargadas de asignar categorías a documentos.

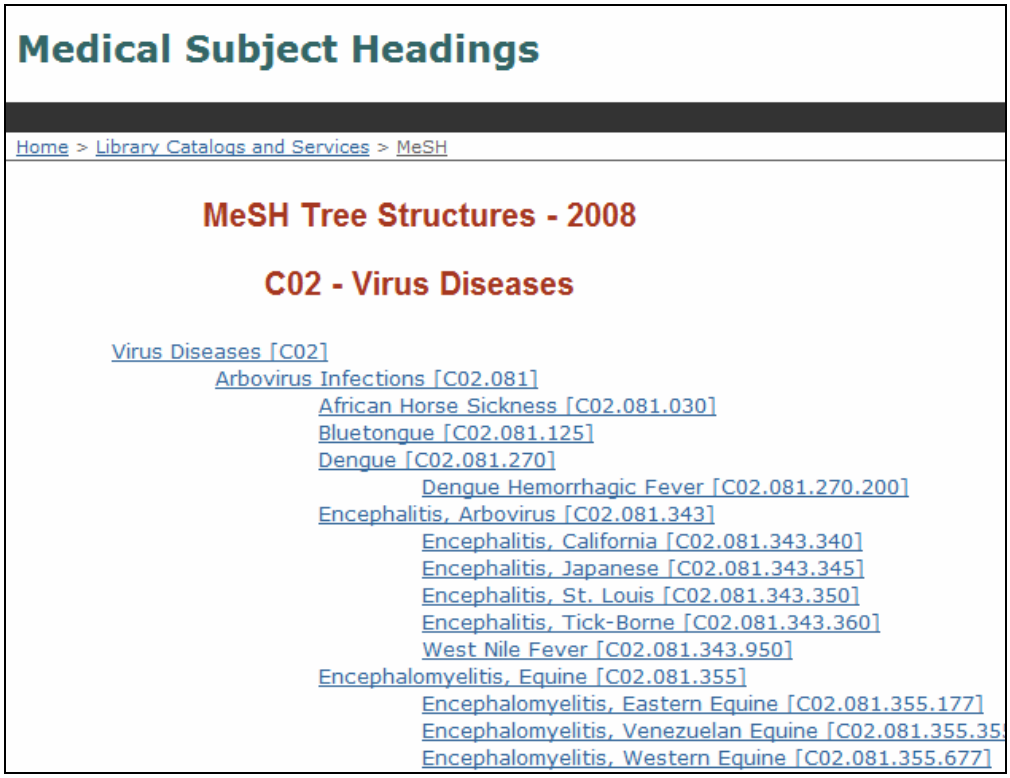

Figura 13 Organización jerárquica de los *Medical Subject Headings*

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Encephalitis, California
Tree Number	C02.081.343.340
Tree Number	C02.182.500.300.300.200
Tree Number	C02.290.310.140
Tree Number	C02.782.147.340
Tree Number	C02.782.310.340
Tree Number	C10.228.228.210.150.300.300.200
Tree Number	C10.228.228.245.340.310.140
Annotation	for La Crosse virus encephalitis, coord IM with LA CROSSE VIRUS (IM); DF: ENCEPH CALIFC
Scope Note	A viral infection of the brain caused by serotypes of California encephalitis virus (ENCEPHALI by the LA CROSSE VIRUS. This condition is endemic to the midwestern United States and prim abdominal pain followed by SEIZURES, altered mentation, and focal neurologic deficits. (From J
Entry Term	California Encephalitis
Entry Term	California Viral Encephalitis
Entry Term	Encephalitis, California, Viral
Entry Term	Viral Encephalitis, California
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX
Entry Version	ENCEPH CALIFORNIA
History Note	1991(1975)
Date of Entry	19741111
Unique ID	D004670

Figura 14 Información asociada al concepto *Encephalitis en MeSH*

En los últimos años, la utilización de ontologías en aplicaciones de procesamiento de lenguaje natural para el dominio biomédico se ha incrementado notablemente. A menudo, cuando se aborda un proyecto de cierta envergadura, y tras realizar un análisis preliminar de las bases terminológicas y de conocimiento existentes, se concluye que no son suficientes para cubrir todos los requisitos del sistema. En estos casos, la solución más obvia es desarrollar una ontología *ad hoc* para la tarea específica que se pretende resolver.

Un ejemplo muy ilustrativo es el sistema *OntoSem,* que proporciona un entorno completo para el procesamiento de textos del dominio biomédico.

Aunque inicialmente se consideró la utilización de MeSH en conjunción con el meta-tesauro de la *NLM*, finalmente se decidió desarrollar una ontología y un léxico, independientes del lenguaje, específicos para el proyecto. La ontología contiene alrededor de 7000 conceptos, cada uno de ellos descritos con una media de 16 propiedades. La siguiente figura muestra la descripción en la ontología del concepto *cáncer de colon.*

Figura 15 Información asociada al concepto *cáncer de colon* en OntoSem

1.4. Evaluación y Selección de la Ontología

A lo largo de este apartado, y habiendo descartado la posibilidad de implementar una ontología propia, analizaremos las ventajas e inconvenientes de las tres ontologías biomédicas introducidas en el apartado anterior,

exponiendo los motivos que nos han inducido a seleccionar *UMLS* para su uso en nuestra tarea de generación automática de resúmenes.

- En primer lugar, el propio propósito para el que han sido desarrolladas las tres ontologías apoya nuestra elección. UMLS está concebido como un sistema multipropósito; es decir, para ser utilizado en la construcción de aplicaciones que creen, procesen, extraigan, integren o agreguen datos biomédicos de muy diversos tipos y formatos: historiales de pacientes, literatura científica, recomendaciones sanitarias públicas, estudios estadísticos, etc. Esta polivalencia puede ser muy ventajosa en un futuro, si se desea ampliar nuestra investigación a otro tipo de documentos y a la realización de nuevas tareas de procesamiento de lenguaje natural en el dominio biomédico. Por su parte, SNOMED- CT se plantea fundamentalmente con fines asistenciales (toma de decisiones, alertas, etc.) y con fines de agregación y análisis de datos. Finalmente, MEsH está pensada para la indexación y la búsqueda en la base de datos de artículos de MEDLINE, y para catalogación de documentos, asignando titulares y tipos a las publicaciones. Tanto los contenidos como la estructuración está pensada para dicho propósito y se muestran inadecuados para la generación automática de resúmenes.
- UMLS está pensado para su uso desde aplicaciones informáticas; y por lo tanto, para ser utilizado por programadores expertos. Por este motivo, incluye un conjunto muy completo de herramientas para asistir a los desarrolladores. Aunque el API para el acceso a los datos del UMLS Knowledge Source Server permite una fácil integración con otras aplicaciones java, también se puede acceder a través de un servidor web o bien a través de sockets TCP/IP desde otro tipo de aplicaciones. Por su parte, SNOMED está más orientado a ofrecer una terminología común para ser utilizada por expertos sanitarios (médicos, investigadores) y enfermos, y no tanto por aplicaciones (aunque también se contempla esta posibilidad).
- Los tres sistemas son de disposición pública. Sin embargo, la obtención de la licencia, descarga e instalación es más sencilla en el caso de UMLS.

- UMLS Cuenta con el respaldo de un considerable número de aplicaciones que lo utilizan (*PubMed*[15], *NLM Gateway*, *ClinicalTrials.gov*, o la *Indexing Initiative* de NLM; *Enterprise Vocabulary Services* del *National Cancer Institute*, y *National Guidelines Clearinghouse* y *National Quality Measures Clearinghouse* de la *Agency for Healthcare Research and Quality's*. Por su parte, SNOMED ha sido ampliamente utilizada en el desarrollo de aplicaciones para el análisis de resultados clínicos y para el apoyo en la toma de decisiones médicas, aunque no tanto en aplicaciones de procesamiento de lenguaje natural. Actualmente ha sido adoptada a nivel nacional en el sistema de salud del Reino Unido, en Estados Unidos (edición en inglés y en castellano), y en Dinamarca. Finalmente, MEsH lo utilizan de manera habitual los catalogadores de la NLM para el análisis de material bibliográfico, la asignación de titulares a los documentos, y su indexación.
- UMLS incluye el vocabulario de SNOMED-CT, además de referencias cruzadas con otros vocabularios. También permite indexar los conceptos con los descriptores de MESH, para la clasificación y catalogación de los documentos.
- Finalmente, frente UMLS, SNOMED-CT adolece de no disponer de herramientas que faciliten el análisis léxico de los textos.

1.1. Utilización de UMLS en OBS

El método de generación automática de resúmenes propuesto utiliza UMLS para extraer los conceptos asociados al documento que se desea resumir. La primera decisión que se debe acatar tiene que ver con la forma en la que se va a acceder a la ontología. Como hemos comentado, las posibilidades se reducen al acceso al UMLSKS en remoto, mediante *web services* o *RMI*, y al acceso en local a una copia de la ontología. Las tres alternativas han sido implementadas en OBS, si bien se ha constatado que el acceso local constituye la opción más

[15] PubMed. URL: http://www.pubmed.gov/

realista, debido al elevado tiempo que consumen las otras alternativas. Sirva para clarificar esta afirmación los experimentos realizados sobre un documento de 58 oraciones utilizando ambos tipos de acceso: mientras que accediendo a la copia local la fase de recuperación de conceptos consume menos de un minuto, accediendo mediante *web services* o *RMI* se necesitan del orden de varias horas. No obstante, el acceso local tiene el inconveniente de que la copia de la ontología debe ser actualizada periódicamente, cada vez que la NLM publique una nueva versión, lo que suele ocurrir en torno a una o dos veces al año.

2. Corpus Biomédicos

El propósito de este apartado es presentar y analizar algunos corpus biomédicos, con el objetivo de seleccionar uno o varios de ellos para su utilización en este proyecto. A la hora de evaluar un corpus se han de tener en cuenta diversos aspectos. En primer lugar, desde el punto de vista biológico, los datos han de ser exactos y adecuados. En segundo lugar, es deseable que presenten anotaciones respecto de las características estructurales y lingüísticas, y que utilicen estándares. Finalmente, es importante que el diseño sea correcto, y que incluya suficientes ejemplos, tanto positivos como negativos, para el entrenamiento y el test.

- **GENIA Corpus**

Desarrollado en la Universidad de Tokio, como parte del proyecto *Tsujii Labwithin*, para la comunidad BioNLP. Está compuesto por un conjunto de más de 2000 abstracts anotados procedentes de la base de datos de la *National Library of Medicine*, MEDLINE; todos ellos del dominio de la biología molecular. El corpus ha sido anotado conjuntamente por expertos del dominio y por lingüistas, e incluye anotaciones semánticas, sintácticas y del discurso. Ha sido utilizado en aplicaciones de recuperación de información y filtrado, extracción de información, categorización y generación automática de resúmenes.

- **Medstract Corpus**

Como parte del proyecto *MEDSTRACT* (Pustejovsky et al., 2002), se ha desarrollado una base de datos de *abstract* procedentes de MEDLINE, del dominio de la genómica y la proteómica. Además de incluir etiquetado sintáctico y semántico, identifica relaciones a nivel de oración (del tipo "inhibe" o "regula") y de nombres (del tipo "inhibidor" o "regulador") e incluye expansión de acrónimos.

- **Yapex Corpus**

Se trata de un corpus reducido, que consta de 200 *abstracts* procedentes de MEDLINE, 99 para el entrenamiento y los 101 restantes para el test. Está pensado para el etiquetado de proteínas y de las posibles interacciones entre ellas (Cohen et al., 2005).

- **Cincinnati Children's Hospital Medical Center's Department of Radiology Corpus**

Desarrollado para el *Computational Medicine Center's 2007 Medical Natural Language Processing Challenge*, está formado por registros médicos, cada uno de los cuales consta de un informe radiológico, y un estudio renal. En ambos casos, se adjunta el historial clínico del paciente y una anotación en lenguaje natural sobre la impresión del médico que ha tratado al paciente. Cada una de las muestras del conjunto de entrenamiento está etiquetada con un código ICD-9-CM.

- **OHSUMED Text Collection**

Utilizado en la *TREC-9 Filtering Track*, está compuesta por un conjunto de 348,566 referencias de MEDLINE, consistentes en títulos y *abstracts* de 270 publicaciones médicas del período comprendido entre 1987-1991.

- **BMC corpus**

Construido y mantenido por *BioMed Central*[16], y concebido para la investigación en minería de texto en el dominio biomédico. Está compuesto por

[16] BioMed Central: http://www.biomedcentral.com/

más de 23900 artículos completos publicados por esta organización, incluyendo una versión estructurada de los mismos en XML.

- **PennBioIE**

Desarrollado como parte del proyecto del mismo nombre de la Universidad de Pennsylvania, para tareas de extracción de información, consta de 2258 *abstracts* de MEDLINE de dos dominios distintos: inhibición de enzimas de la clase CYP450 (1000 ejemplos) y genética molecular en oncología (1158 ejemplos). Todas las muestras están anotadas por párrafos, oraciones, parte del discurso y un conjunto de tipos de entidades biomédicas definidas específicamente para el proyecto y para cada dominio por expertos del *Children's Hospital of Philadelphia*. Además, 324 de las muestras de CYP y 318 de oncología están anotadas sintácticamente.

- **Corpus Técnico del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona**

Recopila textos en cinco idiomas (catalán, castellano, inglés, francés y alemán) de distintos dominios. En concreto, el apartado de medicina recoge artículos completos de MEDLINE, anotados morfológica y sintácticamente, clasificados en distintas áreas de especialización.

2.1. Selección del corpus para OBS

Puesto que para la evaluación de nuestro sistema nos interesa disponer de un corpus de artículos científicos completos, y no sólo de sus *abstracts*, la decisión se reduce al corpus de BioMed Central y al de la Universidad Pompeu Fabra. Nos hemos de decantado por el primero de ellos, debido a que sólo éste se encuentra disponible públicamente.

Capítulo 4

Herramientas Software Utilizadas

El propósito de este capítulo es presentar las diferentes herramientas software que, no habiendo sido desarrolladas expresamente para este proyecto, han sido utilizadas en el mismo. En primer lugar, se describen los objetivos y el funcionamiento de la arquitectura *GATE*, así como su utilización para la realización de diversas tareas de procesamiento de lenguaje natural. En segundo lugar, se presenta *MetaMap*, una aplicación que permite mapear textos a conceptos del meta-tesauro de UMLS. Finalmente, se dedica un breve apartado a comentar la utilidad *MetamorphoSys,* empleada para adaptar el meta-tesauro a los objetivos y necesidades particulares de nuestra aplicación.

1. GATE

GATE[17] (*Generic Architecture for Text Engineering*) es una conocida infraestructura para el desarrollo de software de procesamiento de lenguaje natural, desarrollada por la Universidad de Sheffield desde 1995, y en continua evolución desde entonces. Surge con el objetivo de facilitar el trabajo de científicos y desarrolladores especificando una arquitectura para la construcción de aplicaciones de ingeniería lingüística, y proporcionando un

[17] GATE (Generic Architecture for Text Engineering): http://gate.ac.uk/

framework que implementa dicha arquitectura y un entorno gráfico para el desarrollo de los distintos componentes que generalmente se encuentran presentes en toda aplicación de este tipo.

Para GATE, todos los elementos que componen un sistema software de procesamiento de lenguaje natural pueden clasificarse en tres tipos de componentes, denominados *resources.*

- ***Language Resources (LRs)***, que representan entidades como documentos, corpus u ontologías.
- ***Processing Resources (PRs)***, que representan entidades que son en su mayoría algoritmos, como parsers, generadores, etc.
- ***Visual Resources (VRs),*** que representan la visualización y edición de los componentes de la GUI.

El conjunto de recursos integrados en GATE reciben el nombre de CREOLE (*Collection of REusable Objects for Language Engineering*). Todos los recursos se encuentran empaquetados en un archivo JAR, junto con otros ficheros XML de configuración. Pero además de sus propios componentes GATE incorpora *plugins* a otros desarrollados por diferentes organizaciones.

GATE presenta dos modos de funcionamiento. Un modo gráfico y un API para Java. Tanto si se utiliza el entorno gráfico de desarrollo como si se accede a través del API, los desarrolladores pueden crear recursos de los tres tipos. El entorno de desarrollo puede utilizarse para visualizar las estructuras de datos producidas y consumidas en el procesamiento, para depurar, obtener medidas de rendimiento, etc.

1.1. ANNIE

ANNIE (*A Nearly-New Information Extraction System)* es el componente de GATE orientado a la Extracción de Información. Incorpora un amplio abanico de recursos que acometen tareas de análisis del lenguaje a distintos niveles. Los componentes de ANNIE se organizan secuencialmente, tal y como se observa en la Figura 16.

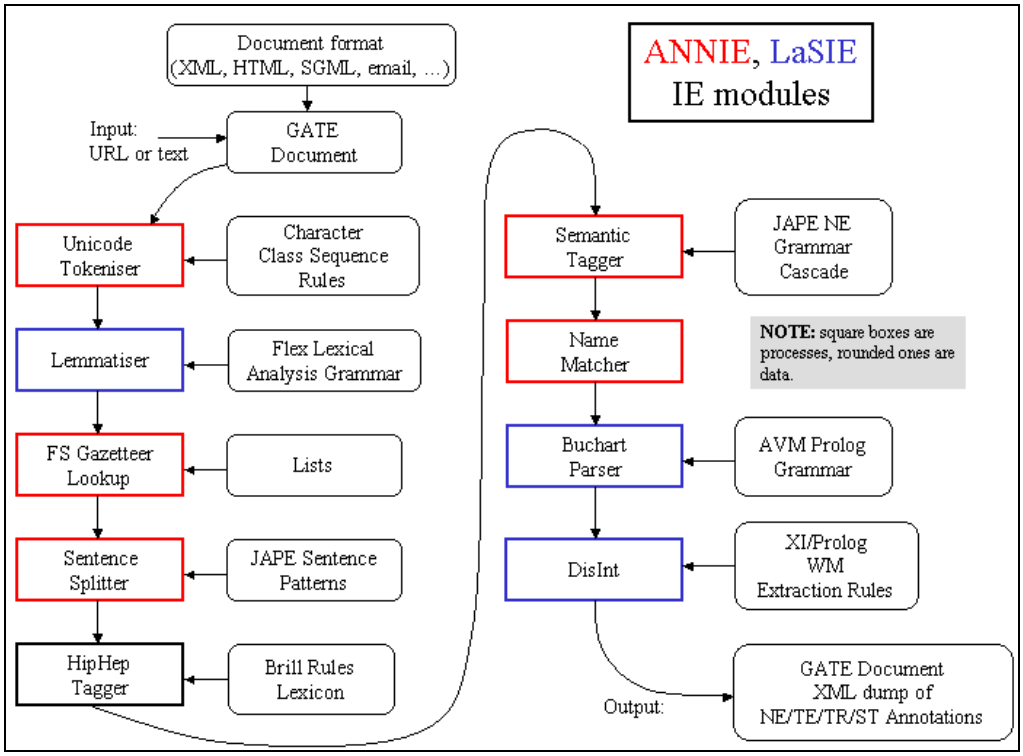

Figura 16 Componentes de GATE

A continuación se proporciona una breve explicación de los módulos más importantes del sistema.

- ***Tokeniser***. Realiza la división del texto en unidades simples (*tokens*) tales como números, símbolos, signos de puntuación, espacios y palabras de distintos tipos.
- ***Gazetter.*** Consiste en un conjunto de listas (ficheros de texto plano) en las que se representan conjuntos de nombres, tales como nombres de ciudades, organizaciones, días de la semana, etc., y que se utilizan para reconocer en el texto dichas entidades. Estas listas pueden ser editadas e incluso se pueden crear otras nuevas. A continuación, se muestra un extracto de la lista disponible en GATE para las unidades de moneda.

• FFr	• New Taiwan dollar
• Fr	• New Taiwan dollars
• German mark	• NT dollar
• German marks	• NT dollars

- ***Sentence Splitter.*** Consiste en un conjunto de transductores de estados finitos que permiten segmentar el texto en oraciones. Utiliza la lista de abreviaturas del *gazzetter* para distinguir los "." que indican el final de una abreviación de aquellos que delimitan las oraciones. Cada oración resultante es anotada con el tipo "Sentence", mientras que a cada delimitador se asocia una etiqueta "Split". Es independiente del dominio y de la aplicación.
- ***Part of Speech Tagger.*** Es una versión modificada del etiquetador *Brill*, que realiza la anotación de cada palabra o símbolo del texto con su categoría morfológica (*POS Tagging*). Para ello, utiliza un léxico por defecto y un conjunto de reglas extraídas del entrenamiento sobre un corpus extenso de noticias del Wall Street Journal. Existe la posibilidad de modificar tanto el léxico como las reglas.
- ***Semantic Tagger.*** El etiquetador semántico está basado en el lenguaje JAPE, y contiene un conjunto de reglas que actúan sobre las etiquetas asignadas en las etapas anteriores para anotar las entidades.
- ***Orthographic Coreference (OrthoMatcher).*** Añade relaciones de identidad entre las entidades anotadas por el etiquetador semántico, con el objetivo de detectar posibles referencias anafóricas. Para ello, utiliza una tabla de cadenas que representan la misma entidad, como por ejemplo, "IBM" y "Big Blue" o "Coca Cola" y "Coke"; y una tabla de cadenas que, erróneamente, se podrían interpretar como representantes de la misma entidad, pero que corresponden a entidades distintas, como por ejemplo, "BT Wireless" y "BT Cellnet".

- ***Pronominal Coreference.*** Realiza la resolución de referencias pronominales; es decir, identifica a qué entidad de las mencionadas anteriormente en el texto se refiere un determinado pronombre. De nuevo, precisa de las anotaciones realizadas por los otros módulos.

1.2. Corpus, Documentos y Anotaciones

GATE presenta un único modelo de información para la descripción de documentos, corpus y anotaciones, basada en pares atributo/valor. Los atributos son cadenas de caracteres, mientras que los valores son objetos JAVA. Tanto los corpus como los documentos en GATE son tipos de *Language Resources*, cada uno de los cuales tiene asociado un conjunto de características que definen información sobre el recurso. Además de estas características, un documento se define como un contenido textual y una serie de anotaciones sobre dicho contenido. En cuanto a formatos de documentos se refiere, GATE soporta texto plano, HTML, SGML, XML, RTF y email.

1.3. Otros recursos en GATE

Además de los anteriores, GATE ofrece soporte a otras tareas comunes en las aplicaciones actuales de PLN.

- Incluye un módulo para Recuperación de Información (IR) basado en *Lucene*[18].
- Permite realizar búsquedas en *WordNet*[19] directamente desde GATE.
- Permite el entrenamiento y la utilización de algoritmos de Aprendizaje Automático, directamente desde GATE, o gracias a distintos *wappers* a librerías como *Open NLP MAXENT*[20] o *WEKA*[21]

[18] http://lucene.apache.org/

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

- Proporciona soporte para el uso de ontologías, a través de una API que provee un modelo unificado de ontología y que permite el acceso a los formalismos de representación más aceptados (*RDF-Schema*, *OWL* y *DAML-OIL*). El modelo definido presenta una jerarquía de clases con un alto nivel de expresividad. Permite representar taxonomías de conceptos, instancias y herencia entre ellos; y definir propiedades (relaciones entre conceptos e instancias y propiedades) con restricciones de cardinalidad, simetría o transitividad. Por último, incluye una interfaz gráfica para la edición y búsqueda en ontologías.

1.4. Utilización de GATE en OBS

Como paso previo a la generación del resumen, el documento de origen debe ser sometido a un proceso destinado a identificar y extraer las oraciones que lo compone; y a etiquetar morfosintácticamente los *tokens*. Para ello, se ha utilizado la herramienta GATE, a través de su interfaz gráfica.

A continuación se detallan los pasos a realizar para completar el proceso de extracción de oraciones.

- En primer lugar, se ha de crear un *Language Resource* de tipo *GATE document*. El documento original se encuentra en formato XML, por lo que es perfectamente compatible con GATE (Figura 17).
- El siguiente paso consiste en crear una aplicación, de tipo *pipeline*, y cargar los componentes necesarios. En este caso, se han de cargar los módulos *English Tokeniser*, *Gazetter*, *Sentence Splitter* y *Part of Speech Tagger*, en el orden especificado (Figura 18).
- Para terminar, ejecutamos la aplicación sobre el documento cargado y guardamos el resultado como documento XML.

[20] http://maxent.sourceforge.net/about

[21] http://www.cs.waikato.ac.nz/ml/weka/

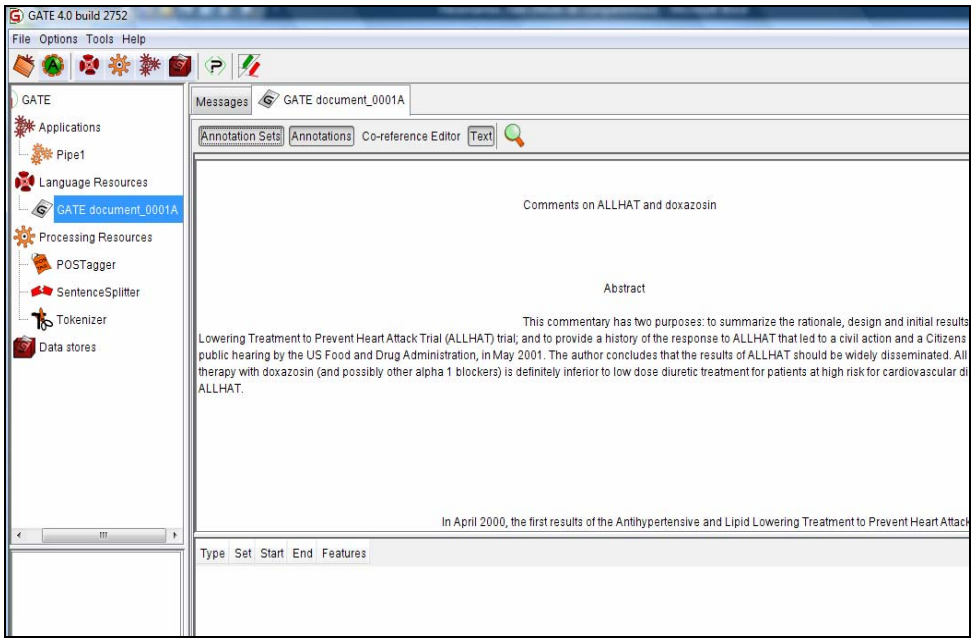

Figura 17 Creación de un recurso lingüístico en GATE

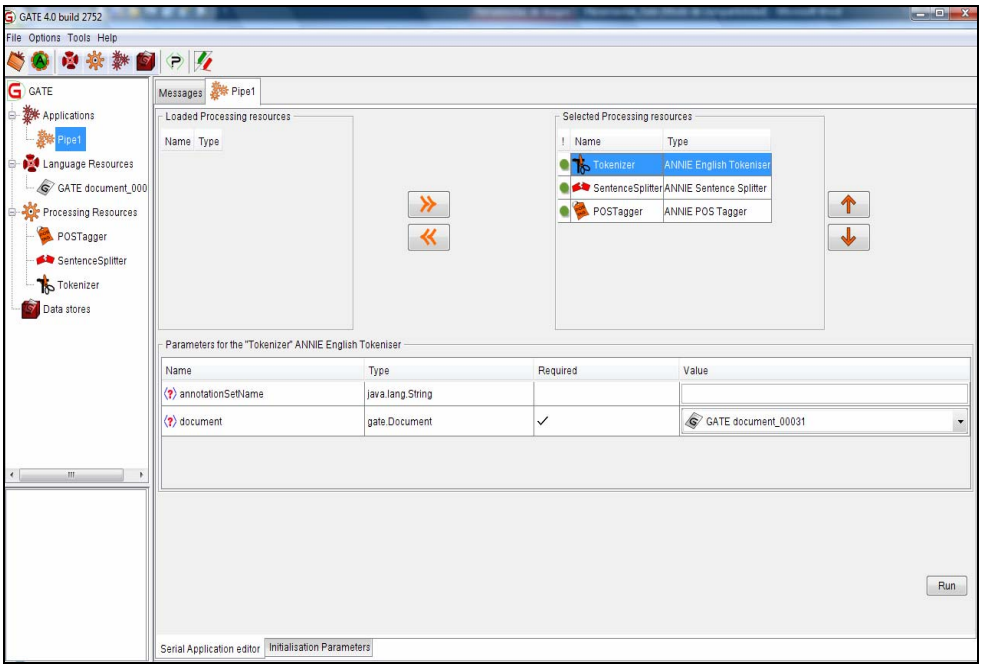

Figura 18 Creación de una aplicación en GATE

- Este documento, será el que se utilice como entrada para el generador de resúmenes. En él se identifican claramente las distintas oraciones, los tokens y su tipo (palabra, signo de puntuación, símbolo, número, espacio, etc.) y la categoría morfológica de cada palabra (nombre común singular, verbo gerundio, adjetivo superlativo, etc.)

2. *MetaMap*

MetaMap[22] es un programa desarrollado por la Biblioteca Nacional de Medicina (*National Library of Medicine,* NLM) de los Estados Unidos, en el marco del *Indexing Initiative System,* para el mapeo de textos biomédicos a conceptos del Meta-tesauro de UMLS.

MetaMap utiliza un enfoque intensivo en conocimiento, basado en técnicas lingüísticas y de procesamiento de lenguaje natural, haciendo uso de herramientas como el léxico especializado de UMLS. Se trata de una herramienta muy configurable, que permite especificar, entre otras cosas, el grado en que se han de ignorar los términos muy genéricos, si se ha de respetar la ordenación de las palabras en el texto, o las terminologías que se desean utilizar para el mapeo de conceptos.

Aunque fuera inicialmente concebido como soporte en la recuperación de material bibliográfico de MEDLINE, a partir de peticiones formuladas en inglés, lo cierto es que su utilidad se extiende a la resolución de todo tipo de problemas de recuperación de información, minería de texto, categorización y clasificación, generación de resúmenes o descubrimiento de conocimiento. A modo de ejemplo, se citan algunos trabajos donde ha sido utilizado:

- Extracción de drogas, genes, y relaciones entre ellas en literatura biomédica (Rindflesch et al., 1999)

[22] MetaMap Transfer(MMTx): http://mmtx.nlm.nih.gov/

- Identificación de terminología anatómica en textos médicos (Sneiderman et al., 1998)
- Categorización de textos biomédicos (Perea et al., 2008)

2.1. Motivación

El mapeo de los términos presentes en un documento a conceptos de la ontología presenta diversos problemas, que pueden solventarse utilizando MetaMap.

- Si únicamente se tienen en cuenta términos individuales, en muchos casos los conceptos indexados no reflejarán la verdadera semántica del texto. Por ejemplo, el sintagma nominal *coronary heart disease* enlazaría con tres conceptos distintos del meta-tesauro: *Coronary, Heart* y *Disease,* en lugar de indexar con el concepto único *Coronary Heart Disease.*
- Si se realiza el mapeo exacto de una palabra o sintagma nominal, sin considerar posibles variantes léxicas o semánticas, frecuentemente no se recupera ningún concepto o bien se recuperan conceptos que no son los adecuados. Por ejemplo, el sintagma *phrenic motoneurones* no tiene ningún concepto asociado en UMLS. Ahora bien, si se consideran sus sinónimos, se obtendría el concepto *Motor Neurons.*
- El mismo problema se presenta si no se consideran coincidencias parciales o complejas. Por ejemplo, *obstructive sleep apnea* se corresponde con el término *Obstructive apnea* en el meta-tesauro. Por su parte, *intensive care medicine* se corresponde con *Intensive Care* y *Medicine.*
- Por último, incluso encontrándose el término exacto en el meta-tesauro, éste puede ser ambiguo y puede tener distintos conceptos asociados. Por ejemplo, el meta-tesauro contiene dos conceptos para el término *ventilation,* uno relacionado con el flujo de aire en los edificios, y otro relacionado con la respiración.

2.2. Funcionamiento del Algoritmo

El programa MetaMap acepta como entrada el documento que se desea procesar. Seguidamente, para cada oración, ejecuta los siguientes pasos:

1. Se parsea el texto para extraer los sintagmas nominales.
2. Para cada sintagma identificado, se generan una serie de variantes, que consisten en combinaciones de una o varias de las palabras que lo forman, junto con sus variaciones morfológicas inflexionales y derivacionales, abreviaturas, acrónimos y sinónimos.
3. Se obtienen del Meta-tesauro de UMLS los posibles candidatos, extrayendo los términos que contienen alguna de las variantes.
4. Para cada candidato, se utiliza una función de evaluación para calcular una medida de la fuerza del mapeo, y se ordenan los candidatos de mayor a menor.
5. Se combinan los candidatos con otras partes no adyacentes del sintagma nominal, se vuelven a calcular las puntuaciones y se seleccionan aquellos candidatos con mayor puntuación, formando un conjunto de "mejores candidatos" para el sintagma original.

A continuación se muestra un ejemplo del conjunto de candidatos extraídos para el sintagma *Heart attack trial*.

```
Phrase: "Heart Attack Trial"
Meta Candidates (8)
   827 Trial (Clinical Trials) [Research Activity]
   734 Heart attack (Myocardial Infarction) [Disease or Syndrome]
   660 Heart [Body Part, Organ, or Organ Component]
   660 Attack, NOS (Onset of illness) [Finding]
   660 Attack (Attack device) [Medical Device]
   660 attack (Attack behavior) [Social Behavior]
   660 Heart (Entire heart) [Body Part, Organ, or Organ
Component]
   660 Attack (Observation of attack) [Finding]
Meta Mapping (901)
   734 Heart attack (Myocardial Infarction) [Disease or Syndrome]
   827 Trial (Clinical Trials) [Research Activity]
```

Puede observarse cómo de todos los posibles candidatos, los que mayor puntuación obtienen son *Trial*, con una puntuación de 827, que además es el único candidato asociado al término *trial*, y *Heart Attack*, con una puntuación de 734.

2.3. Opciones de Configuración

Como ya adelantamos en la introducción, MetaMap presenta un extenso abanico de opciones de configuración, controlables a través de un conjunto de *flags* de estado. A continuación se enumeran las más importantes.

- **Opciones de Datos:** Determinan los vocabularios y los modelos de datos utilizados por MetaMap. Permiten especificar, por ejemplo, las terminologías que se desean indexar o si se utiliza un modelo de datos *estricto*, *moderado* o *relajado*.
- **Opciones de Procesamiento:** Controlan el comportamiento interno de MetaMap. Permiten, por ejemplo, establecer una preferencia por la indexación de términos de mayor longitud, especificar que no se desean considerar las variantes derivacionales, que se debe ignorar el orden de las palabras o que se debe realizar desambiguación siempre que sea posible.
- **Opciones de salida:** Permiten seleccionar el formato y el contenido de la información producida como resultado de la ejecución del programa. Por ejemplo, se pueden mostrar todos los candidatos indexados, o únicamente los mejores candidatos, mostrar los tipos semánticos asociados a los conceptos recuperados o restringir la recuperación a aquellos candidatos cuya evaluación sea superior a un determinado umbral.

2.4. Utilización de MetaMap en OBS

El papel de MetaMap en nuestro generador de resúmenes se limita al mapeo de las oraciones del documento, previamente extraídas utilizando GATE, a

conceptos del meta-tesauro de UMLS. Para ello, se ha optado por utilizar MetaMap a través del API desarrollado a tal efecto, integrando así su funcionalidad dentro del propio generador. La recuperación de los conceptos del documento se realiza como paso previo a su extensión con los conceptos relacionados en la jerarquía (a través de las relaciones *is a* y *associated with*), y a la construcción del grafo del documento.

Para justificar el uso de MetaMap, a continuación se muestra el resultado de extraer los conceptos asociados a la oración *This prospective, randomized trial was designed to compare a diuretic (chlorthalidone) with long-acting (once-a-day) drugs among different classes: angiotensin-converting enzyme inhibitor (lisinopril); calcium-channel blocker (amlodipine); and alpha blocker (doxazosin),* mediante dos procedimientos: utilizando MetaMap e indexando uno a uno los términos en el meta-tesauro.

1. **Utilizando MetaMap.** Se recuperan los siguientes conceptos asociados a la oración:

```
1.  C0008976|Clinical Trials|827
2.  C0012798|Diuretics|1000
3.  C0008294|Chlorthalidone|1000
4.  C0205166|Long|814
5.  C0439228|day|861
6.  C0013227|Pharmaceutical Preparations|1000
7.  C0443199|Differential quality|594
8.  C0456387|Class|861
9.  C0003015|Angiotensin-
    Converting Enzyme Inhibitors|1000
10. C0065374|Lisinopril|1000
11. C0006684|Calcium Channel Blockers|1000
12. C0051696|Amlodipine|1000
13. C0001641|Adrenergic alpha-Antagonists|1000
14. C0114873|Doxazosin|1000
```

2. **Indexando individualmente los términos de la oración en el meta-tesauro.** Se recuperan los siguientes conceptos asociados a la oración:

```
1.  C0012798|Diuretic
2.  C0008294|Chlorthalidone
3.  C0205166|Long
4.  C0439228|Day
5.  C1720092| Once - dosing instruction fragment
6.  C0013227|Pharmaceutical Preparations
7.  C0456387|Class
8.  C0003018|Angiotensins
9.  C0014442|Enzymes
10. C0065374|Lisinopril
11. C0006675|Calcium
12. C0439799|Channel
13. C0051696|Amlodipine
14. C0439095|Greek letter alpha
    (qualifier value)
15. C0114873|Doxazosin
```

El resultado obtenido utilizando el primero de los procedimientos es indudablemente más apropiado para nuestro propósito. En primer lugar, porque recupera un menor número de conceptos, lo que se traducirá en un menor tamaño de los grafos de las oraciones a construir, y en una consiguiente mejora de la eficiencia. En segundo lugar, porque al indexar términos complejos en lugar de únicamente términos individuales, la interpretación semántica de la oración es más correcta. A modo de ejemplo, la recuperación del concepto único *Angiotensin-Converting Enzyme Inhibitors* es más precisa que la recuperación de los dos conceptos distintos *Angiotensin* y *Enzyme.* Otro ejemplo de interpretación semántica más adecuada por parte de *Metamap,* es la recuperación del concepto *Adrenergic alpha- Antagonists* para el sintagma

nominal *alpha blocker,* en lugar de recuperar incorrectamente el concepto *Greek letter alpha.*

3. MetamorphoSys

MetamorphoSys es la herramienta de configuración y personalización del Meta-tesauro, incluida en la propia distribución de UMLS. En general, un usuario puede estar interesado en crear subconjuntos personalizados del meta-tesauro por dos motivos:

- Para excluir vocabularios que no necesita en su aplicación, que puede ser de un dominio restringido a unas pocas terminologías de las incluidas en UMLS. De esta forma, se reduce significativamente el tamaño de la ontología y se facilita su manejo.
- Para modificar el formato de los datos de salida y aplicarles distintos filtros.

3.1. Utilización de MetamophoSys en OBS

La decisión de utilizar MetamorphoSys en nuestro trabajo viene motivada fundamentalmente por la necesidad de acceder a la ontología en local, ya que los experimentos realizados para acceder a través de RMI y de Web Services al UMLS Knowledge Source Server han demostrado un elevado tiempo de cómputo en la comunicación y una excesiva dependencia del servidor, que en ocasiones no se encuentra disponible por tareas de mantenimiento y mejora.

Por otra parte, se ha decidido restringir el vocabulario utilizado a las siguientes terminologías, estableciendo además el siguiente orden de precedencia entre ellas:

1. *SNOMED Clinical Terms* (SNOMED-CT).
2. *Medical Subject Headings* (MeSH).
3. *National Cancer Institute Metathesaurus* (NCI Thesaurus).

Una vez ejecutado *Metamorphosys*, el resultado es un conjunto de archivos de extensión *ORF (Original Release Format)* o *RRF (Rich Release Format)* que contienen los subconjuntos seleccionados del meta-thesauro, el léxico y la red semántica, junto con un script para cargarlos en una base de datos *oracle* o *mysql*.

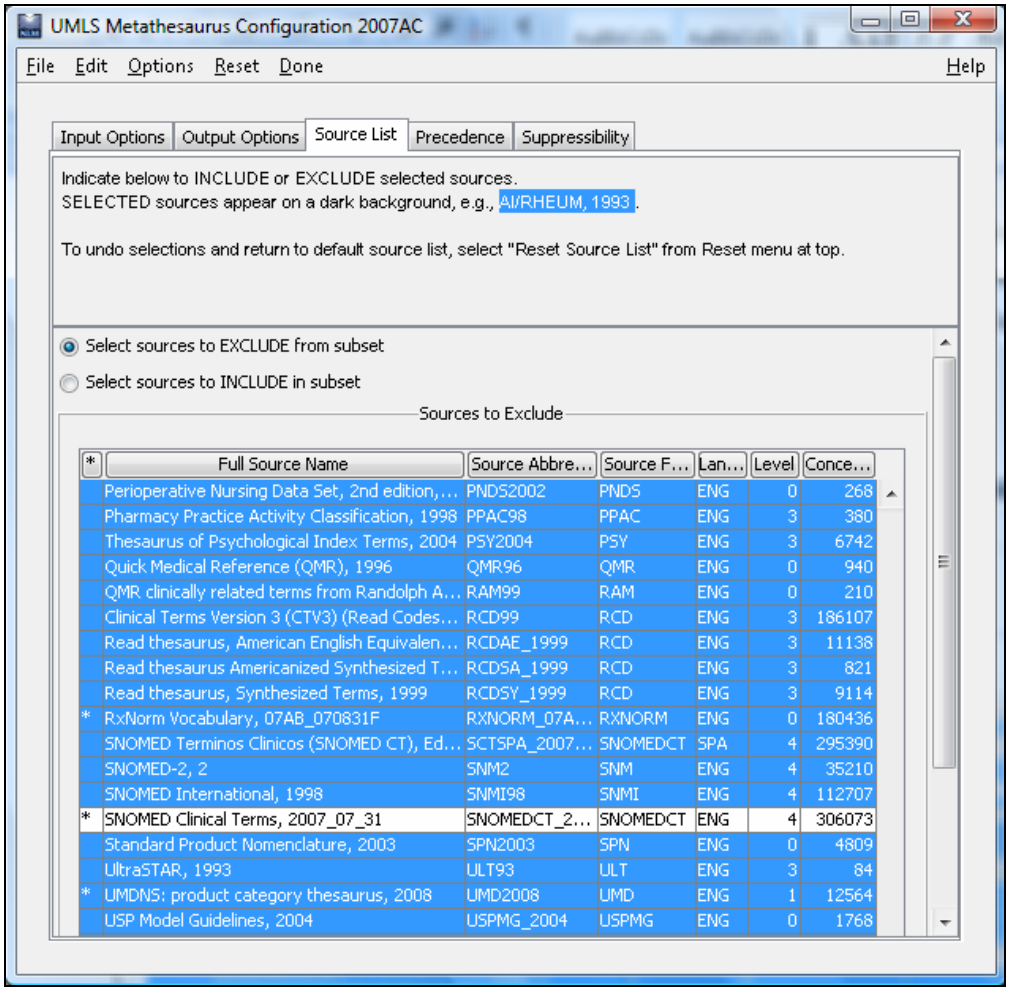

Figura 19 Selección de fuentes en *MetamorphoSys*

Capítulo 5

Método Propuesto

En este apartado se describe el método propuesto para resolver la tarea, a través de las distintas etapas que conducen a la elaboración del resumen (Plaza et al., 2008). El enfoque que se presenta realiza un análisis superficial del documento para la selección de oraciones relevantes, utilizando los conceptos de UMLS para construir una representación basada en grafos de las oraciones y del documento. Esta representación conceptual permite capturar las relaciones semánticas entre los elementos textuales y determinar con mayor precisión el tema central del documento, lo que permite construir resúmenes de mayor calidad.

La Figura 20 muestra los componentes del sistema. En primer lugar, el documento de entrada se somete a un pre-procesamiento lingüístico, del que se obtienen las oraciones que, en la etapa posterior, serán representadas gráficamente utilizando la ontología. Seguidamente, los distintos grafos de las oraciones se combinan en un único grafo del documento. A continuación, se aplica un algoritmo de agrupamiento basado en la detección de vértices *concentradores* o *hub* como centroides de los clusters, para obtener grupos de grafos a los que se asignan las distintas frases del documento. Finalmente, se seleccionan y reordenan las oraciones a partir de las cuales se generará el resumen. Cada una de estas etapas se describe detalladamente en los siguientes apartados.

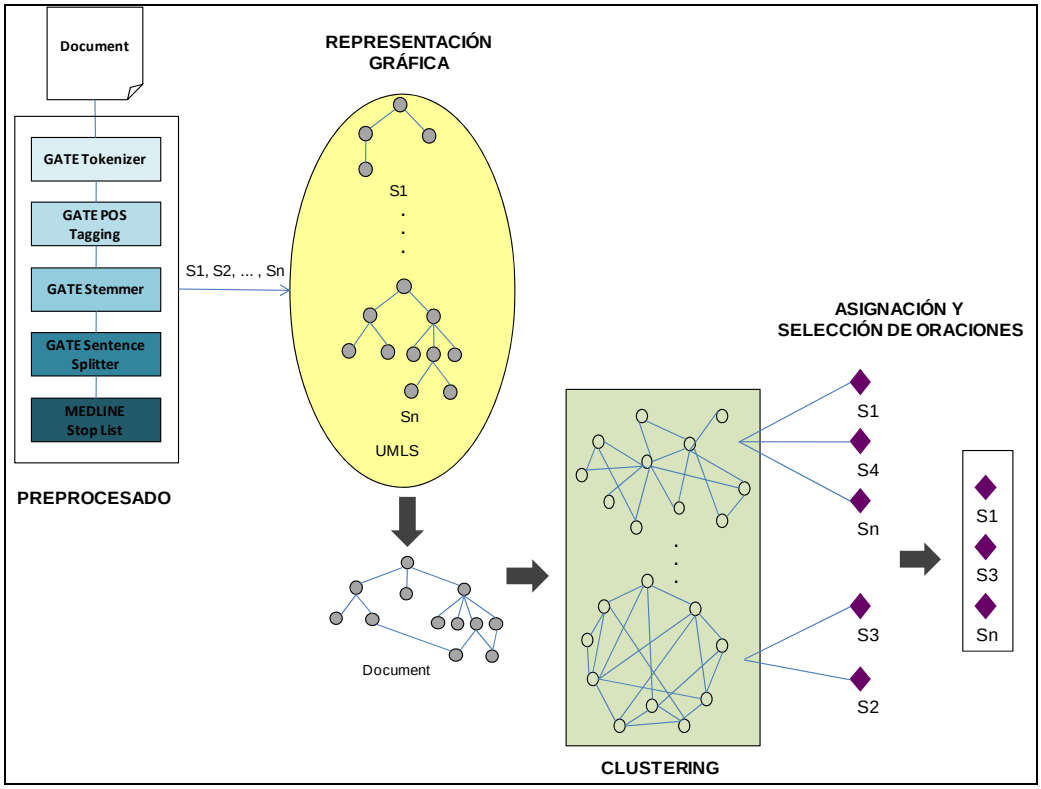

Figura 20 Arquitectura del módulo de extracción de oraciones

Con el objetivo de clarificar el funcionamiento del algoritmo, a lo largo de la exposición se hará referencia a un ejemplo concreto de generación de un resúmen para uno de los documentos del corpus. A continuación se muestra un extracto del documento utilizado.

> In April 2000, the first results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALL- HAT) were published summarizing the comparison between two of the four drugs studied (chlorthalidone and doxazosin) as initial monotherapy for hypertension. This prospective, randomized trial was designed to compare a diuretic (chlorthalidone) with long-acting (once-a-day) drugs among different classes: angiotensin-converting enzyme inhibitor (lisinopril); calcium-channel blocker (amlodipine); and alpha blocker (doxazosin). The diuretic had been the mainstay of several previous trials, particularly the Systolic Hypertension in the Elderly Program (SHEP) study. During the first three years of the trial, the Data Safety and Monitoring Committee became aware of different event rates between two groups and, after considerable deliberation, made the decision to discontinue the arm assigned to doxazosin for two stated reasons. One reason was...

El texto completo presenta un total de 58 oraciones, y puede encontrarse en el anexo II.

1. Etapa I: Preprocesamiento

Como paso previo a la generación del resumen, el documento es sometido a un tratamiento preliminar, con el objetivo de prepararlo para las posteriores etapas: el texto se divide en tokens, se realiza su etiquetado morfosintáctico, y se divide en oraciones. Para ello, se han utilizando los módulos *English Tokenizer*, *Gazetter, Sentence Splitter* y *Part of Speech Tagger* de la librería GATE (ver apartado 1 del capítulo 4). Finalmente, se eliminan las palabras genéricas utilizando una lista de parada construida y publicada por MEDLINE[23], así como los términos que presentan una alta frecuencia en el documento, puesto que no van a ser de utilidad a la hora de discriminar entre contenidos importantes e irrelevantes. Por lo tanto, como resultado de esta etapa, se obtienen las oraciones del documento preparadas para su traducción a la ontología en la siguiente etapa.

2. Etapa II: Traducción de las oraciones a conceptos de la ontología

El objetivo de esta etapa es traducir el léxico del documento a los conceptos de la ontología. Para ello, se ejecuta sobre cada una de las oraciones el programa *Metamap* (ver apartado 2 del capítulo 4), obteniéndose como resultado un listado de los conceptos del meta-tesauro de UMLS presentes en cada oración.

Considérese como ejemplo la oración: *The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat*

[23] PubMed StopWords: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords

hypertension. Como resultado de la ejecución de *Metamap* se obtienen los siguientes conceptos:

```
1.      C0018017|Goals|1000
        (T078|Idea or Concept:T170|Intellectual ProducT)
2.      C0008976|Clinical Trials|1000
        (T062|Research Activity)
3.      C0007226|Cardiovascular system|694
        (T022|Body System)
4.      C0038454|Cerebrovascular accident|1000
        (T047|Disease or Syndrome)
5.      C0010068|Coronary heart disease|1000
        (T047|Disease or Syndrome)
6.      C0018802|Congestive heart failure|1000
        (T047|Disease or Syndrome)
7.      C0178499|Base|627
        (T120|Chemical Viewed Functionally)
8.      C0302614|Guide device|827
        (T074|Medical Device)
9.      C1292734|Treats|966
        (T185|Classification)
10.  C0020538|Hypertensive disease|1000
        (T047|Disease or Syndrome)
```

El primer elemento de cada fila es el *Concept Unique Identifier (CUI)* o identificador único del concepto en el meta-tesauro (*C0018017*). El segundo elemento es un literal o cadena de caracteres que describe a dicho concepto (*Goals*). El tercer elemento es la puntuación que *Metamap* ha asignado al concepto en la traducción y que indica su grado de adecuación en función de los términos de la oración y de su contexto (1000). A continuación, se listan los tipos semánticos a los que pertenecen los conceptos en la red semántica de UMLS, representados por su *Unique identifier of Semantic Type (TUI)* o identificador único del tipo semántico en la red (*T078*), y una cadena de caracteres que describe a dicho tipo semántico (*Idea or Concept*).

Los conceptos con un significado muy general no aportan información a la hora de identificar los temas del documento y de discriminar entre oraciones relevantes e irrelevantes. Por lo tanto, pueden ser ignorados a la hora de construir los grafos de las oraciones y del documento, consiguiendo representaciones más compactas, y en consecuencia, mejorar el rendimiento de la aplicación. Los tipos semánticos de UMLS se pueden utilizar para identificar los términos asociados a conceptos muy generales. La Tabla 3 muestra los niveles superiores de la jerarquía de tipos semánticos en UMLS. La jerarquía completa puede consultarse en el anexo III de este documento.

ENTITY
Physical Object Organism Anatomical Structure Manufactured Object Substance Organic Chemical Conceptual Entity Idea or Concept Finding Organism Attribute Intellectual Product Language Occupation or Discipline Organization Group Attribute Group
EVENT
Activity Behavior Daily or Recreational Activity Occupational Activity Machine Activity Phenomenon or Process Human-caused Phenomenon or Process Environmental Effect of Humans Natural Phenomenon or Process Injury or Poisoning

Tabla 3 Tipos Semánticos en UMLS

Tras un estudio del significado de cada tipo y de los conceptos que agrupan se ha determinado que se podrían ignorar los conceptos pertenecientes a los tipos *Quantitative Concept*, T*emporal Concept*, *Idea or Concept,*

Intellectual Product, Mental Process, Spatial Concept y *Language*. A continuación se muestran algunos de los conceptos que, estando presentes en nuestro documento de ejemplo, pertenecen a los tipos semánticos anteriores y, por tanto, no serán considerados para la construcción del grafo del documento.

- **Quantitative Concept:** Lowered, Two, Four, Several.
- **Temporal Concept:** Previous, Year, Seconds, Frequent.
- **Idea or Concept:** Reasons, Complete, Goal, Accepted.
- **Intellectual Product:** Class, Groups, Agencies, Reports.
- **Mental Process:** Awareness, Initiation, Euphoric mood.
- **Spatial Concept:** Upper, Separate, Address, Over.
- **Language**: Ninguna aparición en el documento. Posibles conceptos de este tipo semántico serían: Spanish, English.

Como resultado de este proceso, los conceptos asociados a la oración del ejemplo se reducen a los siguientes:

```
1. C0008976|Clinical Trials|1000(T062|Research Activity)

2. C0007226|Cardiovascular system|694(T022|Body System)

3. C0038454|Cerebrovascular accident|1000(T047|Disease or Syndrome)

4. C0010068|Coronary heart disease|1000(T047|Disease or Syndrome)

5. C0018802|Congestive heart failure|1000(T047|Disease or Syndrome)

6. C0178499|Base|627(T120|Chemical Viewed Functionally)

7. C0302614|Guide device|827(T074|Medical Device)

8. C1292734|Treats|966(T185|Classification)

9. C0020538|Hypertensive disease|1000(T047|Disease or Syndrome)
```

El número final de conceptos asociados al documento de nuestro ejemplo es de 345. Es importante resaltar que se trata de un número relativamente pequeño. Experimentos realizados sobre artículos más extensos muestran en torno a 1000-2000 conceptos.

3. Etapa III: Representación de las oraciones como grafos

El objetivo de esta etapa es construir una representación de cada oración en forma de grafo, de manera que capture la estructura semántica de los términos y las relaciones entre ellos. Para ello, los conceptos descubiertos para cada oración en la etapa anterior se expanden con los conceptos de niveles superiores en la jerarquía (hiperónimos), a través de las relaciones *is_a* del meta-tesauro de UMLS. De esta forma se consigue tratar de manera homogénea conceptos con un significado similar y dar solución, entre otros, al problema presentado en la introducción.

El proceso de expansión de un determinado concepto con sus hiperónimos comienza obteniendo el CUI (*Concept Unique Indentifier*) correspondiente del meta-tesauro, y extrayendo sus ancestros en la jerarquía de conceptos de la base de datos en la que hemos almacenado la ontología. A modo de ejemplo, la Figura 21 muestra el resultado del proceso para la palabra *cardiovasculars*.

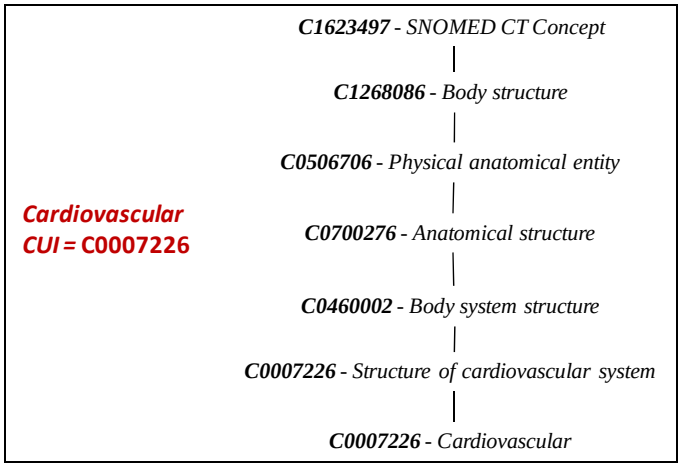

Figura 21 Hiperónimos del concepto *Cardiovascular*

Bajo la hipótesis de que los conceptos que se encuentran en los primeros niveles de la jerarquía representan información muy genérica, se ha decidido ignorar los dos primeros niveles (*SNOMED CT Concept* y *Body structure).*

Una vez expandidos todos los conceptos con sus hiperónimos, se combinan en un único grafo que representa a la oración. La Figura 1 muestra el

grafo construido para la oración que nos ha servido como ejemplo en el apartado anterior: *The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension.*

Seguidamente, a cada una de las aristas del grafo se le asigna un peso que es directamente proporcional a la profundidad de los conceptos que une en la jerarquía; es decir, será tanto mayor cuanto más específicos sean los conceptos que conecte. Para el cálculo de estos pesos se han evaluado dos medidas distintas de similitud entre conjuntos:

- La utilizada en (Yoo et al., 2007), que calcula la similitud según la Ecuación 5, donde α representa el conjunto de todos los ancestros de un concepto determinado, incluido el propio concepto, y β representa el conjunto de todos los ancestros del concepto del nivel inmediatamente superior, incluido el propio concepto.

$$\frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} = \frac{|\beta|}{|\alpha|}$$

Ecuación 5

- El ***índice de Jaccard*** para el cálculo de la similitud entre ambos conjuntos (Ecuación 6), donde el significado de α y β es el mismo que para la ecuación anterior.

$$\frac{|\alpha \cap \beta|}{|\alpha| + |\beta| + |\alpha \cup \beta|}$$

Ecuación 6

La Figura 23 muestra los pesos asociados a las aristas de un subgrafo extraído de la oración del ejemplo anterior, utilizando la métrica de Yoo (Figura 23.a) y el índice de Jaccard (Figura 23.b) respectivamente. Puede observarse que los pesos calculados según el índice de Jaccard son menores y decrecen menos escalonadamente. Por ello, se ha optado por utilizar la primera métrica en el trabajo definitivo.

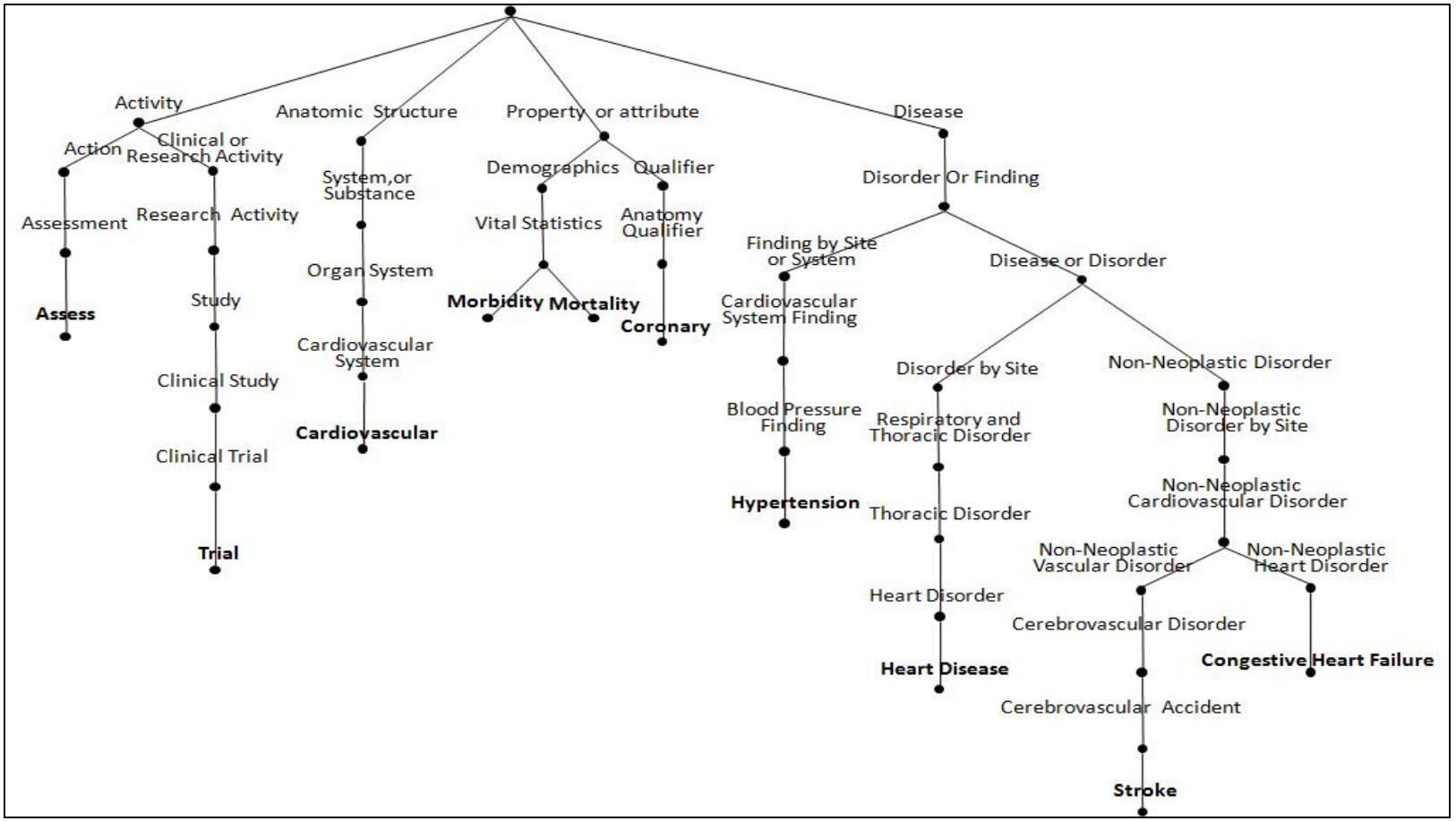

Figura 22 Grafo de una oración

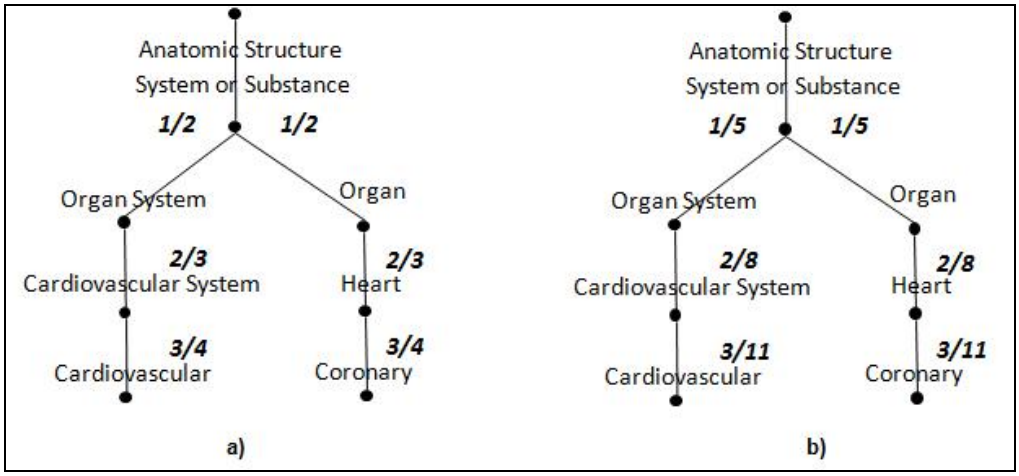

Figura 23 Asignación de pesos

4. *Etapa IV: Construcción del grafo del documento*

La cuarta etapa del algoritmo consiste en fusionar los grafos de las distintas oraciones en un único grafo que represente la relación semántica entre los términos de todo el documento. En (Yoo et al., 2007), el grafo se completa añadiendo relaciones de coocurrencia de conceptos en el conjunto de los documentos. En este proyecto, se han estudiado las relaciones existentes entre los tipos semánticos de UMLS, antes de decidir cuál de ellas utilizar. La Tabla 4 muestra las posibles relaciones a considerar.

isa associated_with physically_related_to part_of consists_of contains connected_to interconnects branch_of tributary_of ingredient_of spatially_related_to location_of adjacent_to surrounds traverses	performs carries_out exhibits practices occurs_in process_of users manifestation_of indicates result_of temporally_related_to co occurs_with precedes conceptually_related_to evaluation_of

functionally_related_to affects manages treats disrupts complicates interacts_with prevents brings_about produces causes	degree_of analyzes assesses_effect_of measurement_of measures diagnoses property_of derivative_of developmental_form_of method_of conceptual_part_of issue_in

Tabla 4 Relaciones en la Semantic Network de UMLS

De entre todas las anteriores, por su significado y por los resultados obtenidos en los experimentos preliminares realizados, se ha optado por utilizar la relación *associated with.* De esta forma, y a modo de ejemplo, los conceptos *trial* y *hypertension* están relacionados en el grafo, ya que sus respectivos tipos semánticos (*Research Activity* y *Disease or Syndrome*) presentan una relación *associated with* entre ellos en UMLS (Figura 24).

Junto a esta relación entre tipos semánticos de la red, también se ha estudiado la relación *other related* entre conceptos del meta-thesauro.

El peso de estos nuevos enlaces en el grafo del documento se calculará siguiendo el mismo criterio que para las relaciones *is a.*

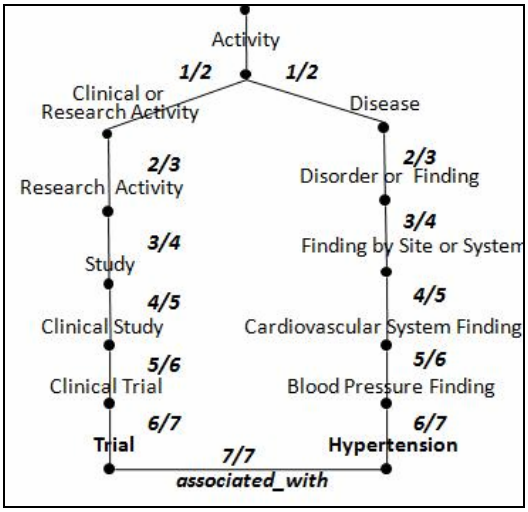

Figura 24 Relaciones *associated_with* entre conceptos

5. *Etapa V: Clustering de conceptos. Identificación de subtemas*

El propósito de esta etapa es realizar una agrupación de los conceptos del grafo del documento, utilizando para ello un algoritmo de clustering basado en la conectividad (*degree-based method*) (Erkan y Radev, 2004), donde cada cluster puede verse como una red de conceptos que mantienen una estrecha relación semántica entre sí. En este contexto, cada cluster representa un *theme* o tema del documento; y dentro de ellos, los conceptos centrales (*centroides*) aportan la información necesaria y suficiente de cada tema.

Ferrer-Cancho y Solé (2001) han demostrado que los grafos que representan la relación entre las palabras en los textos en inglés constituyen una *red libre de escala* (Barabasi y Albert, 1999). Se parte, por lo tanto, de la hipótesis de que el grafo obtenido forma una red de este tipo. Una red libre de escala (*scale-free network)* es un tipo específico de red compleja en la que algunos nodos están altamente conectados (nodos *hub*); es decir, poseen un gran número de enlaces a otros nodos, aunque el grado de conexión de casi todos los nodos es bastante bajo. Esta propiedad surge como consecuencia de dos características que pueden observarse fácilmente en las redes reales: en primer lugar, la red se encuentra en continua construcción, mediante la adición dinámica de nuevos vértices y, en segundo lugar, los nuevos vértices muestran preferencia por relacionarse con los vértices con mayor conectividad. Estas observaciones contrastan con la tradicional teoría aleatoria de grafos (*random graph theory)* de Erdos y Rényi (1959), que asume que las redes comienzan con un número fijo de vértices que no se modifica durante la vida de la red, y que la probabilidad de que dos vértices estén conectados es aleatoria y uniforme.

Barabasi y Albert (1999) muestran que, independientemente del sistema, la probabilidad *P(k)* de que un vértice de la red interactúe con otros *k* vértices, sigue una distribución $P(k) \approx k^{-\gamma}$, lo que indica que las redes de cierta complejidad se auto organizan en redes libres de escala. De hecho, este modelo es muy común en las redes lingüísticas, puesto que se presenta tanto en las redes de co-ocurrencia de palabras, como en las redes de asociación de conceptos o en las de dependencia sintáctica.

Partiendo de esta idea, el algoritmo SFGC (*Scale Free Graph Clustering*) comienza con la localización en el grafo del documento del conjunto de nodos más conectados. Siguiendo a (Yoo et al., 2007), se define el prestigio o *salience* de cada vértice (v_i) como la suma de los pesos de todas las aristas (e_j) que tienen como origen o destino a dicho vértice, de acuerdo con la Ecuación 7.

$$salience(v_i) = \sum_{e_j \mid \exists v_k \wedge e_j conecta(v_i, v_k)} weight(e_j)$$

Ecuación 7

Los vértices de mayor *salience* se denominan *hub vertices,* y representan los más nodos más conectados del grafo, tanto en relación al número de aristas como al peso de las mismas.

El algoritmo de agrupamiento comienza seleccionando los *n* vértices de mayor *salience.* Para determinar el valor de *n*, se han realizado distintos experimentos, en los que se ha considerado, respectivamente, *n*=5%, 10% y 20% del número total de conceptos en el grafo. Para el ejemplo que nos ocupa, por lo tanto, *n* toma los siguientes valores (Tabla 5).

5%	10%	20%
17	34	69

Tabla 5 Valores experimentales de *n*

A continuación, los *hub vertices* se agrupan formando *Hub Vertex Sets (HVS)*, que constituirán los centroides de los clusters a construir. Para ello, en una primera etapa, el algoritmo propuesto busca iterativamente para cada *hub vertice,* y entre los demás, aquel al que se encuentra más conectado, uniéndolos en un único *HVS.* En la segunda etapa, para cada par de *HVS* se comprueba si sus conectividades internas son menores que la conectividad entre ellos. De ser así, los dos *HVS* se fusionan. Esta decisión obedece a la hipótesis de que, idealmente, la conectividad entre los conceptos dentro un cluster ha de ser máxima, mientras que la conectividad entre conceptos de distintos clusters

debe ser mínima. Es necesario, por tanto, definir tales medidas de intra-conectividad e inter-conectividad (Ecuación 8 y Ecuación 9).

$$intraconectividad\ (HVS_i) = \sum_{e_j \mid \exists v,w \in HVS_i \wedge e_j conecta(v,w)} weight(e_j)$$

Ecuación 8

$$interconectividad\ (HVS_i, HVS_j) = \sum_{e_k \mid \exists v \in HVS_i, w \in HVS_j \wedge e_k conecta(v,w)} weight(e_k)$$

Ecuación 9

Con el objetivo de determinar el número de *hub vértices* a utilizar para conseguir una configuración adecuada de *HVS*, se ha realizado una serie de experimentos en los que, además de variar el número de *hub vertices,* se han considerado distintas relaciones a la hora de construir el grafo del documento. Los resultados obtenidos en los distintos experimentos se muestran en la Tabla 6. El número de *hub vertices* viene expresado como tanto por ciento del número total de conceptos en el grafo.

En los experimentos preliminares se ha observado que, a pesar de que el grafo del documento es bastante conexo, muchos de los *hub vertices* no establecen ninguna relación entre sí, por lo que algunos de los *HVS* resultantes sólo contienen un concepto. Puesto que se considera que un único concepto no puede ser representativo de un tema o tópico del documento, los *HVS* unitarios no serán tenidos en cuenta, y los vértices que los componen serán tratados como si no fueran *hub vertices,* y asignados a los *HVS* en una etapa posterior.

Número de *Hub Vertices*	**Relaciones entre conceptos**	**Número de HVS**	**Número de HVS de tamaño mayor que 1**
5%	***associated with***	6	3
10%	***associated with***	13	5
20%	***associated with***	6	4

5%	***other related***	14	3
10%	***other related***	22	8
20%	***other related***	32	13
5%	***associated with other related***	5	2
10%	***associated with other related***	13	5
20%	***associated with other related***	7	3

Tabla 6 Parametrización del número de hub vertices para el clustering

En primer lugar, se observa que la relación *other related*, cuando se utiliza en exclusividad, da como resultado número de *HVS* demasiado elevado. Esto se debe a que esta relación conecta entre sí a pocos conceptos del meta-tesauro, por lo que el grafo del documento resulta poco conexo. En cuanto a utilizar la relación *associated with* únicamente o junto a la relación *other related*, los resultados son similares, pero el tiempo de cómputo se incrementa notablemente en este último caso, como consecuencia de los accesos a la base de datos, que tienen lugar sobre una tabla de gran tamaño. Por último, un examen detallado de los conceptos agrupados en cada uno de los *HVS* nos ha llevado a decidir utilizar un número de *hub vertices* igual al 20% del tamaño del grafo, construido únicamente a partir de relaciones *is a* y *associated with.* A continuación se muestran los conceptos que componen los cuatro *Hub Vertices Sets* generados.

HVS 1 (3 conceptos)

- Study
- View
- Provide

HVS 2 (28 conceptos)

- Analysis of substances
- Blood
- receptor
- Other therapy NOS
- Discontinued
- Administration occupational activities
- Descriptor
- Guide device

- Reduction - action
- Therapeutic procedure
- Finding
- Primary operation
- Entire lung
- Entire heart
- Hepatic
- Entire upper arm
- Base
- Reserpine
- Agent
- Audiological observations
- Age
- Very large
- Related personal status
- In care
- Cardiovascular event
- Cardiovascular system
- heart rate
- Adverse reactions

HVS 3 (4 conceptos)

- Systolic hypertension
- Blood Pressure
- May
- Person

HVS 4 (32 conceptos)

- Duplicate concept
- Articular system
- Entire hand
- Body system structure
- Diastolic blood pressure
- Expression procedure
- Prevention
- Chlorthalidone
- Hopelessness
- Lisinopril
- Doxazosin
- Reporting
- Lower
- Support, device
- Antihypertensive Agents
- Falls
- Angiotensin-Converting Enzyme Inhibitors
- Prazosin
- Immune Tolerance
- Tissue damage
- Ramipril
- Amlodipine
- Clonidine
- Hydralazine
- Adrenergic beta-Antagonists
- Diuretics
- PREVENT
- CONCEPT Drug
- Calcium Channel Blockers
- Assessment procedure
- Qualifier value
- Unapproved attribute

Una vez construidos los *HVS*, el siguiente paso consiste en asignar el resto de vértices (es decir, aquellos que no son *hub vertices*) al *HVS* con respecto al cual presenten una mayor conectividad. De este modo, se obtienen los clusters de conceptos finales. La asignación se realiza calculando el grado de conexión del concepto a asignar con cada *HVS*, según la Ecuación 10, reajustando los *HVS* y los vértices asignados en un proceso iterativo.

$$conectividad(v, HVS_i) = \sum_{e_j \mid \exists w \in HVS_i \wedge e_j conecta(v,w)} weight(e_j)$$

Ecuación 10

Para el ejemplo que nos ocupa, se han generado los siguientes clusters.

Cluster 1 (33 conceptos)

- Study
- View
- Provide
- Inactive concept
- Prevents
- Deficiency
- Functional finding
- System
- Observable entity
- Immunologic function
- General finding of observation of patient
- Nervous system function
- Location within home premises
- Attribute
- Social and personal history finding
- Reason not stated concept
- Chemical categorized structurally
- Topographical modifier
- Rank
- Morphologically abnormal structure
- Proliferation
- Non-metal elements and their compounds or derivatives
- Thoracic structure
- Cardiovascular drug
- Cardiovascular observable
- Neck, chest and abdomen
- Environment or geographical location
- Moved elsewhere
- Chest, abdomen, and pelvis
- Monitoring procedure
- Events
- Chemical compound
- Hypertensive disorder

Cluster 2 (132 conceptos)

- Analysis of substances
- Blood
- receptor
- Other therapy NOS
- Reduction - action
- Therapeutic procedure
- Finding
- Primary operation
- Entire lung
- Entire heart
- Hepatic
- Entire upper arm
- Base
- Reserpine
- Agent
- Discontinued
- Administration occupational activities
- Descriptor
- Guide device
- Age
- Very large
- Related personal status
- In care
- Cardiovascular event
- Cardiovascular system
- heart rate
- Adverse reactions
- Substance
- Educational establishment
- Administrative procedure
- Elderly
- Physicians
- Chest and abdomen
- Drug allergen or pseudoallergen
- Community environment
- Upper extremity part
- Ambiguous concept
- Procedure with explicit context
- Veterinary proprietary
- Extended series patch test substance
- Geographical and/or political region of the world
- Orthopedic device
- Materials
- Organic natural material
- Sequela
- Structural modification
- Congestive heart failure
- Modifier related to clinical specialty AND/OR occupation
- Research administrative procedure
- Body part structure
- Blind Vision
- Liaising with

- Audiological observations
- Clinical equipment and/or device
- Alpha 2 adrenergic agonist
- Plant alkaloid
- Heart AND pericardium structure
- Large
- Cerebrovascular accident
- Cerebrovascular disease
- Minors
- Upper arm structure
- United States
- Exposure to mechanical force
- Biomedical equipment
- Market
- Geographic state
- Healthcare professional
- Upper limb structure
- Allergen or pseudoallergen
- Poisoning by drug AND/OR medicinal substance
- History finding
- Various patch test substance
- Geographical environment
- General site descriptor
- Heart structure
- Post-starting action status
- Thoracic cavity structure
- Lung structure
- Vital signs
- Disorder of cardiac ventricle
- Being organized
- Disorder of cardiac function
- Primary procedure
- Body structure
- Class
- Plant product
- drug AND/OR biological
- Lipids
- Special concept
- Education/welfare/health professions
- Quinazoline
- Alpha-adrenoceptor agonist
- Mediastinal structure
- Clinical Trials
- Contact allergen
- Finding relating to complex and social behaviors
- Confrontational behavior
- Cardiac finding
- Morphologically altered structure
- Instrument
- administrative procedure
- Body part structure
- Blind Vision
- Liaising with
- Social context
- Findings values
- Neck, chest, abdomen, and pelvis
- Sulfur AND/OR sulfur compound
- Physical anatomical entity
- Sulfur compound
- Location inside building
- Sulfonamide
- Drug pseudoallergen by function
- Sympathomimetic agent
- Treats
- Situation with explicit context
- Disorder of the central nervous system
- Coronary heart disease
- Disorder of thorax
- North American country
- Pre-starting action status
- Rauwolfia alkaloid
- Body disability AND/OR failure state
- Structure of musculoskeletal system
- Function
- Person in the healthcare environment
- Social context
- Findings values
- Neck, chest, abdomen, and pelvis
- Sulfur AND/OR sulfur compound
- Physical anatomical entity
- Sulfur compound
- Location inside building
- Cardiac finding
- Morphologically altered structure
- Instrument
- Extended series patch test substance
- Geographical and/or political region of the world
- Orthopedic device
- Materials
- Organic natural material
- Sequela
- Structural modification
- Congestive heart failure
- Modifier related to clinical specialty AND/OR occupation
- Research
- Table - furniture
- Further education establishment
- Person categorized by age
- Anatomical structure
- Heart failure
- General body state finding
- General adjectival modifier
- Procedure
- Commercial premises
- Environment

Cluster 3 (9 conceptos)

- Systolic hypertension
- Person
- Baresthesia
- Thrombocytopenic disorder

- Blood Pressure
- May
- Ventricular Function, Left
- Notable event
- Individual

Cluster 4 (171 conceptos)

- Duplicate concept
- Articular system
- Entire hand
- Body system structure
- Diastolic blood pressure
- Expression procedure
- Prevention
- Chlorthalidone
- Hopelessness
- Lisinopril
- Doxazosin
- Reporting
- Lower
- Support, device
- Antihypertensive Agents
- Falls
- Angiotensin-Converting Enzyme Inhibitors
- Prazosin
- Immune Tolerance
- Tissue damage
- Ramipril
- Amlodipine
- Clonidine
- Hydralazine
- Adrenergic beta-Antagonists
- Diuretics
- PREVENT
- CONCEPT Drug
- Calcium Channel Blockers
- Assessment procedure
- Qualifier value
- Unapproved attribute
- Verbal
- Preventive monitoring
- Meetings
- Finding by site
- Disease
- Clinical finding
- Discontinuation
- Using
- Medical
- Adjectival modifier
- result
- Degree findings
- Duplicate
- Sustained
- Device
- Step
- Disorder by body site
- Body region structure
- Alpha-adrenergic blocking agent
- Organic compound
- Wrist and hand structures
- Medical specialty
- Relative sites
- Nature of procedure values
- Clinical history/examination observable
- Poisoning
- Level of hope - finding
- Disorder of mediastinum
- Mental state, behavior and/or psychosocial function finding
- Mental state finding
- Eye / vision finding
- Techniques
- General information qualifier
- Clinical specialty
- Complication
- Hand structure
- Finding by method
- Allergen class
- Finding of body region
- Organism
- Finding of vision of eye
- Monitoring of patient
- Age AND/OR growth period
- Limb structure
- Megakaryocytic thrombocytopenia
- Failure
- Availability
- Planned
- Body fluid
- Substance categorized structurally
- Residence and
- Procedure by method
- Additional values
- Psychological finding
- Sensory function
- Traumatic AND/OR non-traumatic injury
- Occupation
- Drug pseudoallergen
- Drug allergen
- Residential environment
- Municipal and civic establishment
- Injury of anatomical site
- Alpha 1 adrenergic blocking agent
- Conversions
- Exposure to potentially harmful entity
- Lower respiratory tract structure
- Environmental finding
- Spatial and relational concepts
- Biological substance
- Housing, local environment and transport finding
- Home
- Disorder of nervous system
- Platelet disorder
- Vascular disorder
- General categories of people
- Lipid or lipoprotein
- Disorder of brain
- Sulfone - chemical
- Disorder of artery
- Room of building
- Sulfonamide diuretic
- Myocardial Infarction
- College
- Other and unspecified drug and medicament poisoning
- Disputes
- cardiology discipline
- Body substance
- Disorder of

- Arterial pulse pressure
- Clinical action
- Pharmaceutical / biologic product
- Behavior descriptors
- Numerical descriptors
- Disorder of body system
- Reaction
- More
- Including
- Origins
- Sensibilities
- Natures
- Impaired
- Clinical history and observation findings
- Structure of shoulder and/or upper arm
- Context values
- Manipulation
- Country
- accommodation circumstances - finding
- Disorder of cardiovascular system
- Heterocyclic compound
- Finding of trunk structure
- Mood finding
- Growth alteration
- Patch test substance
- Disorder of trunk
- Finding of personal status
- Cardiovascular finding
- Feature of left ventricle
- Behavior finding
- Meetings and conferences
- Emotional state finding
- Pulse rate
- Linkage concept
- hemostatic system
- Classification
- Physical object
- Changing
- Heart disease
- Patients
- Vasodilator
- Vasodilator antihypertensive drugs
- Trade and service environment
- Context values for actions
- Regimes and therapies
- Medical practitioner
- Natural material
- Ended
- Plant material
- Aging
- Lead
- Cardiac feature

6. Etapa VI: Asignación de oraciones a grafos

El propósito de esta etapa es asignar cada una de las oraciones a uno de los clusters anteriores. Para ello, es preciso definir una medida de la similitud entre el cluster y el grafo de la oración. Es importante aclarar que, puesto que ambas representaciones son muy distintas en cuanto a tamaño se refiere, las métricas clásicas de similitud entre grafos (i.e. la distancia de edición) no resultan adecuadas. En su lugar, se utiliza un mecanismo de votos (Yoo et al., 2007), por el que cada vértice (v_k) de una oración (O_j) asigna a cada cluster (C_i) en el que se encuentra presente una puntuación ($w_{k,j}$) distinta dependiendo de si pertenece o no al HVS de dicho cluster (Ecuación 11).

$$similitud(C_i, O_j) = \sum_{v_k | v_k \in O_j} w_{k,j}$$

Ecuación 11

$$\text{donde } \begin{cases} v_k \notin C_i \Rightarrow w_{k,j} = 0 \\ v_k \in HVS(C_i) \Rightarrow w_{k,j} = \gamma \\ v_k \notin HVS(C_i) \Rightarrow w_{k,j} = \delta \end{cases}$$

Los valores de γ y δ se han establecido a 1.0 y 0.5 respectivamente, lo que significa que se atribuye el doble de importancia a los conceptos que pertenecen a los HVS que a los restantes.

La Tabla 7 muestra la similitud entre las oraciones y los diferentes clusters.

Oración	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	98.0	172.0	74.0	208.0
2	13.0	21.0	9.0	18.0
3	8.0	13.0	4.0	17.0
4	8.0	19.0	4.0	16.0
5	9.0	13.0	4.0	16.0
6	6.0	4.0	4.0	8.0
7	1.0	1.0	1.0	3.0
8	8.0	15.0	4.0	25.0
9	3.0	9.0	2.0	10.0
10	9.0	10.0	6.0	10.0
11	4.0	9.0	2.0	14.0
12	5.0	7.0	3.0	8.0
13	16.0	29.0	11.0	25.0
14	7.0	5.0	5.0	15.0
15	3.0	10.0	1.0	9.0
16	11.0	11.0	7.0	15.0
17	4.0	5.0	2.0	6.0
18	5.0	11.0	3.0	17.0
19	5.0	6.0	2.0	4.0
20	17.0	24.0	10.0	25.0
21	16.0	31.0	12.0	27.0
22	11.0	23.0	8.0	27.0
23	3.0	8.0	4.0	14.0
24	14.0	17.0	9.0	27.0
25	7.0	27.0	6.0	20.0
26	5.0	7.0	3.0	12.0
27	5.0	6.0	2.0	8.0
28	2.0	5.0	1.0	2.0

29	10.0	22.0	8.0	18.0
30	6.0	7.0	4.0	7.0
31	9.0	14.0	5.0	12.0
32	4.0	6.0	2.0	5.0
33	3.0	3.0	2.0	2.0
34	18.0	18.0	9.0	20.0
35	8.0	15.0	5.0	17.0
36	1.0	6.0	1.0	2.0
37	8.0	12.0	4.0	12.0
38	6.0	6.0	4.0	9.0
39	0.0	2.0	0.0	6.0
40	2.0	2.0	1.0	2.0
41	1.0	4.0	1.0	8.0
42	9.0	14.0	6.0	17.0
43	14.0	14.0	8.0	26.0
44	11.0	12.0	7.0	22.0
45	3.0	6.0	2.0	9.0
46	5.0	16.0	2.0	18.0
47	3.0	5.0	2.0	11.0
48	4.0	11.0	3.0	20.0
49	7.0	9.0	4.0	12.0
50	5.0	10.0	4.0	13.0
51	3.0	9.0	3.0	6.0
52	2.0	6.0	1.0	6.0
53	2.0	2.0	2.0	4.0
54	0.0	0.0	0.0	0.0
55	0.0	0.0	0.0	0.0
56	2.0	5.0	2.0	4.0
57	12.0	14.0	7.0	22.0
58	4.0	7.0	2.0	7.0

Tabla 7 Similitud entre oraciones y clusters

Por lo tanto, el mapeo final de oraciones a clusters queda de la siguiente manera:

	Tamaño	**Oraciones**
Cluster 1	**2**	33 40
Cluster 2	**18**	2 4 10 13 15 19 21 25 28 29 30 31 32 36 37 51 52 56 58
Cluster 3	**2**	54 55
Cluster 4	**36**	1 3 5 6 7 8 9 11 12 13 14 16 17 18 20 22 23 24 26 27 34 35 38 39 41 42 43 44 45 46 47 48 49 50 53 57

Tabla 8 Asignación de oraciones a clusters

7. Etapa VII: Selección de las oraciones relevantes

El último paso del algoritmo consiste en extraer las oraciones completas del texto original que constituirán el resumen, en función de su distancia semántica respecto a los distintos clusters. Aunque en general el tamaño del resumen dependerá de las características del texto a resumir y del uso deseado del mismo, la extensión idónea podría oscilar entre un 20 y un 30 por ciento del texto de referencia. En esta etapa, se han investigado tres heurísticas para la selección de las oraciones:

- **Heurística 1:** El cluster de mayor tamaño (esto es, el que representa el *theme* principal en el documento), es el único que debería tenerse en cuenta para la generación del resumen. Por lo tanto, se seleccionan las *N* oraciones con las que presenta mayor similitud.
- **Heurística 2:** Todos los clusters contribuyen a la construcción del resumen con un número de oraciones (n_i) proporcional a su tamaño. Por lo tanto, para cada uno de los clusters, se seleccionan las n_i oraciones con las que presenta mayor similitud.
- **Heurística 3:** Para cada oración, se calcula el total de sus votaciones a todos los clusters, ponderadas por el tamaño de estos últimos, según la Ecuación 12. Se seleccionan las oraciones con mayor puntuación total.

$$score(O_j) = \sum_{C_i} \frac{similitud(C_i, O_j)}{|C_i|}$$

Ecuación 12

El problema de la ordenación de las oraciones en el resumen es trivial al tratarse de un resumen monodocumento, y se resolvería tomando las oraciones en el mismo orden en el que aparecen en el documento original.

La Tabla 9 recoge las oraciones seleccionadas por cada heurística, junto con su puntuación.

Heurística	Oraciones	Puntuación
Heurística 1	1	208.0
	22	27.0
	24	27.0
	43	26.0
	8	25.0
	20	25.0
	44	22.0
	57	22.0
	34	20.0
	48	20.0
	46	18.0
	3	17.0
Heurística 2	1	208.0
	21	31.0
	13	29.0
	22	27.0
	24	27.0
	25	27.0
	43	26.0
	8	25.0
	20	25.0
	29	22.0
	44	22.0
	57	22.0
Heurística 3	1	101.33
	21	16.47
	13	15.81
	35	15.53
	42	15.06
	43	13.2
	25	12.67
	20	12.5
	2	11.53
	10	10.89
	31	10.72
	8	10.28

Tabla 9 Oraciones seleccionadas por cada heurística y puntuación

Con respecto a la heurística 1, cabe decir que si bien la oración número 3 presenta la misma puntuación que las oraciones 18, 35 y 42, se ha decidido seleccionar esta por ser la que ocupa una posición más cercana al comienzo del documento, siguiendo el criterio posicional que afirma que las posiciones iniciales del documento generalmente contienen la información más significativa. Con respecto a la heurística 2, el número de oraciones que se extraen de cada cluster son 0 para el cluster 1, 4 para el cluster 2, 0 para el cluster 3 y 8 para el cluster 4.

A continuación, se muestra el resumen generado por cada una de las heurísticas.

Heurística 1

Nº Oración	Texto
1	In April 2000, the first results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) were published summarizing the comparison between two of the four drugs studied (chlorthalidone and doxazosin) as initial monotherapy for hypertension.
3	The diuretic had been the mainstay of several previous trials, particularly the Systolic Hypertension in the Elderly Program (SHEP) study.
8	While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke and a definite trend for heart failure to occur more often in the doxazosin group, than in the group taking chlorthalidone.
20	For the past 20 years, the proliferation of new antihypertensive drug classes has led to speculation (based on various kinds of evidence) that for equal reduction of blood pressure, some classes might be more beneficial than others with regard to effectiveness in preventing cardiovascular disease, and lesser adverse reactions of either minor or major significance.
22	On the other hand, the 'null' hypothesis for treatment of hypertension had been that the benefit of treatment is entirely related to reduction of arterial pressure and that the separate actions of the various drug classes are of no importance.
24	ALLHAT was conceived and designed to provide meaningful comparisons of three widely used newer drug classes to a 'classic' diuretic, as given in daily practice by primary care physicians for treatment of hypertension.
34	While a placebo arm was not included (and would have been unethical) there is every reason to accept the view that doxazosin did reduce arterial pressure (i.e. it remains an antihypertensive drug), but slightly less so than the

	diuretic.
43	Instead, clinical research implies that, like prazosin, doxazosin has no sustained hemodynamic benefit for congestive heart failure, due to development of tolerance (ie. the lack of a sustained hemodynamic effect in those with impaired left ventricular systolic function).
44	This has led to the suggestion that emergence of heart failure in the doxazosin cohort of ALLHAT was the expression of 'latent' heart failure at baseline, or soon thereafter, which either had been kept in check by previous treatment or was prevented from appearing by the diuretic or other therapy.
46	The case-fatality rates for heart failure, however, were also similar for the arms receiving doxazosin or chlorthalidone; once heart failure appeared, its consequences were disastrous.
48	Clinicians who treat hypertension should be aware that doxazosin, certainly as monotherapy, may be ineffective, perhaps little better than a placebo, for patients at higher risk for heart failure.
57	If even a small fraction of this large population is given a drug that fails to prevent heart failure, when effective medication might have been prescribed, our system of heathcare is just as deficient as if a more dramatic toxic adverse reaction (such as severe hepatic reactions to troglitizone) had occurred.

Heurística 2

Nº Oración	Texto
1	In April 2000, the first results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) were published summarizing the comparison between two of the four drugs studied (chlorthalidone and doxazosin) as initial monotherapy for hypertension.
8	While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke and a definite trend for heart failure to occur more often in the doxazosin group, than in the group taking chlorthalidone.

13 There was, however, no action taken by either the United States Food and Drug Administration (FDA) or the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, which authored the most recent advisory guideline for the National Heart Lung and Blood Institute.

20 For the past 20 years, the proliferation of new antihypertensive drug classes has led to speculation (based on various kinds of evidence) that for equal reduction of blood pressure, some classes might be more beneficial than others with regard to effectiveness in preventing cardiovascular disease, and lesser adverse reactions of either minor or major significance.

21 For example, the Heart Outcomes Prevention Evaluation (HOPE) trial reported that an angiotensin-converting enzyme inhibitor, ramipril, reduced fatal and nonfatal cardiovascular events in high-risk patients, irrespective of whether or not they were hypertensive.

22 On the other hand, the 'null' hypothesis for treatment of hypertension had been that the benefit of treatment is entirely related to reduction of arterial pressure and that the separate actions of the various drug classes are of no importance.

24 ALLHAT was conceived and designed to provide meaningful comparisons of three widely used newer drug classes to a 'classic' diuretic, as given in daily practice by primary care physicians for treatment of hypertension.

25 The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension.

29 Thereafter, they were treated similarly with addition of a beta blocker or other allowed agents (reserpine or clonidine for second step and hydralazine for third step) when needed.

43 Instead, clinical research implies that, like prazosin, doxazosin has no sustained hemodynamic benefit for congestive heart failure, due to development of tolerance (ie. the lack of a sustained hemodynamic effect in those with impaired left ventricular systolic function).

44 This has led to the suggestion that emergence of heart failure in the doxazosin cohort of ALLHAT was the expression of 'latent' heart failure at baseline, or soon

	thereafter, which either had been kept in check by previous treatment or was prevented from appearing by the diuretic or other therapy.
57	If even a small fraction of this large population is given a drug that fails to prevent heart failure, when effective medication might have been prescribed, our system of heathcare is just as deficient as if a more dramatic toxic adverse reaction (such as severe hepatic reactions to troglitizone) had occurred.

Heurística 3

Nº Oración	Texto
1	In April 2000, the first results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) were published summarizing the comparison between two of the four drugs studied (chlorthalidone and doxazosin) as initial monotherapy for hypertension.
2	This prospective, randomized trial was designed to compare a diuretic (chlorthalidone) with long-acting (once-a-day) drugs among different classes: angiotensin-converting enzyme inhibitor (lisinopril); calcium-channel blocker (amlodipine); and alpha blocker (doxazosin).
8	While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke and a definite trend for heart failure to occur more often in the doxazosin group, than in the group taking chlorthalidone.
10	ALLHAT continues with ongoing comparisons for amlodipine, lisinopril, and chlorthalidone.
13	There was, however, no action taken by either the United States Food and Drug Administration (FDA) or the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, which authored the most recent advisory guideline for the National Heart Lung and Blood Institute.
20	For the past 20 years, the proliferation of new antihypertensive drug classes has led to speculation (based on various kinds of evidence) that for equal reduction of blood pressure, some classes might be more beneficial than others with regard to effectiveness in preventing

cardiovascular disease, and lesser adverse reactions of either minor or major significance.

21 For example, the Heart Outcomes Prevention Evaluation (HOPE) trial reported that an angiotensin-converting enzyme inhibitor, ramipril, reduced fatal and nonfatal cardiovascular events in high-risk patients, irrespective of whether or not they were hypertensive.

25 The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension.

31 Despite a uniform goal of treatment for all enrolled, a small difference in systolic pressure was found between the two groups soon after entry and persisted until the doxazosin arm was discontinued.

35 The difference of 2–3 mmHg in systolic pressure between the two arms cannot account for doubling of the heart failure rate by doxazosin, compared to the diuretic as shown in Table 1.

42 The available studies support the concept that doxazosin or alpha blockers have a direct cardiotoxic effect.

43 Instead, clinical research implies that, like prazosin, doxazosin has no sustained hemodynamic benefit for congestive heart failure, due to development of tolerance (ie. the lack of a sustained hemodynamic effect in those with impaired left ventricular systolic function).

El análisis y la comparación de los resultados de las distintas heurísticas se realizarán en el capítulo 6, dedicado a la evaluación.

8. Resumen

En este capítulo se abordó la problemática de la generación automática de resúmenes monodocumento en el dominio de la biomedicina. El método propuesto consiste en crear una representación del documento y de sus oraciones en forma de grafo, en el que los vértices son los conceptos de UMLS

descubiertos en el texto y sus ancestros, mientras que las aristas se corresponden con distintos tipos de relaciones entre ellos.

Se realizó un análisis exhaustivo del significado de las relaciones entre conceptos y tipos semánticos en UMLS, y se decició utilizar las relaciones *is a* y *associated with* para enlazar los conceptos en el grafo. También se decició eliminar de la representación aquellos conceptos pertenecientes a tipos semánticos muy generales.

Para configurar correctamente el algoritmo de clustering, se llevó a cabo una extensa experimentación, de la que se obtuvieron los valores adecuados de los distintos parámetros que intervienen en el algoritmo. El número de *hub vertices* se estableció a un 20% del total de conceptos del grafo del documento, mientras que las ponderaciones de los votos a los *hub vertices* y los *non hub vertices* para la asignación de las oraciones a los cluster se fijaron a 1.0 y 5.0 respectivamente.

Finalmente, se propusieron tres heurísticas para la selección de las oraciones relevantes, y se generó el resumen correspondiente a cada una de ellas.

Capítulo 6

Evaluación

Para poder determinar la adecuación y la eficacia del método propuesto, resulta imprescindible realizar una evaluación de la calidad, en cuanto a contenido se refiere, de los resúmenes generados. Para ello, se ha realizado un análisis preliminar de los resultados obtenidos con las tres heurísticas descritas, aplicadas sobre el documento utilizado como ejemplo. Seguidamente, se ha llevado a cabo una evaluación formal utilizando la metodología ROUGE.

1. Configuración del Entorno de Experimentación

En este apartado se describen los pasos necesarios para configurar adecuadamente el entorno de ejecución, de acuerdo con la experimentación que se desee realizar. Para una mayor simplicidad, hemos dividido el proceso en tres etapas, según vaya encaminado a configurar la ontología, a pre-procesar el documento o a configurar el generador de resúmenes.

1.1. Configuración de UMLS

En primer lugar, debemos asegurarnos de que disponemos de la última versión de UMLS publicada por la NLM, que en el momento de realizar este trabajo se corresponde con la 2008AA. El *UMLS Knowledge Source* se puede obtener a

través de la página de descargas del *UMLS Knowledge Source Server*[24], o puede solicitarse en DVD a la NLM.

A continuación, debe ejecutarse la aplicación *MetamorphoSys,* que se distribuye junto con las fuentes de UMLS, para construir un subconjunto de la ontología con las terminologías que se deseen utilizar para la traducción de los documentos a conceptos. La ejecución de *MetamorphoSys* producirá también los *scripts* de carga para la base de datos *mysql* que almacenará la ontología (*populate_mysql_db.bat* y *populate_net_mysql_db.bat)*.

Seguidamente, se deben modificar estos *scripts* para indicar la ruta y el nombre de la base de datos, y el nombre de usuario y contraseña, y finalmente se ejecutarán.

1.2. Procesado preliminar de los documentos

Antes de ejecutar OBS, es necesario pre-procesar con GATE el documento cuyo resumen se desea realizar, para anotarlo sintácticamente y delimitar sus oraciones. Para más información sobre cómo realizar esta tarea, remitirse al apartado 1 del capítulo 4.

1.3. Configuración del sistema OBS

Finalmente, es necesario especificar los valores deseados para los siguientes parámetros de configuración, en el archivo *config.xml.*

- **DOCUMENT/PROCESSED:** Ruta del documento a resumir.
- **DOCUMENT/IGNORED_FIELDS:** Etiquetas XML que identifican a campos del documento que se desean ignorar a la hora de realizar el resumen (por ejemplo, tablas o gráficos).

[24] Descargas UMLS: http://kswebp1.nlm.nih.gov/uPortal/tag.a6eace6aebcd409.render.userLayoutRootNode.uP?uP_fname=umls-download

- **STOPLIST/FILE:** Ruta de la lista de parada que se utiliza para definir las palabras vacías.
- **ONTOLOGY/LEVEL_THRESHOLD:** Número de niveles superiores de la jerarquía de conceptos a ignorar.
- **UMLS/IGNORED_SEMANTIC_TYPES:** Tipos semánticos cuyos conceptos se ignoran en la construcción del grafo del documento.
- **UMLSLOCAL/DB_NAME:** Nombre de la base de datos que almacena UMLS.
- **UMLSLOCAL/DB_URL:** URL de la base de datos que almacena UMLS.
- **UMLSLOCAL/USER:** Nombre del usuario de la base de datos que almacena UMLS.
- **UMLSLOCAL/PW:** Password de la base de datos que almacena UMLS.
- **SNFC/PERCENTAGE_HUB_VERTICES:** Porcentaje de los vértices del grafo del documento que serán clasificados como *hub vertices.*
- **EXTRACTION/COMPRESSION_RATE:** Tasa de compresión del resumen a generar.
- **EXTRACTION/HUB_SCORE:** Número de votos asignados por una oración a los *hub vertices.*
- **EXTRACTION/NON_HUB_SCORE:** Número de votos asignados por una oración a los *non hub vertices.*

2. Evaluación Preliminar

En este apartado se analizan los extractos generados con las distintas heurísticas, para el documento utilizado como ejemplo en la explicación de la metodología.

Si bien los resultados obtenidos no son estadísticamente significativos, su análisis muestra algunos aspectos en los que el algoritmo se comporta satisfactoriamente. En primer lugar, llama la atención que las tres heurísticas presentan a la oración 1 como la más relevante, con una puntuación muy superior al resto de oraciones. Esto concuerda con el criterio posicional, adoptado en muchos trabajos, de seleccionar la primera oración del documento para el resumen, por ser la que generalmente contiene la información más significativa.

La oración número 58 presenta un claro ejemplo de oración que, estando posicionada al final del documento, recoge conclusiones sobre la exposición, y por lo tanto, tiene un alto contenido informativo. Esta oración es seleccionada tanto por la heurística 1 como por la heurística 2, aunque no por la heurística 3.

Por su parte, la oración número 20 ejemplifica la sobrevaloración de oraciones de gran longitud. Llama la atención el hecho de que es seleccionada por las tres heurísticas, aunque en principio parece tener una importancia relativa similar a otras oraciones que no se seleccionan. En esta oración, el mapeo al meta-tesauro de UMLS da como resultado un total de 23 conceptos, cuando el resto de oraciones presentan en torno a 10-12 conceptos. Por ello, y a pesar de que la mayoría de los conceptos que contiene no son centrales (es decir, no pertenecen a los *HVS*), las puntuaciones que asigna a los distintos clusters son altas, y por tanto, la probabilidad de ser seleccionada aumenta.

Otro aspecto a destacar es que las heurísticas 1 y 3 comparten un gran número de oraciones; en concreto, 8 de las 12 seleccionadas. Si analizamos las diferencias, observamos que la primera heurística selecciona oraciones de poca importancia relativa (3 y 26), aunque sí incluye otras muy relevantes como la oración 34. Un análisis detallado del resultado de esta primera heurística demuestra que con ella se ignoran algunos tópicos importantes del texto original, como los relacionados con los ensayos ALLHAT. Por su parte, la heurística 2 extrae oraciones cuyo contenido informativo se aleja un poco del tema central del documento, como las oraciones 13 y 21. Sin embargo, a falta de una evaluación formal, no es posible discernir de manera objetiva cuál de las dos heurísticas produce el mejor resumen.

La heurística 3, por su parte, se aleja bastante de la heurística 1, con la que comparte únicamente 4 oraciones, aunque no tanto de la heurística 2, con la que tiene 7 oraciones en común. Las oraciones en las que difiere de ambas heurísticas son las número 2, 10, 31, 35 y 42. Mientras que la oración 2 puede considerarse relevante para el resumen, la oración 10 no aporta ninguna información adicional, sino que está subsumida en la oración 2. Las oraciones 31 y 35, de nuevo no son suficientemente significativas como para formar parte del resumen, mientras que la número 42 sí contiene información relevante y, además, condensa la información en muy pocas palabras. No obstante, esta tercera estrategia no cubre todos los *themes* del documento, ya que no contempla ningún tipo de información a cerca de los resultados de la experimentación ALLHAT.

Algunas oraciones como la número 43, que además es seleccionada por todas las heurísticas, ponen de manifiesto los problemas de inconsistencia típicos de los métodos de extracción. Nótese que no tiene sentido incluir esta oración en el resumen si no se incluye también la anterior, ya que completa el contenido de esta. Para solucionar, al menos parcialmente, este tipo de inconsistencias, podría implementarse una estrategia basada en detectar palabras o grupos de palabras que actúan como conectores de oraciones (en este caso, *instead*), y no seleccionar las oraciones que las contengan a menos que también se seleccione la oración inmediatamente anterior.

Dado que los artículos del corpus utilizado se presentan acompañados del resumen elaborado por el propio autor, resulta interesante realizar una comparación entre éste y los resultados obtenidos por las distintas heurísticas. A pesar de que las longitudes de ambos resúmenes varían significativamente (de las 3 oraciones del resumen del autor a las 12 oraciones del resumen automático), se observa que las oraciones 1, 34 y 57 cubren la totalidad de la información presente en el *abstract* del autor, por lo que en base a este criterio, la estrategia número 1 es la que mejor cubre el contenido del documento, ya que incluye las tres oraciones; seguida por la estrategia 2, que incluye las oraciones 1 y 57; y finalmente, la estrategia 3, que sólo incluye la oración 1.

3. Evaluación Formal

El presente apartado muestra los resultados de la evaluación realizada para medir la calidad del método de generación de resúmenes desarrollado en esta tesis, en comparación con otros métodos propuestos en la literatura. Como ya se ha comentado en el apartado 2 del segundo capítulo, la evaluación de los sistemas automáticos de generación de resúmenes es una tarea compleja y muy controvertida, para la que no existe común acuerdo sobre las medidas y los procedimientos que se deben utilizar, a la hora de determinar objetivamente la calidad de los resúmenes. No obstante, y puesto que esta evaluación es necesaria si se desea comparar los distintos sistemas, existe un conjunto de métricas más o menos aceptadas por la comunidad investigadora. Si nos referimos a la evaluación intrínseca de estos sistemas, ROUGE es sin duda una de las metodologías más utilizadas (ver apartado 7 del capítulo 2), y es por ello que se ha elegido para la evaluación de OBS. En este tipo de evaluación, el procedimiento habitual consiste en comparar los resúmenes generados automáticamente con otros resúmenes "ideales" redactados por humanos (*jueces)*, a los que se les pide que extraigan las *n* oraciones que consideren más significativas. Por lo general, y para evitar discrepancias entre los jueces, se suelen incluir en el resumen ideal las oraciones elegidas por la mayoría.

En nuestra experimentación se ha contado con la colaboración de dos personas que practican la medicina, y se han elaborado dos extractos independientes de cada artículo, que posteriormente se han combinado en un único extracto incluyendo las oraciones seleccionadas por ambos y eligiendo por consenso las restantes. El resultado servirá como resumen "modelo" en relación al cual realizar la evaluación.

Para cada documento se han generado diez extractos aleatoriamente, cuya evaluación media definirá un límite inferior por debajo del cuál el resultado de cualquier sistema no sería aceptable, pues no aporta ninguna mejora respecto a una selección aleatoria de las oraciones. Por último, la comparación se realizará también con extractos generados seleccionando las *n* primeras oraciones de cada documento, práctica muy habitual que se conoce como *lead-based method* (Radev et al., 2002).

Jing (1998) demostró que, dependiendo de la tasa de compresión utilizada, la evaluación de un mismo sistema puede variar significativamente. Por ello, se realizaran experimentos para resúmenes de diferentes tamaños, y se calculará una evaluación media. Finalmente, para cada documento se evaluarán los resúmenes generados por cada una de las tres heurísticas propuestas.

El procedimiento de evaluación se llevará a cabo de la siguiente manera. En primer lugar, se calcularán las medidas ROUGE-1, ROUGE-2, ROUGE- L y ROUGE-W de los resúmenes generados con OBS, para cada una de las tres heurísticas y para los cuatro documentos del corpus de evaluación. En una primera fase, los resúmenes elaborados por los expertos se tomarán como resúmenes "ideales" con respecto a los que comparar para calcular las métricas ROUGE. En una segunda fase, la comparación se realizará con respecto a los *abstracts* que acompañan a los artículos originales. De este modo, tendremos una medida de la comparación con respecto a un extracto y una medida de la comparación con respecto a un resumen generado por abstracción. En segundo lugar, se calculará la medida ROUGE-1 para los resúmenes aleatorios y posicionales, tomando en este caso como resúmenes "ideales" los elaborados por los expertos. En todos estos experimentos, la tasa de compresión utilizada para generar los resúmenes será del 20%. Por último, se calculará ROUGE-1 para los resúmenes generados por OBS con la heurística 1 y con diferentes ratios de compresión (20%, 30% y 50% del tamaño del documento original), comparando una vez más con los resúmenes de los jueces.

Los documentos utilizados para la evaluación, en total cuatro, junto con sus resúmenes automáticos y manuales, se adjuntan en el Anexo II de esta memoria.

La Tabla 12 muestra el resultado de evaluar los resúmenes generados por nuestro sistema con respecto a los resúmenes elaborados por los jueces. Los valores promedio para la métrica ROUGE-1 son superiores a 0.8, lo que supone una evaluación muy superior a la obtenida por los resúmenes aleatorios (Tabla 14) y posicionales (Tabla 10), cuyos valores, como cabía esperar, se encuentran en torno a 0.4-0.5. Se observa que no existen diferencias significativas entre las tres heurísticas, si bien en promedio, la heurística 1 es la que mejor se comporta, seguida de la heurística 2 y, finalmente, la heurística 3.

Al realizar la evaluación con respecto al *abstract,* no obstante, los resultados empeoran sensiblemente, tal y como se puede observar en la Tabla 13. Obviamente, esto es debido a que, en lugar de realizar la comparación con resúmenes generados mediante la extracción de oraciones en el documento original, se está comparando respecto a resúmenes generados mediante abstracción y reescritura. Por este motivo, los resultados de las métricas ROUGE en el primer grupo de experimentos (evaluación con respecto a los extractos elaborados por los expertos) son muy elevados si se comparan con los obtenidos, por ejemplo, por los sistemas participantes en las conferencias *DUC,* que suelen situarse en torno a 0.4 – 0.5. En estas conferencias, la evaluación se realiza con respecto a un abstracto, por lo que los resultados sí serían equiparables a nuestras evaluaciones respecto del *abstract.* Por lo tanto, si consideramos que el *abstract* del propio autor del documento es un buen resumen, podemos afirmar que nuestro algoritmo presenta resultados a la altura de los mejores sistemas de generación de resúmenes que participan en estas conferencias. No obstante, es importante tener en cuenta que tan sólo se está evaluando la calidad del resumen en cuanto a cobertura de contenidos se refiere. Habría que completar esta evaluación con otras encaminadas a determinar la calidad gramatical y otros aspectos como la redundancia o la claridad referencial.

Si estudiamos de forma aislada la Tabla 10, correspondiente a la evaluación de los resúmenes posicionales, observamos que los resultados son mejores cuando la comparación se realiza con respecto a los *abstracts* que cuando se hace con respecto a los resúmenes de los jueces. Esto se debe, fundamentalmente, a que el resumen posicional contiene siempre la primera oración del documento, que generalmente condensa gran parte de lo contenido en el *abstract,* aunque sólo cubre una pequeña parte del contenido del resumen del experto.

En cuanto a la tasa de compresión se refiere, de nuevo y en contra de lo esperado, las diferencias encontradas son poco importantes, aunque la evaluación mejora ligeramente conforme aumenta el tamaño del resumen generado. Los resultados de la heurística 1 para distintos factores de compresión pueden verse en la Tabla 11.

Un ejemplo del buen funcionamiento del algoritmo lo encontramos en el documento número 2. En este caso, su longitud es de 11 oraciones, por lo que utilizando una tasa de compresión del 20% tan sólo se deben seleccionar dos de ellas. Las tres heurísticas realizan la misma selección, que además coincide con la de los jueces.

	Heurística 1	
Doc.	**Jueces**	**Abstract**
1	0.40183	0.55263
2	0.45833	0.25424
3	0.43284	0.47594
4	0.48894	0.52000
Promedio	0,44549	0,45070

Tabla 10 Resultados de la evaluación de los resúmenes posicionales (ROUGE-1, t=0,2)

	Heurística 1			
Doc.	**20%**	**30%**	**50%**	**Promedio**
1	1.00000	0.90840	0.88854	0,93231
2	1.00000	1.00000	0.93583	0,97861
3	0.88060	0.68108	0.80442	0,78870
4	0.56000	0.87925	1.00000	0,81308
Promedio	0,86015	0,86718	0,90720	0,87818

Tabla 11 Resultados de la evaluación para distintas tasas de compresión (ROUGE-1)

	HEURÍSTICA 1				HEURÍSTICA 2				HEURÍSTICA 3			
Doc.	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**
1	1.00000	0.99771	0.32063	1.00000	0.85845	0.78490	0.27017	0.84932	0.60731	0.41876	0.17837	0.58219
2	1.00000	1.00000	0.43497	1.00000	1.00000	1.00000	0.43497	1.00000	1.00000	1.00000	0.43497	1.00000
3	0.88060	0.83459	0.35347	0.87313	0.88060	0.82707	0.35347	0.87313	0.88060	0.83459	0.35347	0.87313
4	0.79853	0.70690	0.21687	0.79115	0.81572	0.72167	0.24245	0.80590	0.84521	0.77094	0.27353	0.84029
Promedio	**0,91978**	**0,88480**	**0,33149**	**0,91607**	**0,88869**	**0,83341**	**0,32527**	**0,88209**	**0,83328**	**0,75607**	**0,31009**	**0,82390**

Tabla 12 Resultados de la evaluación frente a los resúmenes de los jueces con las distintas heurísticas (t =0,20)

	HEURÍSTICA 1				HEURÍSTICA 2				HEURÍSTICA 3			
Doc.	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**	**ROUGE-L**
1	0.67544	0.24779	0.17888	0.58772	0.69298	0.24779	0.17709	0.59649	0.65789	0.20354	0.15615	0.51754
2	0.33898	0.31250	0.10933	0.33898	0.33898	0.31250	0.10933	0.33898	0.33898	0.31250	0.10933	0.33898
3	0.43316	0.23118	0.13054	0.40642	0.43316	0.23118	0.13054	0.40642	0.43316	0.23118	0.13054	0.40642
4	0.56000	0.18593	0.16152	0.51500	0.61000	0.23116	0.17432	0.57000	0.53500	0.16583	0.15333	0.49500
Promedio	**0,50190**	**0,24435**	**0,14507**	**0,46203**	**0,51878**	**0,25566**	**0,14782**	**0,47797**	**0,49126**	**0,22826**	**0,13734**	**0,43949**

Tabla 13 Resultados de la evaluación frente al *abstract* con las distintas heurísticas (t =0,20)

	Documento 1		Documento 2		Documento 3		Documento 4		
No. Resumen	Jueces	Abstract	Jueces	Abstract	Jueces	Abstract	Jueces	Abstract	Promedio
1	0.35388	0.50877	0.00000	0.19774	0.36567	0.28342	0.48894	0.63000	0,3535525
2	0.41324	0.61404	0.09722	0.20904	0.29104	0.36364	0.50369	0.53000	0,37773875
3	0.33562	0.57895	1.00000	0.33898	0.29851	0.39572	0.53808	0.51000	0,4994825
4	0.69635	0.63158	0.56944	0.25424	0.47015	0.37433	0.50123	0.58500	0,51029
5	0.39726	0.53509	0.45833	0.25424	0.41045	0.09189	0.40295	0.60000	0,39377625
6	0.42694	0.60526	0.30556	0.20339	0.44030	0.41176	0.56020	0.23618	0,39869875
7	0.49772	0.69298	0.19444	0.25424	0.29851	0.36898	0.54300	0.65500	0,43810875
8	0.46804	0.60526	0.52778	0.31638	0.51493	0.36898	0.51597	0.68000	0,4996675
9	0.40183	0.61404	0.09722	0.20904	0.35821	0.25668	0.51351	0.63500	0,38569125
10	0.51142	0.63158	0.04167	0.16949	0.43284	0.32620	0.45455	0.64000	0,40096875
Media	0,45023	0,601755	0,329166	0,240678	0,388061	0,32416	0,502212	0,570118	0,4257975

Tabla 14 Resultados de la evaluación frente a los jueces de los resúmenes generados aleatoriamente para la heurística 1 (ROUGE-1, t=0,20)

Capítulo 8

Conclusiones Y Trabajos Futuros

1. Principales Aportaciones

En este trabajo se ha presentado un método para la generación automática de resúmenes de textos de biomedicina, basado en la representación del documento como grafo extendido de conceptos y relaciones de UMLS, y en el cálculo de la relevancia de las oraciones a extraer en relación al prestigio o *salience* de los conceptos en el grafo del documento. De este modo, se construye una representación más rica en conocimiento que la que se tendría utilizando un modelo del espacio vectorial, y se consiguen solventar los problemas identificados en el capítulo introductorio.

Se ha presentado *OBS,* un sistema desarrollado para implementar el método propuesto, y se ha utilizado para generar resúmenes automáticos sobre algunos de los artículos científicos del corpus de *BioMed Central.* Así mismo, se han evaluado diferentes heurísticas para la extracción de las oraciones, primero mediante un análisis informal de los resúmenes generados y después, calculando distintas métricas ROUGE para cada uno de ellos. En ambos casos se concluye que, si bien las diferencias no son muy acusadas, la primera de las heurísticas se comporta mejor que las dos restantes, seguida por la segunda heurística y, finalmente, la tercera heurística.

En el capítulo 2, se ha realizado una extensa revisión del estado del arte en generación automática de resúmenes, tanto monodocumento como

multidocumento, gracias a la cual ha sido posible comprender cuáles son las principales limitaciones y problemas por resolver a día de hoy y hacia dónde han de dirigirse los esfuerzos de las nuevas investigaciones. La principal conclusión extraída es que la generación de resúmenes de calidad se perfila aún como uno de los grandes retos de la investigación en procesamiento de lenguaje natural. Es una tarea compleja, en la que confluyen otras tareas típicas del tratamiento automático de textos, como la recuperación de información, la indexación, la categorización, el agrupamiento, la extracción de información o la detección de temas; y que por lo tanto, puede y debe nutrirse de los métodos y técnicas utilizados en cada una de ellas.

Por otra parte, y a pesar de que la evaluación del método propuesto ha mostrado resultados satisfactorios en relación a la calidad de los resúmenes generados, también es cierto que hereda los principales inconvenientes de las aproximaciones basadas en extracción de oraciones. Como ya se ha comentado, uno de los principales problemas al que nos enfrentamos es el de la inconsistencia del resumen resultante. Una forma de paliar este problema, que además sería completamente compatible con nuestro enfoque, es utilizar el párrafo en lugar de la frase como unidad de extracción (Salton et al., 1994), con la esperanza de que al proporcionar un contexto más amplio se mejoren los problemas de legibilidad. En cuanto a la resolución de referencias anafóricas se refiere, algunas investigaciones nos muestran posibles soluciones. Se podría optar por no incluir las frases que contengan estas referencias (Brandow et al., 1995), incluir la frase anterior a la seleccionada (Namba y Okumura, 2000) o las frases que resuelven la anáfora, aunque no sean la inmediatamente anterior (Paice, 1990). El principal inconveniente de estas soluciones es que al seleccionar las oraciones anteriores, debemos dejar de añadir otras frases que pueden ser más importantes, para respetar el ratio de compresión deseado. Por último, podríamos intentar resolver inconsistencias detectando las expresiones que conectan oraciones (“sin embargo”, “por tanto”, etc.) y eliminándolas si aparecen al principio de una frase y la frase anterior no forma parte del resumen (Mateo et al., 2003).

Otro inconveniente de este tipo de técnicas y, en particular del método diseñado, es que el tamaño del resumen generado, aún utilizando una misma tasa de compresión, puede variar mucho dependiendo de la longitud de las

oraciones seleccionadas. Esto se debe a que el tamaño del resumen se calcula aplicando el ratio de compresión sobre el número de oraciones en lugar de sobre el número de palabras. Por lo tanto, una solución directa a este problema sería calcular el tamaño en términos de palabras e ir seleccionando oraciones hasta completar el número de palabras deseado.

Finalmente, la investigación desarrollada ha dado lugar a una serie de publicaciones que se detallan en el anexo IV de este documento

2. Trabajo futuro

El trabajo realizado hasta el momento, presentado como tesis para optar al título de *Máster en Investigación Informática*, será el punto de partida para la Tesis Doctoral que se pretende desarrollar en los próximos años. Como fruto del esfuerzo realizado, se vislumbran algunas posibles mejoras, así como nuevas líneas de investigación en las que centrar nuestro trabajo futuro.

En primer lugar, el método presentado extrae oraciones completas, lo que implica que las de mayor longitud, al contener un mayor número de conceptos, tienen mayor posibilidad de ser seleccionadas. Una solución a considerar sería dividir la puntuación de las oraciones entre el número de conceptos que las componen.

Por otra parte, tal y como se ha podido comprobar empíricamente, uno de los factores que más influye en el tamaño y composición de los grupos de conceptos generados por el algoritmo *SFGC* es el grado de conectividad del grafo del documento. Esta conectividad depende directamente de las relaciones ontológicas consideradas entre los conceptos. En este trabajo se han analizado las relaciones *is a, other related* y *associated with* de UMLS, pero sería conveniente estudiar nuevas relaciones como las de coocurrencia entre conceptos.

Sería recomendable experimentar con nuevos algoritmos de clustering. Por ejemplo, se podría tener en cuenta la distancia (en términos de número de saltos) entre los nodos del grafo a la hora de hacer el agrupamiento. También

se podría aplicar un algoritmo basado en lo que se conoce como *betweenness centrality*, una medida que describe el grado en que un nodo se encuentra entre otros dos, y que en nuestro caso, describiría el grado en que un concepto sirve para enlazar otros dos conceptos.

Desde el punto de vista lingüístico, habría que abordar la resolución de referencias anafóricas y pronominales. En ocasiones, se observa cómo en una oración un concepto relevante es referenciado mediante un pronombre. Sin embargo, al no resolverse estas referencias en la implementación actual, el concepto no se reconoce como tal y la puntuación asignada a la oración se ve penalizada. En este sentido, se estudiará la posibilidad de aplicar las distintas soluciones comentadas en el apartado anterior.

Quedarían por resolver problemas como el tratamiento de tablas y gráficos a la hora de construir el resumen, ya que estos elementos se presentan muy frecuentemente en los textos que nos ocupan y, generalmente contienen información importante y que debería ser incluida en el resumen.

Por otra parte, la extracción de las oraciones relevantes es sólo la primera etapa de un proceso en el que se pretende elaborar un resumen (*abstract*) donde no se mencione explícitamente el contenido del documento original, sino que sea reescrito íntegramente utilizando técnicas de generación de lenguaje. Por ello, el trabajo más inmediato se centrará en investigar cómo abordar las fases de análisis de las oraciones seleccionadas y de generación. En la fase de análisis, para extraer la estructura de la oración, se estudiará la posibilidad de aplicar un análisis de dependencias, utilizando para ello herramientas como *Minipar*[25] o *MaltParser*[26]. También se podría aplicar un análisis de constituyentes, utilizando algunos de los parsers disponibles como *Charniak*[27]. En la etapa de generación, se traducirá la estructura de representación intermedia obtenida como resultado del análisis aplicando técnicas de generación de lenguaje natural, y se estudiará la posibilidad de reutilizar recursos existentes.

[25] Parser *MiniPar:* http://www.cs.ualberta.ca/~lindek/minipar.htm

[26] Parser *MaltParser: http://w3.msi.vxu.se/~nivre/research/MaltParser.html*

[27] Parser Charniak: ftp://ftp.cs.brown.edu/pub/nlparser

Otra línea de trabajo futura sería la extensión del método para poder realizar resúmenes a partir de múltiples documentos sobre un mismo tema. En un principio, las modificaciones no supondrán cambios sustanciales, si bien la tarea es mucho más compleja que la generación a partir de un solo documento y plantea retos adicionales. En primer lugar, debería estudiarse cuidadosamente la selección de documentos que comparten una relación semántica y que contribuirán a la redacción de un mismo resumen, para evitar mezclar en un mismo resumen informaciones inconexas. En segundo lugar, el hecho de contener los documentos información común puede dar lugar a resúmenes redundantes, por lo que la detección y eliminación de redundancias es uno de los principales problemas que se tendrían que resolver. Tercero, es igual de crítico reconocer las diferencias importantes entre documentos, que pueden deberse a la consideración de información adicional o al planteamiento de la misma bajo distintos puntos de vista. Por último, se debe asegurar la coherencia del resumen, tomando en cuenta que las diferentes porciones de información provienen de diferentes fuentes.

Asimismo, se está estudiando una posible modificación del algoritmo para generar resúmenes adaptados al usuario. Para ello, se necesitaría disponer de un modelo del usuario, entendido como una representación de sus intereses y preferencias (Díaz y Gervás, 2004). En el dominio particular que nos ocupa, esta posibilidad resulta muy interesante, ya que permitiría que, para un mismo artículo, se generasen resúmenes distintos dependiendo de si el destinatario es un profesional interesado en conocer posibles tratamientos para una enfermedad, o un estudiante que necesita información sobre los síntomas que presenta y los grupos de riesgo.

Finalmente, resultaría muy interesante aplicar el método propuesto a otros dominios distintos de la biomedicina, como podría ser la economía o las finanzas, y comparar la calidad de los resúmenes obtenidos en ambos dominios. Para ello, tan sólo sería necesario disponer de una ontología adecuada que formalice el conocimiento del dominio en cuestión, ya que la validez de nuestro método es independiente del dominio considerado. Por otro lado, el propio diseño de la aplicación facilita su posterior reutilización en otros dominios.

Anexo I: OBS

El propósito de esta sección es presentar *OBS (An Ontology-Based BioMedical Summarizer),* un sistema diseñado con el objetivo de resolver la tarea de generar resúmenes automáticos de artículos sobre biomedicina, y que implementa el método expuesto en el apartado anterior.

Para ello, en primer lugar se describe la arquitectura general del sistema. Seguidamente, se explica cada uno de los módulos que lo constituyen, siguiendo el orden lógico secuencial en el que intervienen en la generación del resumen.

Nótese que, puesto que el método ya ha sido expuesto en detalle en el capítulo 6, en este apartado nos centraremos en los aspectos relacionados con la arquitectura y la implementación de la aplicación.

1. Arquitectura del Sistema

Para la descripción de la arquitectura general del sistema se ha optado por elaborar un diagrama de paquetes UML. Se ha elegido UML por ser el lenguaje de modelado de sistemas software más conocido y utilizado en la actualidad,

por ofrecer un conjunto de diagramas estándar para la especificación de la vista estática del sistema.

En UML, un diagrama de paquetes muestra cómo un sistema se divide en agrupaciones lógicas, y cuáles son las dependencias existentes entre esas agrupaciones, suministrando una descomposición de su jerarquía lógica. Además, los paquetes están normalmente organizados para maximizar la coherencia interna dentro de cada paquete y minimizar el acoplamiento externo entre los paquetes. La Figura 25 muestra el diagrama de paquetes del sistema OBS.

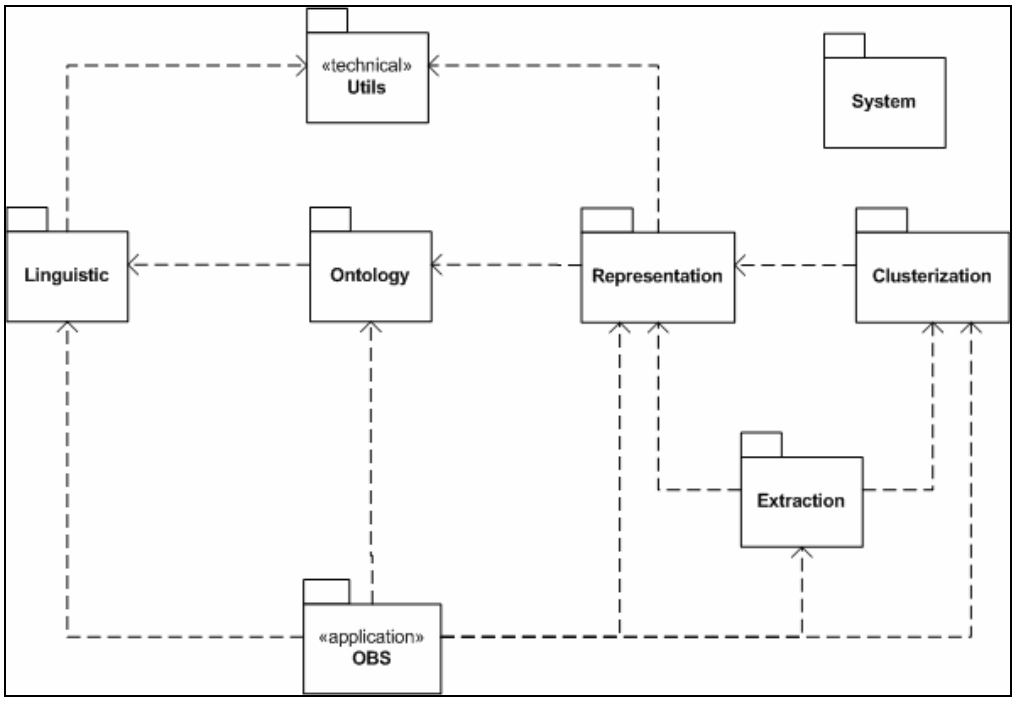

Figura 25 Diagrama de paquetes de OBS

En los siguientes apartados se presentan sucesivamente los diagramas elaborados para cada uno de los subsistemas anteriores, mostrando los paquetes y clases contenidos en cada uno de ellos. Por su simplicidad, para los paquetes *OBS* y *Utils* no se han realizado diagramas de clases, por lo que su funcionalidad se explica someramente a continuación.

- **Paquete *Utils:*** Contiene una colección de clases que son utilizadas por el resto de paquetes para implementar la funcionalidad principal

del sistema, como por ejemplo, estructuras de datos y algoritmos de ordenación.

- **Paquete *OBS:*** Contiene las clases ejecutables del sistema.

2. *Módulo de Configuración e Inicialización (Paquete System)*

Se trata del paquete que engloba las clases encargadas de la configuración del sistema y de la inicialización de los distintos componentes que lo conforman. También se responsabiliza del manejo de excepciones. La Figura 26 muestra el diagrama de clases para este paquete.

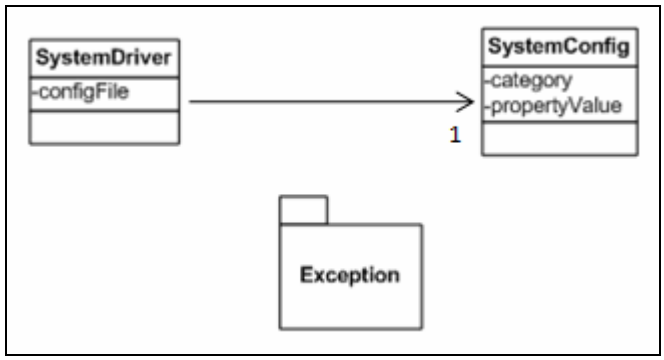

Figura 26 Diagrama de clases del paquete *System*

Clase SystemConfig

Se encarga de leer el fichero *config.xml,* en el que se especifican los valores de los distintos parámetros de configuración del algoritmo, organizados por categorías. En concreto, las principales propiedades de configuración, junto con los valores utilizados para la ejecución del ejemplo desarrollado, se muestran en la Tabla 15.

Categoría	Propiedad	Valor
DOCUMENT	ORIGINAL	cvm-2-6-254.xml
	PROCESSED	cvm-2-6-254-proc.xml
	FRECUENCY_THRESHOLD	2.0

	IGNORED_FIELDS	XML_TABLE;XML_ST
DOCUMENT_TAG	XML_TITLE XML_ABSTRACT XML_BODY …	title abs bdy …
STOPLIST	FILE	StopWords.txt
ONTOLOGY	ONTOLOGY_NAME	UMLS
UMLS	IGNORED_SEMANTIC_TYPES	Quantitative Concept,Temporal Concept,Idea or Concept,Intellectual Product,Mental Process,Spatial Concept,Language
UMLSLOCAL	DB_NAME	Umls
	DB_URL	//localhost/umls
	USER	umlsuser
	PW	umlsuser
UMLSKS	AUTHENTICATION_HOST	http://umlsks.nlm.nih.gov
	AUTHENTICATION_URI	/authorization/services/ AuthorizationPort
	CERTIFICATE	C:\\Users\\Laura\\workspace \\IL\\OntoBioSum\\ laurapm_cer.txt
	USER	laurapm
	PW	lpm.umls
	TICKET	http://umlsks.nlm.nih.gov
	KS_HOST	http://umlsks.nlm.nih.gov
	KS_URI	/UMLSKS/services/UMLSKSService
SNFC	NUM_PERCENTAGE_VERTICES	20
EXTRACTION	COMPRESSION_RATE	0.20
	HUB_SCORE	1.0
	NO_HUB_SCORE	0.5

Tabla 15 Parámetros de configuración de OBS

Clase SystemDriver

Esta clase sirve de fachada al sistema completo, con el objetivo de hacerlo transparente a la interfaz. Se encarga de inicializar los componentes de OBS que así lo requieren (Metamap, la ontología y la lista de parada).

Paquete Exception

Agrupa las clases encargadas del manejo de las condiciones de error que puedan surgir durante la ejecución de OBS.

3. Módulo de Procesamiento Lingüístico (Paquete Linguistic)

Este paquete contiene las clases que representan al documento y las distintas unidades lingüísticas que lo componen. La Figura 27 muestra el diagrama de clases correspondiente.

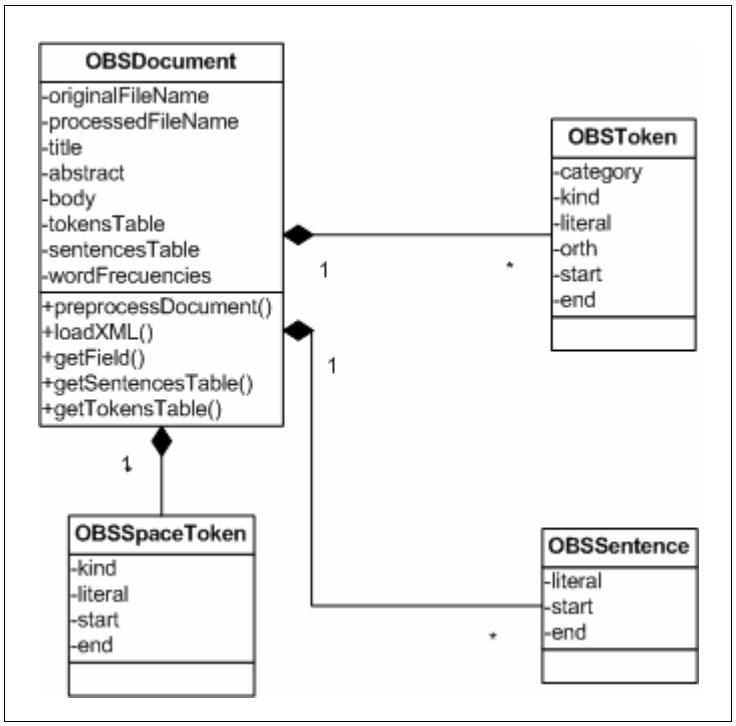

Figura 27 Diagrama de clases del paquete *Linguistic*

Clase OBSDocument

Clase que representa al documento cuyo resumen se desea generar. Contiene el documento original, tal y como se presenta en el corpus de BioMed Central, y el documento xml resultante de aplicar los distintos módulos de GATE en la etapa de pre-procesado. Proporciona las clases necesarias para extraer la información relevante del documento (el cuerpo del artículo, el título y el abstract), y para extraer las oraciones que previamente han sido delimitadas por el *sentece splitter* de GATE.

Clase OBSSentence

Clase que representa a una oración del documento.

Clase OBSToken

Clase que representa a un token del texto, incluyendo signos de puntuación y excluyendo espacios y caracteres de control. Incluye información sobre su categoría sintáctica, sobre el tipo de token (palabra o signo de puntuación) y sobre si está escrito en mayúsculas o minúsculas.

Clase OBSSpaceToken

Clase que representa a un token de tipo espacio o carácter de control.

4. Módulo de Acceso y Traducción a la Ontología (Paquete Ontology)

Se trata del paquete que contiene las clases encargadas de la inicialización y el acceso a la ontología. La Figura 28 muestra el diagrama de clases correspondiente.

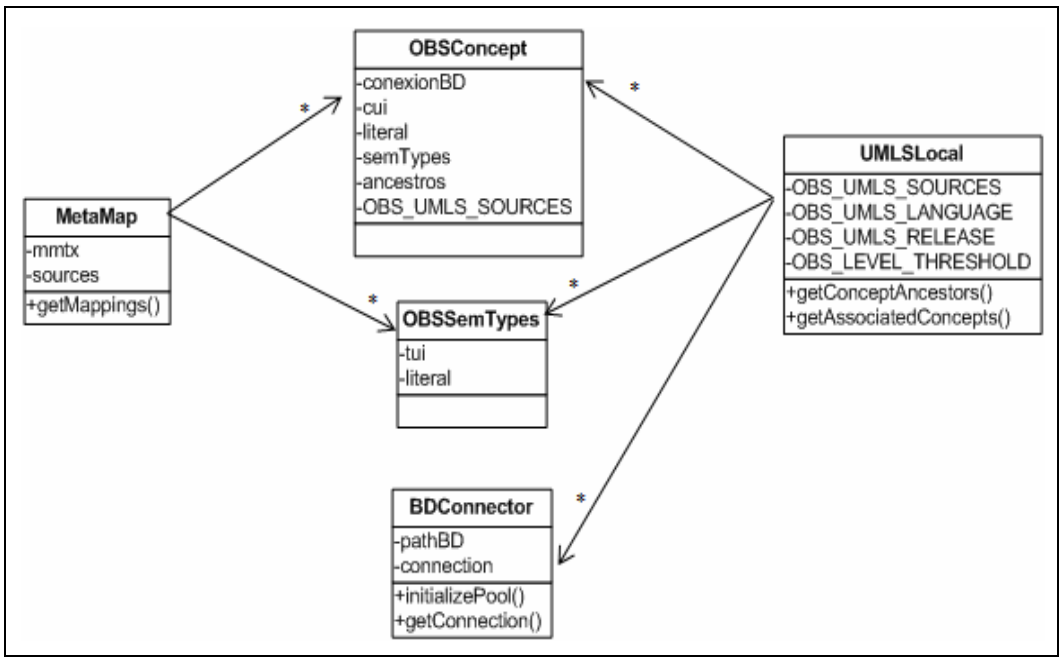

Figura 28 Diagrama de clases del paquete *Ontology*

Clase MetaMap

Clase que encapsula el acceso a la aplicación *MMTx*, a través del API correspondiente, creando una instancia de tipo *MMTxAPI*. Proporciona métodos para realizar el mapeo de una oración a los conceptos en UMLS que contiene.

Clase UMLSLocal

Clase que encapsula el acceso local al subconjunto de UMLS generado con MetaMap. Crea una instancia de la base de datos que almacena la ontología y una conexión a la misma. Proporciona métodos para extraer los hiperónimos de un concepto y los conceptos con los que se relaciona a través de las relaciones *associated with* entre sus tipos semánticos.

Clase OBSConcept

Clase que representa a un concepto en UMLS. Contiene los atributos que identifican de manera unívoca a cada concepto en la ontología.

Clase OBSSemType

Clase que representa a un tipo semántico en UMLS. Contiene los atributos que identifican de manera unívoca a cada tipo en la ontología.

Clase BDConnector

Clase que crea una conexión a la base de datos *MySql* donde se almacena la ontología generada con *MetamorphoSys*, e inicializa el pool de conexiones.

5. Módulo de Representación Gráfica (Paquete Representation)

Este paquete engloba las clases encargadas de la representación de las oraciones y del documento como grafos de conceptos. La Figura 29 muestra el diagrama de clases del paquete *Representation.*

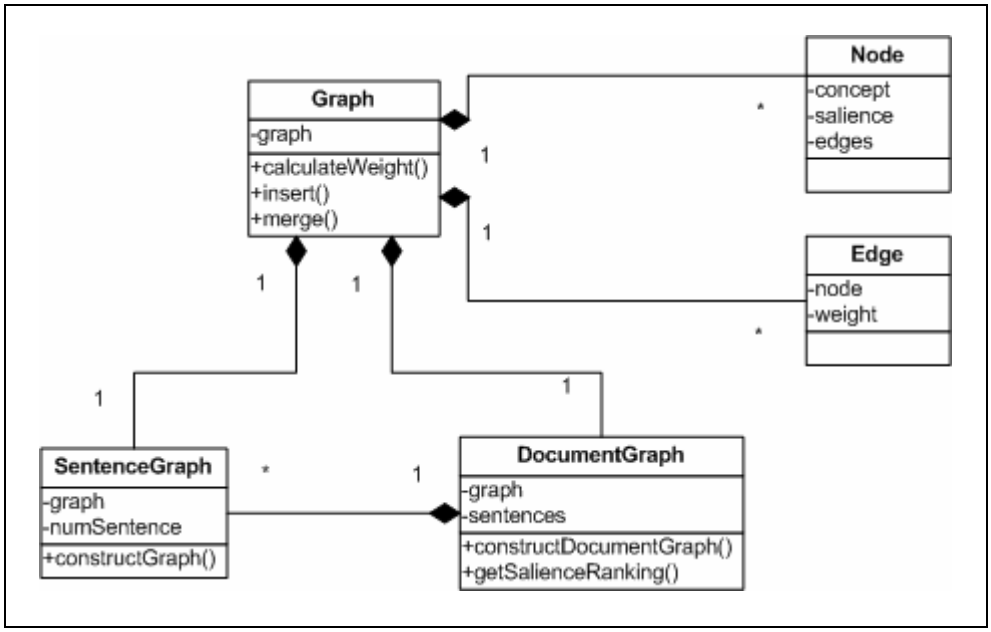

Figura 29 Diagrama de clases del paquete *Representation*

Clase Graph

Clase que representa a un objeto de tipo grafo en el que los nodos representan conceptos de UMLS, y las aristas, las relaciones establecidas entre ellos. Asociado a cada nodo se tiene un *salience* o relevancia, que indica el grado de concentración de aristas. Por su parte, cada arista tiene un peso asociado.

Clase SentenceGraph

Clase que representa el grafo de una oración. Contiene, por lo tanto, un objeto de tipo *Graph,* y una referencia a la oración correspondiente del documento.

Clase DocumentGraph

Clase que representa el grafo de un documento. Contiene, por lo tanto, un objeto de tipo *Graph,* y un vector de grafos de las oraciones a partir de los cuales se construye.

6. Módulo de Agrupamiento de Conceptos (Paquete Clusterization)

Este paquete engloba las clases encargadas de implementar el algoritmo de agrupamiento de conceptos. La Figura 30 muestra el diagrama de clases del paquete *Clusterization.*

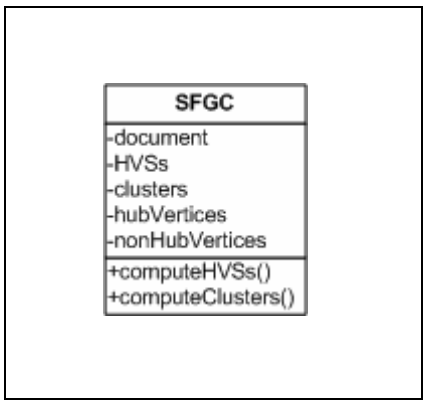

Figura 30 Diagrama de clases del paquete *Clusterization*

Clase SFGC

Clase que encapsula el algoritmo de agrupamiento *scale-free graph clustering.* Contiene una referencia al objeto de tipo *documentGraph* que representa al grafo del documento. La invocación del método *computeClusters* permite generar el conjunto de clusters de conceptos, de acuerdo con el método descrito en la sección 5.

7. Módulo de Generación del Resumen (Paquete Extraction)

Este paquete contiene las clases encargadas de extraer las oraciones que constituirán el resumen final. La Figura 31 muestra el diagrama de clases del paquete *Extraction.*

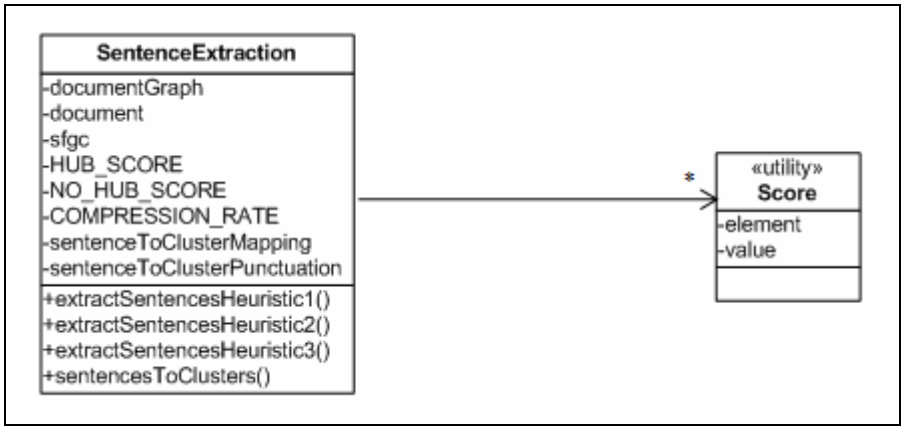

Figura 31 Diagrama de clases del paquete *Extraction*

Clase SFGC

Esta clase presenta métodos para la asignación de las oraciones del documento a los distintos clusters de conceptos, así como para extraer las oraciones que formarán el resumen, de acuerdo con las tres heurísticas definidas.

Anexo II: Documentos

Documento I: Comments on ALLHAT and Doxazosin

- **División en oraciones del documento original**

Nº Oración	Texto
1	In April 2000, the first results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) were published summarizing the comparison between two of the four drugs studied (chlorthalidone and doxazosin) as initial monotherapy for hypertension.
2	This prospective, randomized trial was designed to compare a diuretic (chlorthalidone) with long-acting (once-a-day) drugs among different classes: angiotensin-converting enzyme inhibitor (lisinopril); calcium-channel blocker (amlodipine); and alpha blocker (doxazosin).
3	The diuretic had been the mainstay of several previous trials, particularly the Systolic Hypertension in the Elderly Program (SHEP) study.
4	During the first three years of the trial, the Data Safety and Monitoring Committee became aware of different event rates between two groups and, after considerable deliberation, made the decision to discontinue the arm assigned to doxazosin for two stated reasons.
5	One reason was an extremely low likelihood that doxazosin would prove superior to chlorthalidone when the study would be completed as planned.

6 The second reason was a pattern of increased morbid events in comparing doxazosin to the diuretic, which was highly significant on statistical analysis.

7 This pattern is shown in Table 1.

8 While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke and a definite trend for heart failure to occur more often in the doxazosin group, than in the group taking chlorthalidone.

9 For heart failure, the curves for event rates diverged quite early in the trial, within the first year, but continued to separate over the three-year period of analysis.

10 ALLHAT continues with ongoing comparisons for amlodipine, lisinopril, and chlorthalidone.

11 The results of ALLHAT regarding doxazosin were first made public by a presentation at the meeting of the American College of Cardiology, March 2000, and the subsequent publication.

12 Several news agencies published reports of the study, and comments appeared in a few medical journals.

13 There was, however, no action taken by either the United States Food and Drug Administration (FDA) or the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, which authored the most recent advisory guideline for the National Heart Lung and Blood Institute.

14 In May 2001, however, after initiation of a class action lawsuit and a Citizens Petition, the FDA held a hearing on the issue.

15 Sidney Wolfe and the author gave presentations recommending that all physicians receive a warning with an interpretation of the ALLHAT results, and that the labeling and indications for doxazosin be changed by Pfizer, the company which developed and markets doxazosin as Cardura ®.

16 Pfizer representatives argued that doxazosin is a safe and effective antihypertensive drug, based on their own accumulated studies, with no need for any additional warning or change in labeling.

17 The remainder of this article will summarize the basis for a warning and address the arguments of those who conclude that nothing further needs to be done.

18 There is no longer any doubt that treatment of hypertension is beneficial and prevents stroke, coronary heart disease, and congestive heart failure.

19 This is particularly so for high-risk populations, such as those over 50 years of age, with other risk factors and target organ damage.

20 For the past 20 years, the proliferation of new antihypertensive drug classes has led to speculation (based on various kinds of evidence) that for equal reduction of blood pressure, some classes might be more beneficial than others with regard to effectiveness in preventing cardiovascular disease, and lesser adverse reactions of either minor or major significance.

21 For example, the Heart Outcomes Prevention Evaluation (HOPE) trial reported that an angiotensin-converting enzyme inhibitor, ramipril, reduced fatal and nonfatal cardiovascular events in high-risk patients, irrespective of whether or not they were hypertensive.

22 On the other hand, the 'null' hypothesis for treatment of hypertension had been that the benefit of treatment is entirely related to reduction of arterial pressure and that the separate actions of the various drug classes are of no importance.

23 While published guidelines suggest uniform treatment for hypertension, practice by individual physicians varies considerably.

24 ALLHAT was conceived and designed to provide meaningful comparisons of three widely used newer drug classes to a 'classic' diuretic, as given in daily practice by primary care physicians for treatment of hypertension

25 The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension.

26 ALLHAT is a large multicenter study (44,000 enrolled), with randomization, a double blind design, nearly complete follow-up of enrolled subjects, and other features that qualify it as compelling evidence for clinical decisions.

27 At entry, the baseline characteristics of the doxazosin and chlorthalidone groups in ALLHAT were similar.

28 Ninety percent of both groups had been previously treated

and were then changed to their assigned medication.

29 Thereafter, they were treated similarly with addition of a beta blocker or other allowed agents (reserpine or clonidine for second step and hydralazine for third step) when needed.

30 In both groups, blood pressure fell significantly from entry levels and remained lower.

31 Despite a uniform goal of treatment for all enrolled, a small difference in systolic pressure was found between the two groups soon after entry and persisted until the doxazosin arm was discontinued.

32 Overall, the doxazosin group had a 2–3 mmHg higher systolic pressure during the trial.

33 Diastolic pressures were equal in the two groups.

34 While a placebo arm was not included (and would have been unethical) there is every reason to accept the view that doxazosin did reduce arterial pressure (i.e. it remains an antihypertensive drug), but slightly less so than the diuretic.

35 The difference of 2–3 mmHg in systolic pressure between the two arms cannot account for doubling of the heart failure rate by doxazosin, compared to the diuretic as shown in Table 1.

36 Furthermore, clinicians treating patients in the two groups were not aware of any difference in response.

37 Adherence to assigned treatment was 10% higher for chlorthalidone, compared to doxazosin over 4 years of observation, but discontinuation of medication was similar for the two groups (20% for chlorthalidione and 19% for doxazosin).

38 Is doxazosin (or any other alpha receptor blocker) a dangerous drug?

39 Did the results of ALLHAT reveal drug toxicity or the 'lesser charge' of ineffectiveness?

40 The former status clearly requires a warning and, in some circumstances, may lead to withdrawal of approval.

41 The latter requires widespread information to advise all clinicians who treat hypertension of better and worse strategies.

42 The available studies support the concept that doxazosin or alpha blockers have a direct cardiotoxic effect.

43 Instead, clinical research implies that, like prazosin, doxazosin has no sustained hemodynamic benefit for congestive heart failure, due to development of tolerance (ie. the lack of a sustained hemodynamic effect in those with impaired left ventricular systolic function).

44 This has led to the suggestion that emergence of heart failure in the doxazosin cohort of ALLHAT was the expression of 'latent' heart failure at baseline, or soon thereafter, which either had been kept in check by previous treatment or was prevented from appearing by the diuretic or other therapy.

45 The basis for a diagnosis of heart failure was uniform between the two groups.

46 The case-fatality rates for heart failure, however, were also similar for the arms receiving doxazosin or chlorthalidone; once heart failure appeared, its consequences were disastrous.

47 In other words, heart failure as an endpoint in ALLHAT was an important clinical event.

48 Clinicians who treat hypertension should be aware that doxazosin, certainly as monotherapy, may be ineffective, perhaps little better than a placebo, for patients at higher risk for heart failure.

49 Further more, the greater adherence rate for chlorthalidone compared to doxazosin should not be overlooked.

50 In a double blind trial, such a pattern implies that the alpha blocker was, for unknown reasons, less acceptable that the diuretic alternative.

51 This is new and important information from a very large trial, that is not likely to be replicated.

52 The FDA panel heard the opposing viewpoints on a warning for doxazosin as a result of the ALLHAT trial.

53 They concluded that something should be done, but could not take any further action and, instead, asked for additional analyses.

54 This may be all that current FDA policies allow.

55 There is, however, a compelling mandate emerging from those

	who see the need for greater safety in the provision of healthcare.
56	Beneficial therapy should be achieved for as many hypertensive patients as possible.
57	If even a small fraction of this large population is given a drug that fails to prevent heart failure, when effective medication might have been prescribed, our system of heathcare is just as deficient as if a more dramatic toxic adverse reaction (such as severe hepatic reactions to troglitizone) had occurred.
58	Results of large and well-conducted clinical trials, such as ALLHAT, must be the basis for optimal healthcare policy.

- **Abstract**

This commentary has two purposes: to summarize the rationale, design and initial results of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) trial; and to provide a history of the response to ALLHAT that led to a civil action and a Citizens Petition that was the basis for a public hearing by the US Food and Drug Administration, in May 2001. The author concludes that the results of ALLHAT should be widely disseminated. All clinicians must be warned that initial therapy with doxazosin (and possibly other alpha1 blockers) is definitely inferior to low dose diuretic treatment for patients at high risk for cardiovascular disease, such as those enrolled in ALLHAT.

- **Resúmenes utililizados para la evaluación**

Resumen	Oraciones
Heurística 1 (20%)	1, 3, 8, 20, 22, 24, 34, 43, 44, 46, 48, 57
Heurística 2 (20%)	1, 8, 13, 20, 21, 22, 24, 25, 31, 35, 42, 43
Heurística 3 (20%)	1, 2, 8, 10, 13, 20, 21, 25, 31, 35, 42, 43
Juez (20%)	1, 8, 20, 21, 22, 24, 34, 43, 44, 46, 48, 57
Heurística 1 (30%)	1, 2, 4, 5, 8, 13, 20, 21, 22, 24, 25, 29, 34, 35, 43, 46, 57

Heurística 2 (30%)	1, 2, 4, 5, 8, 13, 20, 21, 22, 24, 25, 29, 34, 35, 43, 46, 55
Heurística 3 (30%)	1, 2, 4, 5, 8, 13, 20, 21, 22, 24, 25, 29, 34, 35, 43, 46, 57
Juez (30%)	1, 5, 8, 14, 20, 21, 22, 24, 31, 34, 35, 43, 44, 46, 48, 57
Heurística 1 (50%)	1, 2, 3, 4, 5, 8, 11, 13, 14, 16, 18, 20, 21, 22, 23, 24, 25, 29, 31, 34, 35, 37, 42, 43, 44, 46, 48, 50, 57
Heurística 2 (50%)	1, 2, 3, 4, 5, 8, 11, 13, 14, 16, 18, 20, 21, 22, 23, 24, 25, 29, 31, 34, 35, 37, 42, 43, 44, 46, 48, 55, 57
Heurística 3 (50%)	1, 2, 3, 4, 5, 8, 11, 13, 14, 16, 18, 20, 21, 22, 23, 24, 25, 29, 31, 34, 35, 37, 42, 43, 44, 46, 48, 50, 57
Juez (50%)	1, 2, 4, 5, 6, 8, 14, 15, 16, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 31, 34, 35, 37, 43, 44, 46, 48, 57, 58
Posicional (20%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Posicional (30%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17
Posicional (50%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29

Documento II: Coronary artery-pulmonary artery fistula: case report

- **División en oraciones del documento original**

Nº Oración	Texto
1	Coronary artery fistulas are rare congenital or acquired coronary artery anomalies that can originate from any of the three major coronary arteries and drain in all the cardiac chambers and great vessels.
2	An 11-year-old boy was referred for evaluation of an exertional dyspnoea.
3	He reported recent history of few episodes of shortness of breath associated with moderate entity physical activity.
4	At physical examination a mild continuous murmur could be heard mainly at the level of the second intercostal space of the left parasternal area.
5	A transthoracic echocardiogram (Figure 1A) showed a continuous flow at color Doppler analysis in the high parasternal short axis view, originating from a small entry site on the wall of the main pulmonary artery (arrow).
6	A selective left coronary angiography (Figure 1B) revealed a fistula connecting the proximal portion of the left anterior descending coronary artery (solid white arrow) with the main pulmonary artery (blank white arrow); black arrows indicated drainage into the pulmonary circulation.
7	Most fistulas originate from the right coronary artery or the left anterior descending and commonly drain in low-pressures structures including right-sided chambers, pulmonary artery, superior cava vein and coronary sinus.
8	A combination like the one described in the present case is unusual since fistulas originate from the left coronary artery in about 35% of cases and drainage into the pulmonary artery occurs in only 17%.
9	The author(s) declare that they have no competing interests.

10	IQ collected the data relative to the Case Report; VZ composed the draft of the manuscript; SM is head of the Echo Lab and conceived the Case report and participated in the coordination, data analysis and elaboration and drafting of the manuscript.
11	All authors read and approved the final manuscript.

➢ Abstract

Background: Coronary artery fistulas are rare congenital or acquired coronary artery anomalies that can originate from any of the three major coronary arteries and drain in all the cardiac chambers and great vessels.
Case presentation: An 11-year-old boy was referred for evaluation of an exertional dyspnoea. He reported recent history of few episodes of shortness of breath associated with moderate entity physical activity. At physical examination a mild continuous murmur could be heard mainly at the level of the second intercostal space of the left parasternal area. A transthoracic echocardiogram showed a continuous flow at color Doppler analysis in the high parasternal short axis view,originating from a small entry site on the wall of the main pulmonary artery. A selective left coronary angiography revealed a fistula connecting the proximal portion of the left anterior descending coronary artery with the main pulmonary artery.
Conclusion: A combination like the one described in the present case is unusual since fistulas originate from the left coronary artery in about 35% of cases and drainage into the pulmonary artery occurs in only 17%.

➢ Resúmenes utililizados para la evaluación

Resumen	Oraciones
Heurística 1 (20%)	1, 6
Heurística 2 (20%)	1, 6
Heurística 3 (20%)	1, 6
Juez (20%)	1, 6

Heurística 1 (30%)	1, 6, 7
Heurística 2 (30%)	1, 6, 7
Heurística 3 (30%)	1, 6, 7
Juez (30%)	1, 6, 7
Heurística 1 (50%)	1, 5, 6, 7, 8, 10
Heurística 2 (50%)	1, 5, 6, 7, 8, 10
Heurística 3 (50%)	1, 5, 6, 7, 8, 10
Juez (50%)	1, 2, 5, 6, 7, 8
Posicional (20%)	1, 3
Posicional (30%)	1, 2, 3
Posicional (50%)	1, 2, 3, 4, 5, 6

Documento III: An unusual cause of chest pain: case report

- **División en oraciones del documento original**

Nº Oración	Texto
1	Sarcomas form a heterogenous group of relatively uncommon malignant tumours which are derived from connective tissue components.
2	In total they comprise approximately 1% of all new cancers diagnosed per year in the United Kingdom (UK).
3	As subset of this, the 'Unclassified' Sarcoma forms approximately 4% of the total.
4	They often present with as relatively slow growing, asymptomatic masses and as such may often be misdiagnosed as in this case.
5	A 52 year old Caucasian gentleman presented to his general practitioner (G.P) with moderate left sided chest pain.
6	This pain had developed suddenly that morning and was localised to the chest.
7	Due to a strong family history of ischemic heart disease and the nature of the pain, he was then referred to the local casualty department.
8	All investigations including chest radiographs, electrocardiograms, cardiac enzymes, echocardiograpy and subsequent coronary angiography proved negative for coronary artery disease.
9	He was advised by his hospital practitioner that he may have had a chest infection, and as part of a general lifestyle change was advised to try to take more exercise and lose weight.
10	Over the next four months Mr DP succeeded in losing 20 kg, but still complained of worstening left sided chest pain.
11	This pain was not affected by exercise, movement or heavy lifting.

12 He then re-presented to his G.P with a hard mass in the left upper chest over the site of his longstanding chest pain.

13 He was referred for a surgical consultation and underwent excision of a lesion arising from the left pectoralis major muscle 1 week later.

14 Histology from this lesion, revealed features suggestive of Hodgkin's Lymphoma, but due to the unusual appearance of the lesion the biopsies were sent for a second opinion.

15 A second opinion noted mixed chronic inflammatory cells with occasional germinal centres which might suggest an inflammatory pseudotumour.

16 The ususual features of very large bizarre, and atypical polygonal cells however led to a final diagnosis of an unclassified sarcoma, morphologically low grade, with lymphoma like features (Fig 1).

17 Personal communication from Professor Fletcher recommended annual clinical follow-up with annual MRI scanning.

18 Of note, Mr DP's chest pain completely disappeared following excision of his lesion.

19 Annual scans were clear of recurrent local disease but Mr DP represented in November 2005 with a nodule adjacent to the scar from previous surgery.

20 Excision of this lesion revealed recurrent tumour of the same morphology and he underwent a radical compartectomy at a third operation in December 2005.

21 Sarcomas form a heterogenous group of relatively uncommon malignant tumours which are derived from connective tissue components.

22 In total they comprise approximately 1% of all new cancers diagnosed per year in the United Kingdom (UK).

23 As subset of this, the 'Unclassified' Sarcoma forms approximately 4% of the total and may often be misdiagnosed.

24 The majority of patients present with asymptomatic, deep seated, slow growing, ill defined masses.

25 In this case, the chest pain may have been due to perineural invasion and expansion of the tumour with compression of the surrounding muscle.

26 It became palpable as a result of intentional weight loss

	and increasing size of the tumour.
27	Tumours < 5 cm (as in this case) can be effectively treated with wide local excision ensuring margins of at least 2 cm, with 5 year disease free survival rates of nearly 78% in low grade lesions.
28	Larger more aggressive tumours or recurrent tumours may benefit from more radical surgery or radiotherapy.

➢ **Abstract**

Background: Sarcomas form a heterogenous group of relatively uncommon malignant tumours which are derived from connective tissue components. In total they comprise approximately 1% of all new cancers diagnosed per year in the United Kingdom (UK). As subset of this, the 'Unclassified' Sarcoma forms approximately 4% of the total [1]. They often present with as relatively slow growing, asymptomatic masses and as such may often be misdiagnosed as in this case.

Case presentation: A 52 year old man presented to his general practitioner (GP) with left sided chest pain. A strong family history of ischaemic heart disease prompted hospital referral and further investigations which all proved negative for coronary artery disease. Following weight loss and ongoing chest pain, he represented to his GP with a hard mass arising from the left pectoralis major muscle at the site of the previous pain. Surgical excision followed by later compartectomy revealed an unclassified low grade Sarcoma with lymphoma like features.

Conclusion: In this case, chest pain masquerading as ischaemia, may have been caused by perineural infiltration or compression of adjacent muscle bulk by tumour, with eventual surgical resection providing a good long term prognosis.

➢ **Resúmenes utililizados para la evaluación**

Resumen	Oraciones
Heurística 1 (20%)	1, 7, 8, 16, 19, 25
Heurística 2 (20%)	1, 7, 8, 16, 19, 25
Heurística 3 (20%)	1, 7, 8, 16, 19, 25

Juez (20%)	1, 5, 8, 16, 19, 25
Heurística 1 (30%)	1, 7, 8, 13, 15, 16, 19, 25
Heurística 2 (30%)	1, 5, 7, 8, 13, 16, 19, 25
Heurística 3 (30%)	1, 7, 8, 11, 13, 16, 19, 25
Juez (30%)	1, 5, 8, 16, 19, 25, 27, 28
Heurística 1 (50%)	1, 5, 7, 8, 9, 11, 12, 13, 15, 16, 19, 25, 26, 27
Heurística 2 (50%)	1, 5, 7, 8, 9, 11, 13, 15, 16, 19, 25, 26, 27, 28
Heurística 3 (50%)	1, 5, 7, 8, 9, 11, 12, 13, 15, 16, 19, 25, 26, 27
Juez (50%)	1, 4, 5, 8, 12, 14, 15, 16, 19, 20, 25, 26, 27, 28
Posicional (20%)	1, 2, 3, 4, 5, 6
Posicional (30%)	1, 2, 3, 4, 5, 6, 7, 8
Posicional (50%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Documento IV: Tongue lesions in psoriasis: a controlled study

- **División en oraciones del documento original**

Nº Oración	Texto
1	The occurrence of psoriatic lesions on oral mucous membranes was a subject of controversy.
2	Some investigators stated that they do not occur; others, have claimed that they are uncommon.
3	Still others say that they occur only in generalized pustular psoriasis (GPP).
4	Nowadays, there is sufficient evidence that a subset of patients have oral lesions in association with skin disease.
5	Oppenheim, in 1903, was the first to substantiate oral psoriasis with biopsy.
6	Since then, various lesions have been described, including grey, yellowish, white or translucent plaques or annular forms, diffuse areas of erythema, geographic tongue and fissured tongue.
7	In all the cases reported in the literature, a positive biopsy showing a psoriasiform pattern has been the crucial component of the diagnosis.
8	Thus hyperkeratosis, parakeratosis, and an inflammatory infiltrate consisting of lymphocytes, polymorphonuclear leukocytes and histiocytes have been noted as well as Munro's microabscesses and spongiform pustules of Kogoj.
9	In addition, many investigators believe that the presence of cutaneous lesions with a course parallel to that of oral lesions is necessary for establishing the diagnosis of oral psoriasis.
10	However, it is impossible to perform an oral biopsy in psoriatic cases in everyday clinical practice.
11	On the other hand, some of the lesions seen more frequently in psoriatic patients are not specific histologically.

12 In fact, similar changes are seen in otherwise healthy people (although with a lower frequency) leading to an underestimation of the value of these findings in psoriatic patients.

13 In order to substantiate further the relationship between these oral disorders and psoriasis, we compared 200 patients with psoriasis to a matched control group.

14 Two hundred psoriatic patients (70 women and 130 men) attending the dermatology clinics of Razi Hospital, a major referral center in Tehran, from September 2000 till February 2001, were enrolled in this study using simple nonrandom (sequential) sampling.

15 The diagnosis was made mainly on clinical data.

16 The control group included 200 healthy subjects among the visitors of Surgery wards in a general hospital, matched one by one for age and sex.

17 The skin and oral mucosa were examined in the two groups and, in addition to demographic and clinical data, PASI score 21 was recorded in plaque-type psoriasis.

18 The data were analyzed by Epi-Info (version 6) software, and frequency, mean, standard deviation, OR and p-value were calculated.

19 The mean age of the patient group was 33.8+/- 18.2 years (4–79 years).

20 The mean age of onset of disease was 26+/-17.7 years (0–74 years), 23 +/-18.8 years in women and 27.6 +/- 17.0 years in men.

21 Age and sex were matched between patients and control subjects.

22 Family history of psoriasis was positive in 34 patients.

23 Different clinical types of psoriasis were as follows: Chronic plaque-type psoriasis (n = 140); generalized pustular psoriasis (n = 10); flexural psoriasis (n = 10); erythrodermic psoriasis (n = 9); localized pustular psoriasis (n = 3); guttate psoriasis (n = 9); palmoplantar psoriasis (n = 15); scalp (n = 95); nail alone (n = 3).

24 Oral findings were detected in 87 (43.5%) and 39 (19.5%) cases in the psoriatic and control groups, respectively.

25 They are presented in table 1 .

26 FT was seen more frequently in psoriatic patients (66 patients, 33%) than the control group (19, 9.5%) (OR: 4.69; 95% CI: 2.61–8.52) (p-value < 0.0001).

27 BMG, too, was significantly more frequent in psoriatic patients (28 cases, 14%) than the control group (12, 6%) (OR: 2.55; 95% CI: 1.20–5.50) (p-value < 0.012).

28 BMG was seen in 18.2% of patients with FT, and 42.9% of patients with BMG suffered from FT.

29 In other words, in 12 patients (6%) FT and BMG coexisted.

30 In the control group, FT and BMG coexisted in 2 cases (1%).

31 One hundred eighty-four patients (92%) suffered from erythemato-squamous lesions and 13 cases (6.5%) from pustular lesions.

32 The frequency of FT in the erythemato-squamous and pustular groups was 30.4% (56 cases) and 53.8% (7 cases), respectively.

33 On the other hand, the frequency of BMG in the erythemato-squamous and pustular groups was 14.1% (26 cases) and 15.4% (2 cases), respectively.

34 The severity of chronic plaque-type psoriasis cases assessed by PASI score was as follows: mild, 53 cases (37.9%); moderate, 60 cases (42.9%); and severe 27 cases (19.3%).

35 The corresponding frequency of FT and BMG in the three severity groups is presented in table 2 .

36 The frequency of BMG increased with the severity of skin lesions (p-value < 0.001).

37 In general oral lesions in psoriasis can be divided into two major categories.

38 The first one includes authentic psoriatic lesions proved by biopsy and with a parallel clinical course with skin lesions.

39 It not known whether these lesions are truly rare, or they remain undetected, as mucosal biopsy is seldom done in known psoriatic cases.

40 The second group comprises the majority of oral findings in psoriasis and includes nonspecific lesions such as FT and psoriasiform lesions such as BMG.

41 These lesions are underestimated in the literature, but deserve more attention due to their high frequency.

42 We will discuss the main oral findings observed in our study as well as those reported in the literature.

43 Fissured tongue, also termed lingua fissurata, lingua plicata, scrotal tongue, and grooved tongue is recognized clinically by an antero-posteriorly oriented fissure, often with branch fissures extending laterally.

44 It believed by most authors to be an inherited trait.

45 The frequency of FT increases with age and has been associated with Down syndrome and the Melkerson-Rosenthal syndrome.

46 According to our study, FT was the most common oral finding in the psoriasis group: Nearly one-third of patients suffered from FT.

47 It was significantly more frequent in psoriasis patients than the control group (9.5%) (p-value < 0.0001).

48 The previously reported figures of the frequency of FT in the general population vary markedly in the literature depending on the study design and the target study.

49 Axell reported a figure of 6.5% 10 and Morris found FT in 20.3% of its target study.

50 Aboyans et al reported a frequency of 2.56% in Iran.

51 On the other hand, FT was reported in 6–16.7% of psoriatic patients in different studies.

52 BMG or geographic tongue presents clinically as one or more erythematous patches with a raised white or yellow serpiginous border.

53 Lesions may migrate across the tongue by healing on one edge while extending on another.

54 BMG has no known cause, but it has been associated with atopic conditions, diabetes mellitus, reactive bronchitis, anemia, stress 20 , hormonal disturbances, Down syndrome and lithium therapy.

55 Lesions identical to BMG have been described in patients with Reiter syndrome and psoriasis.

56 The association of both psoriasis and BMG with HLA-CW6 provides further evidence that the two disorders are

related.

57 In our study, BMG was significantly more frequent in psoriatic patients (14%) than the control group (6%) (p-value < 0.012).

58 According to the literature, the estimated frequency of BMG in the general population is from 1–5% 1 10 20 26 and varies from 1–10.3% in psoriatic patients.

59 Only Hietanen found a figure of 1% in psoriasis.

60 FT was more frequent in patients with pustular lesions compared with the erythemato-squamous types.

61 Contrary to previous studies, this finding was not seen for BMG, a disease generally considered accompanied with GPP.

62 This may be due to the low frequency of GPP in our study group.

63 On the other hand, the frequency of BMG increased with the severity of psoriasis in plaque-type disease, a finding not seen in Morris study 20 , perhaps due to different definition for the severity of the disease.

64 According to our study, the frequency of FT increase by increasing severity of psoriasis.

65 SAM was first described by Cooke in 1955 as an idiopathic inflammatory condition of the nonlingual oral mucosa.

66 It is also denoted using different terms: Geographic stomatitis, ectopic geographic tongue, erythema circinate migrans, and migratory stomatitis.

67 These lesions are similar in appearance to BMG, but occur on the oral mucosal surfaces as well as the dorsum of the tongue.

68 As seen in Van der Wal study, 14 we find SAM in psoriatic patients.

69 The reported frequency of this oral finding in psoriatic patients in the literature is between 0–19%.

70 Furthermore, this lesion seems to be very rare in the general population, too: Bouquot found no patients with SAM in 231616 white American dental patients.

71 Diffuse oral and tongue erythema was another positive finding in the psoriasis group with a frequency of 5.5%.

72	This lesion, too, was reported previously in the literature, although with a lower frequency (1%).
73	An association between FT and BMG is well established in the literature.
74	In our study, BMG was seen in 18.2% of patients with FT (results consistent with Pindorf).
75	Overall, although oral lesions might not be considered authentic oral psoriasis unless proven histologically and with a parallel clinical course, nonspecific tongue lesions are significantly more frequent in psoriatic cases.
76	Further studies are recommended to evaluate the clinical significance of these seemingly nonspecific lesions in a suspected psoriatic case.
77	Furthermore, more thorough studies are recommended regarding the relationship of oral psoriasis and disease severity in plaque-type psoriasis.

➢ Abstract

Background: Our objective was to study tongue lesions and their significance in psoriatic patients.
Methods: The oral mucosa was examined in 200 psoriatic patients presenting to Razi Hospital in Tehran, Iran, and 200 matched controls.
Results: Fissured tongue (FT) and benign migratory glossitis (BMG) were the two most frequent findings. FT was seen more frequently in psoriatic patients (n = 66, 33%) than the control group (n = 19, 9.5%) [odds ratio (OR): 4.69; 95% confidence interval (CI): 2.61–8.52] (p-value < 0.0001). BMG, too, was significantly more frequent in psoriatic patients (28 cases, 14%) than the control group (12 cases, 6%) (OR: 2.55; 95% CI: 1.20–5.50) (p-value < 0.012). In 11 patients (5.5%), FT and BMG coexisted. FT was more frequent in pustular psoriasis (7 cases, 53.8%) than erythemato-squamous types (56 cases, 30.4%). On the other hand, the frequency of BMG increased with the severity of psoriasis in plaque-type psoriasis assessed by psoriasis area and severity index (PASI) score. Conclusions: Nonspecific tongue lesions are frequently observed in psoriasis. Further studies are recommended to substantiate the clinical significance of these seemingly nonspecific findings in suspected psoriatic cases.

- **Resúmenes utililizados para la evaluación**

Resumen	Oraciones
Heurística 1 (20%)	1, 6, 11, 14, 17, 20, 23, 43, 51, 52, 56, 63, 68, 70, 71
Heurística 2 (20%)	1, 6, 8, 13, 16, 23, 43, 51, 52, 54, 57, 60, 63, 68, 70
Heurística 3 (20%)	1, 6, 8, 14, 17, 23, 43, 51, 52, 54, 63, 68, 70, 71, 77
Juez (20%)	1, 6, 13, 17, 20, 23, 43, 46, 51, 52, 54, 63, 68, 70, 71
Heurística 1 (30%)	1, 6, 11, 13, 14, 16, 17, 21, 23, 24, 43, 46, 51, 54, 57, 63, 65, 66, 68, 70, 71, 72, 77
Heurística 2 (30%)	1, 6, 8, 11, 13, 14, 16, 21, 23, 24, 40, 43, 51, 52, 54, 55, 56, 57, 63, 68, 70, 76, 77
Heurística 3 (30%)	1, 6, 8, 11, 13, 14, 16, 17, 21, 23, 43, 46, 51, 52, 54, 56, 57, 63, 65, 68, 70, 71, 77
Juez (30%)	1, 6, 11, 13, 14, 16, 17, 20, 23, 24, 26, 43, 46, 51, 52, 54, 57, 63, 68, 70, 71, 76, 77
Heurística 1 (50%)	1, 3, 5, 6, 7, 11, 13, 14, 16, 17, 20, 21, 22, 23, 33, 34, 39, 42, 45, 46, 48, 49, 51, 52, 54, 57, 61, 63, 64, 66, 67, 68, 69, 70, 71, 74, 75, 76, 77
Heurística 2 (50%)	1, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 21, 23, 24, 27, 33, 38, 42, 43, 46, 48, 49, 51, 52, 54, 56, 57, 63, 64, 65, 66, 68, 69, 70, 71, 74, 75, 76, 77
Heurística 3 (50%)	1, 3, 6, 7, 8, 11, 12, 13, 14, 16, 17, 21, 23, 24, 26, 27, 33, 34, 42, 43, 46, 47, 49, 51, 52, 54, 56, 57, 63, 64, 65, 66, 68, 69, 70, 71, 75, 76, 77
Juez (50%)	1, 6, 7, 8, 9, 11, 13, 14, 16, 17, 19, 20, 23, 24, 26, 27, 29, 32, 33, 37, 38, 40, 43, 46, 51, 52, 53, 54, 56, 57, 61, 63, 64, 68, 70, 71, 74, 75, 76, 77
Posicional (20%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
Posicional (30%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23
Posicional (50%)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

Anexo III: Tipos Semánticos en UMLS

ENTITY
Physical Object Organism Plant Alga Fungus Virus Rickettsia or Chlamydia Bacterium Archaeon Animal Invertebrate Vertebrate Amphibian Bird Fish Reptile Mammal Human Anatomical Structure Embryonic Structure Anatomical Abnormality Congenital Abnormality Acquired Abnormality Fully Formed Anatomical Structure Body Part, Organ, or Organ Component Tissue Cell Cell Component Gene or Genome Manufactured Object Medical Device

```
            Research Device
            Clinical Drug
        Substance
            Chemical
                Chemical Viewed Functionally
                    Pharmacologic Substance
                        Antibiotic
                    Biomedical or Dental Material
                    Biologically Active Substance
                        Neuroreactive Substance or Biogenic Amine
                        Hormone
                        Enzyme
                        Vitamin
                        Immunologic Factor
                        Receptor
                    Indecator, Reagent, or Diagnostic Acid
                    Hazardous or Poisonous Substance
                Chemical Viewed Structurally

    Organic Chemical
                        Nucleic Acid, Nucleoside, or Nucleotide
                        Organophosphorus Compound
                        Amino Acid, Peptide, or Protein
                        Carbohydrate
                        Lipid
                             Steroid
                             Eicosanoid
                    Inorganic Chemical
                    Element, Ion, or Isotope
            Body Substance
            Food
     Conceptual Entity
        Idea or Concept
            Temporal Concept
            Qualitative Concept
            Quantitative Concept
            Functional Concept
                Body System
            Spatial Concept
                Body Space or Junction
                Body Location or Region
                Molecular Sequence
                    Nucleotide Sequence
                    Amino Acid Sequence
                    Carbohydrate Sequence
                Geographic Area
        Finding
            Laboratory or Test Result
```

```
            Sign or Symptom
        Organism Attribute
            Clinical Attribute
        Intellectual Product
            Classification
            Regulation or Law
        Language
        Occupation or Discipline
            Biomedical Occupation or Discipline
        Organization
            Health Care Related Organization
            Professional Society
            Self-help or Relief Organization
        Group Attribute
        Group
            Professional or Occupational Group
            Population Group
            Family Group
            Age Group
            Patient or Disabled Group
```

EVENT

```
Activity
        Behavior
            Social Behavior
            Individual Behavior
        Daily or Recreational Activity
        Occupational Activity
            Health Care Activity
                Laboratory Procedure
                Diagnostic Procedure
                Therapeutic or Preventive Procedure
            Research Activity
                Molecular Biology Research Technique
            Governmental or Regulatory Activity
            Educational Activity
        Machine Activity
    Phenomenon or Process
        Human-caused Phenomenon or Process

Environmental Effect of Humans
        Natural Phenomenon or Process
            Biologic Function
                Physiologic Function
                    Organism Function
                        Mental Process
                    Organ or Tissue Function
                    Cell Function
```

Molecular Function
Pathologic Function
Disease or Syndrome
Mental or Behavioral Dysfunction
Neoplastic Process
Cell or Molecular Dysfunction
Experimental Model or Disease
Injury or Poisoning

Tabla 16 Tipos Semánticos en UMLS

Anexo III: Publicaciones

El trabajo realizado en este proyecto ha dado lugar a la aceptación de los siguientes artículos.

- Plaza, L., Díaz, A. y Gervás, P. (2008). Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina. En *Sociedad Española de Procesamiento de Lenguaje Natural.*

- Plaza, L., Díaz, A. y Gervás, P. (2008). Concept-graph based Biomedical Automatic Summarization using Ontologies. En *Coling 2008 Workshop TextGraphs-3: Graph-based Algorithms for Natural Language Processing.* Manchester, UK.

Bibliografía

[1] Alonso, L. (2005). Representing discourse for automatic text summarization via shallow NLP. *Tesis*. Universidad de Barcelona.

[2] Alterman, R. (1985). A Dictionary Based on Concept Coherence. En *Artificial Intelligence*, 25(2), pp. 153-186.

[3] Alterman, R. y Bookman, L. (1990). Some computational experiments in summarization. En *Discourse Processes*, núm. 13, pp. 143–174.

[4] Amini, M.-R. y Gallinari, P. (2002). The Use of Unlabeled Data to Improve Supervised Learning for Text Summaries. En *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 105-112. Finland.

[5] Ando, R., Boguraev, B., Byrd, R. y Neff, M. (2000). Multi-Document Summarization by Visualizing Topical Content. En *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA.

[6] Antequeira, L., Pardo, T.A.S., Nunes, M.G.V. y Oliveira, O.N. (2007). Some Issues on complex networks for autor characterization. En *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial,* 11(36), pp. 51-58.

[7] Aone, C., Okurowski, M. E., Gorlinsky, J. y Larsen, B. (1999). A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. En I.

Mani y M. T. Maybury, *Advances in Automatic Text Summarization*, pp. 71-80. The MIT Press.

[8] Baeza-Yates, R. y Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. En *ACM Press Books*, New York.

[9] Barabási A.L. y Albert, R. (1999). Emergence of scaling in random networks. En *Science.* 268, pp. 509-512.

[10] Barzilay, R., McKeown, K. R. y Elhadad, M. (1997). Using Lexical Chains for Text Summarization. En *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 11-121. Madrid, Spain.

[11] Barzilay, R. y Elhadad, M. (1999). Using Lexical Chains for Text Summarization. En I. Mani y M. T. Maybury, *Advances in Automatic Text Summarization*, pp. 111-121. The MIT Press.

[12] Baxendale, P. B. (1958). Man-Made Index for Technical Literature - an Experiment. En *IBM Journal of Research and Development*, 2(4), pp. 354-361.

[13] Brandow, R., Mitze, K. y Rau, L. F. (1995). Automatic Condensation of Electronic Publications by Sentence Selection. En *Information Processing and Management*, 31(5), pp. 675-685.

[14] Bretonnel K., Fox, L., Ogren, F.V. y Hunter, L. (2005). Corpus design for biomedical natural language processing. En *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pp. 38–45. Detroit, USA.

[15] Bretonnel, K., Ogren, P., Fox, L. y Hunter, L. (2005). Empirical data on corpus design and usage in biomedical natural language processing. En *AMIA Annu Symp Proc*, pp. 156–160.

[16] Buckley, C. y Cardie, C. (1997). Using EMPIRE and SMART for high-precision IR and summarization. En *Proceedings of the TIPSTER Text Phase III 12-MonthWorkshop*, pp. 107-121, San Diego, CA, USA.

[17] Carbonell, J., Geng, Y. y Goldstein, J. (1997). Automated Query-Relevant Summarization and Diversity-Based Reranking. En *Proceedings of the IJCAI-97 Workshop on AI in Digital Libraries*, pp. 12-19.

[18] Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y. y Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. En *IJCAI* (1), pp. 708-715.

[19] Carbonell, J. G. y Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. En *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335-336. Melbourne, Australia

[20] Carrero, F., Gómez, J.M., Buenaga, M., Mata, J. y Maña, M. (2007). Acceso a la Información Bilingüe Utilizando Ontologías Específicas del Dominio Biomédico. En *Sociedad Española para el Procesamiento del Lenguaje Natural.*

[21] Chandrasekaran, B., Josephson, J. R. y Benjamins V. R. (1999). What Are Ontologies, and Why Do We Need Them? En *IEEE Intelligent Systems*, 14(1), pp. 20-26.

[22] Chang, Y. K., Cirillo, C. y Razon, J. (1971). Evaluation of Feedback Retrieval Using Modified Freezing, Residual Collection, and Test and Control Groups. En G. Salton, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 355-370. Prentice-Hall, Inc.

[23] Collier, N., Hyun, S.P., Ogata, N., Tateisi, Y., Nobata, C., Sekimizu, T., Imai, H. y Tsujii, J. (1999). The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. En *EACL*.

[24] DeJong, G. F. (1982). An Overview of the FRUMP System. En W. G. Lehnert y M. H. Ringle, *Strategies for Natural Language Processing*, pp. 149-176.

[25] Díaz, A., Gervás, P. y García, A. (2005). System-oriented Evaluation for Multi-tier Personalization of Web Contents. En *Proceedings of the Workshop on Intelligent Information Processing*. Las Palmas, España.

[26] Díaz, A., Gervás, P. y García, A. (2005). Evaluation of a System for Personalized Summarization of Web Contents. En *Proceedings of UM2005 User Modeling: Proceedings of the Tenth International Conference*, LNAI, Edinburgh.

[27] Donaway, R. L., Drummey, K. W. y Mather, L. A. (2000). A Comparison of Rankings Produced by Summarization Evaluation Measures. En *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 69-78.

[28] Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2), pp. 264-285.

[29] Endres-Niggemeyer, B. (1998). *Summarizing Information*. Springer Verlag, Berlin.

[30] Erdös, P. y Rényi, A. (1959). On random graphs. En *I.Publ, Math.* 6, pp. 290-297.

[31] Ferrer-Cancho, R. y Solé, R.V. (2001). The small world of human language. En *Proceedings of the Royal Society of London,* 268, pp. 2261-2266.

[32] Fum, D., Gmda, G. y Tasso, C. (1985) Evaluating Importance: A step towards Text Summarization. En *IJCAI Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pp. 840-844.

[33] Goldstein, J., Mittal, V. O., Carbonell, J. y Kantrowitz, M. (2000). Multi-Document Summarization by Sentence Extraction. En *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA.

[34] Gómez-Pérez, A., Fernández-López , M. y Corcho , O. (2004). *Ontological Engineering*, 1ª ed., Hardcover.

[35] Gruber, T. R. A Translation Approach to Portable Ontologies. (1993). En *Knowledge Acquisition*, 5(2), pp. 199-220.

[36] Halliday, M. y Hasan, R. (1996). *Cohesion in English*. Longmans, London.

[37] Halliday, M.A.K. (1985). An Introduction to Functional Grammar. *Edward Arnold*.

[38] Hahn, U. y Reimer, U. (1999). Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. En Inderjeet Mani and Mark T. Maybury, *Advances in Automatic Text Summarization*, pp. 215–232. The MIT Press.

[39] Hahn, U. y Mani, I. (2000). The Challenges of Automatic Summarization. En *Computer,* 33(11), pp. 29-36.

[40] Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. En *Computational Linguistics*, 23(1), pp. 33-64.

[41] Hersh, W., Buckley, C., Leone, T.J. y Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. En *SIGIR94*, pp. 192-201.

[42] Hobbs, J. (1985). On the Coherence and Structure of Discourse. En *CSLI Technical Report,* pp. 85--37.

[43] Hovy, E. y Lin, C. Y. (1999). Automated Text Summarization in SUMMARIST. En I. Mani y M. T. Maybury, *Advances in Automatic Text Summarization*, The MIT Press.

[44] Hovy, E., Lin, C.-Y. y Zhou, L. (2005). Evaluating DUC 2005 Using Basic Elements. En *DUC 2005. Document Understanding Workshop.* Vancouver, B.C., Canada.

[45] Hovy, E. (2001). Automated Text Summarisation. En *Handbook of Computational Linguistics.* Oxford University Press.

[46] Jacobs, P. y Rau, L. (1990). SCISOR: Extracting Information from On-line News. En *Communications of the ACCM.*

[47] Jannink, J. y Wiederhold, G. (1999). Thesaurus entry extraction from an on-line dictionary. En *Fusion.*

[48] Jing, H., McKeown, K. R., Barzilay, R. y Elhadad, M. (1998). Summarization Evaluation Methods: Experiments and Analysis. En *Proceedings of the AAAISymposium on Intelligent Text Summarization*, pp. 60-68. Stanford, CA, USA.

[49] Knight, K. y Marcu, D. (2000). Statistics-Based Summarization - Step One: Sentence Compression. En *National Conference on Artificial Intelligence (AAAI).*

[50] Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Winters, S. y White, P. (2004). Integrated Annotation for Biomedical Information Extraction. En *HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases,* pp. 61-68.

[51] Kupiec, J., Pedersen, J. O. y Chen, F. (1995). A Trainable Document Summarizer. En *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68-73.

[52] Lee, L. (1999). Measures of distributional similarity. En *ACL.*

[53] Lin, D. (1998). An information-theoretic definition of similarity. En *ICML*.

[54] Lin, C.-Y. y Hovy, E. (2002). From Single to Multi-document Summarization: A Prototype System and its Evaluation. En *Proceedings of the ACL conference.* Philadelphia, PA.

[55] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. En *Proceedings of the ACL-04Workshop: Text Summarization Branches Out*, pp. 74-81, Barcelona, Spain.

[56] Longacre, R. (1979). The Discourse Structure of the Flood Narrative. En *Journal of the American Academy of Religion*, 47(1), pp. 89-133.

[57] Lowe, H. y Barnett, G. (1994). Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. En *Journal of the American Medical Association*, 271(14), pp. 1103-1108.

[58] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. En *IBM Journal of Research Development*, 2(2), pp. 159-165.

[59] Mani, I. y Bloedorn, E. (1998). Machine Learning of Generic and User-Focused Summarization. En *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98)*, pp. 821-826. Menlon Park, CA, USA.

[60] Mani, I., Bloedorn, E. y Gates, B. (1998). Using Cohesion and Coherence Models for Text Summarization. En *Proceedings of the AAAI Symposium on Intelligent Text Summarization*, pp. 69-76. Stanford, CA, USA.

[61] Mani, I. y Maybury, M. T. (1999). *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.

[62] Mani, I. y Bloedorn, E. (1999). Summarizing Similarities and Differences Among Related Documents. En *Information Retrieval*, 1(1), pp. 35-67.

[63] Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company, Amsterdam /Philadephia.

[64] Mani, I. (2001). Summarization Evaluation: An Overview. En *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*.

[65] Mann, W. y Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. En *Text*, 8(3), pp. 243-281.

[66] Maña, M.J., Buenaga, M. y Gómez, J.M. (1999). Using and Evaluating User Directed Summaries to Improve Information Access. En *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Lecture Notes in Computer Science, 1696, pp. 198-214.

[67] Marcu, D. (1997). From Discourse Structures to Text Summaries. En *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the*

8th Conference of the European Chapter of the Association for Computational Linguistics, pp. 82-88. Madrid, Spain.

[68] Marcu, D. (1999). Discourse trees are good indicator of importance in text. En I. Mani y M. T. Maybury, *Advances in Automatic Text Summarization*, pp. 123-136. The MIT Press.

[69] Marcu, D. (1999). The Automatic Construction of Large-Scale Corpora for Summarization Research. En *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 137-144. University of California, Berkeley, USA.

[70] Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

[71] Marcu, D. (2001). Discourse-based Summarization in DUC-2001. En *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New Orleans, LA.

[72] Marcu, D. y Gerber, L. (2001). An inquiry into the nature of multidocument abstracts, extracts and their evaluation. En *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, pp. 1-8, Pittsburg, USA.

[73] Mateo, P.L., González, J.C., Villena, J. y Martínez, L.L. (2003). Un sistema para resumen automático de textos en castellano. En *Procesamiento de Lenguaje Natural*, 31, pp. 29-36

[74] McKeown, K. y Radev, D. (1995). Generating summaries of multiple news articles. En *Proceedings of the ACM Conference on Research and Development in Information Retrieval SIGIR'95*, pp. 74–82. Seattle, Washington, USA.

[75] McKeown, K., Robin, J. y Kukich, K. (1995). Generating concise natural language summaries. En *Information Processing and Management*, 31(5), pp. 703-733.

[76] Mikheev, A. (1998). Part-of-speech guessing rules: learning and evaluation.

[77] Morris, J. y Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. En *Computational Linguistics*, 17(1), pp. 21-43.

[78] Nanba, H. y Okumura, M. (2000). Producing More Readable Extractsby Revising Them. En *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 1071-1075.

[79] Neto, J.L., Freitas, A. A. y Kaestner, C. A. A. (2002). Automatic Text Summarization Using a Machine Learning Approach. In *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*, pp. 205-215, Porto de Galinhas/Recife, Brazil.

[80] Nirenburg, S., McShane, M., Zabludowski, M., Beale, S. y Pfeifer, C. (2005). Ontological Semantic text processing in the biomedical domain. *Working Paper 03-05*, Institute for Language and Information Technologies, University of Maryland Baltimore County.

[81] Otterbacher, C., Radev, D. y Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. En *Proceedings of the Workshop on Automatic Summarization (including DUC 2002)*, pp. 27–36.

[82] Otterbacher, J., Erkan, G. y Radev, D. (2005). Using random walks for question-focused sentence retrieval. En *Proceedings of Human Language Technology Conference and Empirical Methods in Natural Language Processing (HLT/EMNLP).*

[83] Paice, C. D. (1981). The Automatic Generation of Literary Abstracts: An Approach Based on Identification of Self-Indicating Phrases. En O. R. Norman, S. E. Robertson, C. J. van Rijsbergen y P. W. Williams, *Information Retrieval Research*. London.

[84] Paice, C. D. (1990). Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, 26(1), pp. 171-186.

[85] Paice, C. y Jones, P. A. (1993). The Identification of Important Concepts in Highly Structured Technical Papers. En *Proceedings of the 16th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 69-78.

[86] Passonneau, R. y Nenkova, A. (2003). Evaluating content selection in human- or machine-generated summaries: The pyramid method. En *Technical Report CUCS025 -03.* Columbia University.

[87] Passonneau, R.J., Nenkova, A., McKeown, K. y Sigelman, S. (2005). Applying the pyramid method in DUC 2005. En *Proceedings of the Fifth Document Understanding Conference (DUC).* Vancouver, Canada.

[88] Perea, J.M., Martín, M.T., Montejo, A. y Díaz, M.C. (2008). Categorización de textos biomédicos usando UMLS. En *Procesamiento del Lenguaje Natural, Revista nº40*, pp. 121-127.

[89] Plaza, L., Díaz, A. y Gervás, P. (2008). Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina. En *Sociedad Española de Procesamiento de Lenguaje Natural.*

[90] Plaza, L., Díaz, A. y Gervás, P. (2008). Concept-graph based Biomedical Automatic Summarization using Ontologies. En *Coling 2008 Workshop TextGraphs-3: Graph-based Algorithms for Natural Language Processing.* Manchester, UK.

[91] Pustejovsky, J., Cochran, B., Castaños, J., Zhang, J., Morrell, M. y Luo, W. (2002). Medstract: Natural language tools for mining the biobibliome. En *Pacific Symposium on Biocomputing*.

[92] Radev, D. R., Jing, H. y Budzikowska, M. (2000). Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. En *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, USA.

[93] Radev, D. R. y McKeown, K. R. (1998). Generating Natural Language Summaries from Multiple On-Line Sources. En *Computational Linguistics*, 4, pp. 469-500.

[94] Radev, D. R., Hovy, E. y McKeown, K. (2002). En *Computational Linguistics* 28(4). The MIT Press.

[95] Riloff, E. y Jones, R. (1996). Automatically Generating Extraction Patterns from Untagged Text. En *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp. 1044-1049.

[96] Rindflesch, T.C., Tanabe, L., Weinstein, J.N. y Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. En *Pacific Symposium on Biocomputing,* pp. 517-528.

[97] Rush, J. E., Zamora, A. y Salvador, R. (1971). Automatic Abstracting and Indexing. II, Production of Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. En *Journal of the American Society for Information Science*, 22(4), pp. 260-274.

[98] Salton, G., Allan, J., Buckley, C. y Singhal, A. (1994). Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts. En *Science* , 264, pp. 1421-1426.

[99] Salton, G., Singhal, A., Mitra, M. y Buckley, C. (1997). Automatic Text Structuring and Summarization. En *Information Processing & Management*, 33(2), pp. 193-207.

[100] Skorochod'ko, E. F. (1972). Adaptive method of automatic abstracting and indexing. En *Processing 71: Proceedings of the IFIP Congress,* 71, pp. 1179-1182.

[101] Sneiderman, C.A., Rindflesch, T.C. y Bean, C.A. (1998). Identification of anatomical terminology in medical text. En *Proc AMIA Symp,* pp. 428-432.

[102] Sparck-Jones, K. y Galliers, J. R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review. En *Lecture Notes in Artificial Intelligence*, 1083.

[103] Sparck-Jones, K. (1999). Automatic Summarizing: Factors and Directions. En I. Mani y M.T. Maybury, *Advances in Automatic Text Summarization*. The MIT Press.

[104] Spackman, K.A., Campbell, K.E. y Côté, R.A. (1997). SNOMED-RT: A Reference Terminology for Health Care. En *Proceedings of the AMIA Annual Fall Symposium*, pp. 640-644.

[105] Steve, G., Gangemi, A. y Pisanelli, D. (1998). Integrating Medical Terminologies with ONIONS Methodology. http://saussure.irmkant.rm.cnr.it

[106] Tanabe, L., Xie, N., Thom, L.H., Matten, W. y Wilbur, W.J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. En *BMC Bioinformatics*, 6(1).

[107] Teufel, S. y Moens, M. (1997). Sentence Extraction as a Classification Task. En *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain.

[108] Teufel, S. y Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. En *Computational Linguistics*, 28.

[109] Tombros, A. y Sanderson, M. (1998). Advantages of Query Biased Summaries in Information Retrieval. En *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 2-10.

[110] Van Dijk, T. (1988) *News as discourse*. Erlbaum Associates.

[111] Van Dijk, T. (1988) News Analysis: Case Studies of International and National News in the Press. *Erlbaum Associates*.

[112] Verna, R., Chen, P. y Lu, W. (2007). A Semantic Free-text Summarization Using Ontology Knowledge. En *Proceedings of the Document Understanding Conference*.

[113] Watts, D.J. y Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. En *Nature*, 393, pp. 440-442.

[114] Weigand, H. (1997). Multilingual Ontology-Based Lexicon for News Filtering. En *The TREVI Project*, pp. 138-159.

[115] White, M., Cardie, C., Ng, V. y McCullough, D. (2002). Detecting discrepancies in numeric estimates using multidocument hypertext summaries. En *Proceedings of the Second International Conference on Human Language Technology Research*.

[116] Witbrock, M. y Mittal, V. (1999). Ultra-summarization: a Statistical Approach to Generating Highly Condensed Non-extractive Summaries. En *Proceedings of the 22nd Annual iInternational ACM SIGIR Conference on Research and Development in Information Retrieval.*

[117] Zhou, L., Ticrea, M. y Hovy, E. (2004). Multi-document Biography Summarization. En *Proceedings of Proceedings of theConference on Empirical Methods in Natural Language Processing (EMNLP-04)*. Barcelona, Spain.

Índice de Figuras

Índice de Tablas