UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

USO DE GRAFOS SEMÁNTICOS EN LA GENERACIÓN AUTOMÁTICA DE RESÚMENES Y ESTUDIO DE SU APLICACIÓN EN DISTINTOS DOMINIOS: BIOMEDICINA, PERIODISMO Y TURISMO.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Laura Plaza Morales

Bajo la dirección de los doctores

Pablo Gervás Gómez-Navarro
Alberto Díaz Esteban

Madrid, 2011

ISBN: 978-84-694-2887-0

Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo

Tesis doctoral

Presentada por

Laura Plaza Morales

para optar al grado de Doctor en Informática

Dirigida por

Prof. Dr. D. Pablo Gervás Gómez-Navarro

Prof. Dr. D. Alberto Díaz Esteban

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Madrid, diciembre de 2010

Agradecimientos

Desearía agradecer sinceramente a mis directores de tesis, Alberto Díaz y Pablo Gervás, por haberme acordado su confianza a lo largo de estos años de investigación. Sin su apoyo y ayuda, cabe decir, la realización de este trabajo no hubiese sido posible.

Al Departamento de Ingeniería del Software e Inteligencia Artificial de la Universidad Complutense de Madrid, y en especial, al grupo de investigación NIL, gracias por haberme acogido tan calurosamente.

Agradezco igualmente a todas esas personas con las que he tenido el placer de colaborar, en la distancia y en la cercanía, en proyectos relacionados con esta tesis doctoral. Gracias, en especial, a Mark Stevenson y Ahmet Aker, del grupo de Procesamiento de Lenguaje Natural de la Universidad de Sheffield, y a Elena Lloret, del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante.

Por último, y no por ello menos importante, agradezco a mi familia, en especial a Jorge, el haberme expresado su cariño y apoyo en todo momento.

Resumen

En la sociedad en la que vivimos, la información se ha convertido en un bien necesario, a la vez que altamente cotizado, que nos acompaña en todas y cada una de nuestras actividades sociales, culturales y económicas cotidianas. Sin embargo, el crecimiento exponencial de la información a nuestro alcance se presenta como una amenaza a su uso efectivo para la formación de conocimiento, ya que, si bien la disponibilidad de esta información parece estar garantizada, no ocurre lo mismo con el tiempo necesario para procesarla.

La *Generación Automática de Resúmenes (GAR)* trata, en cierta medida, de paliar los efectos negativos de la sobrecarga de información sobre la capacidad de los usuarios para obtener aquella que realmente les interesa y transformarla en conocimiento. Los resúmenes generados automáticamente pueden utilizarse como sustitutos de los documentos originales o simplemente como referencia en la selección de documentos para una lectura más profunda. Más aún, resultan de gran utilidad como paso intermedio en otras tareas de Procesamiento de Lenguaje Natural (PLN).

La generación de resúmenes es una de las tareas más complejas de las encuadradas dentro de la más amplia disciplina del PLN, debido a la gran cantidad de otras tareas que implícitamente conlleva, como la detección de temas, la desambiguación léxica, la resolución de referencias, la simplificación de oraciones o la eliminación de redundancia. Si bien casi todas ellas han sido ampliamente estudiadas en la literatura, en el momento de escribir esta memoria no se conoce ningún trabajo que analice el efecto de la ambigüedad en el texto a resumir sobre la calidad de los resúmenes generados automáticamente. Es por ello que en esta tesis doctoral se presta especial atención a la resolución de ambigüedades como un paso previo a la generación del resumen. Tal y como demuestran los resultados, la ambigüedad repercute negativamente en la generación automática de resúmenes, de tal modo que es posible mejorar significativamente la calidad de los resultados

mediante el uso de los algoritmos apropiados de desambiguación léxica.

El trabajo se completa con tres casos de estudio en los que el método diseñado se configura y utiliza para generar distintos tipos de resúmenes de textos de diferentes dominios y con unas características de estructura y estilo muy dispares: artículos científicos de biomedicina, noticias periodísticas y páginas web de información turística. Los resúmenes generados son evaluados utilizando las métricas ROUGE y los criterios de legibilidad adoptados en las *Document Understanding Conferences*, y se comparan con los generados por otros sistemas automáticos y con los elaborados por seres humanos. Los resultados corroboran la adecuación del método propuesto a la tarea que nos ocupa.

Abstract

In recent years, with the increasing publication of online information, providing mechanisms to facilitate finding and presenting textual information has become a critical issue. New technologies, such as high-speed networks and massive storage, are supposed to improve work efficiency by assuring the availability of data everywhere at anytime. However, the exorbitant volume of data available threatens to undermine the convenience of information if no effective access technologies are provided. In this context, automatic text summarization may undoubtedly help to optimize the treatment of electronic documentation and to tailor it to the needs of users.

Automatic summarization is one of the most complex Natural Language Processing (NLP) tasks, and this is due to the number of other tasks that implicitly entails, such as topic detection, word sense disambiguation, anaphoric resolution, acronym expansion, sentence simplification and redundancy detection. In particular, this thesis studies a crucial issue that has been previously unexplored, as is the effect of lexical ambiguity in the knowledge source on semantic approaches to summarization, and demonstrates that using word sense disambiguation techniques leads to an improvement in summarization performance.

A controversial decision when designing a summarization system is whether it should be general (i.e. able to produce summaries for any type of document) or whether it should be changed by text types (i.e. be specific to documents of a given genre and structure). The advantage of the former is obvious, but the latter strategy has proved to be more effective and capable of improving the quality of the summaries. The main contribution of this thesis is the development of a generic summarization method that combines the advantages of both approaches, by taking into account the structure, genre and domain of the document to be summarized, but is easily configurable to work with new types of documents. The method proposed addresses

the problem of identifying salient sentences in a document by representing the text as a semantic graph, using concepts and relations from a knowledge source. This way it gets a richer representation than the one provided by traditional models based on terms. A degree-based clustering algorithm is then used to discover different themes or topics within the text. Different heuristics for sentence selection aiming to generate different types of summaries are tested.

The thesis also presents three case studies, in which the summarizer has been configured and used to generate summaries of texts from different domains and with very distinct structure and style: biomedical scientific articles, news items and tourism-related websites. The system is evaluated using the ROUGE metrics and the legibility criteria followed in the DUC conferences. It has been found that it compares favorably with existing approaches.

Glosario de Acrónimos y Abreviaturas

BC Base de Conocimiento

BE Basic Elements (Elementos Básicos)

CIE Clasificación Internacional de Enfermedades

CAP College of American Pathologists (Colegio de Patólogos Americanos)

CSIS Cross-Sentence Information Subsumption (Subsunción de Información entre Oraciones)

CST Cross-Document Structure Theory (Teoría de la Estructura Inter-Documento)

DUC Document Understanding Conferences (Conferencias sobre Comprensión de Documentos)

GAR Generación Automática de Resúmenes

GATE Generic Architecture for Text Engineering (Arquitectura General para Ingeniería de Texto)

HRCA Highly Referenced Concept Assumption (Suposición de Conceptos Altamente Referenciados)

HVS Hub Vertex Set (Conjunto de Vértices Concentradores)

IS Information Subsumption (Subsunción de Información)

IHTSDO International Health Terminology Standards Development Organisation (Organización Internacional para el Desarrollo de Estándares sobre Terminologías de Salud)

IDF Inverse Document Frequency (Frecuencia Inversa del Documento)

LCS Least Common Subsumer (Ancestro Común Más Específico)

LKB Lexical Knowledge Base (Base de Conocimiento Léxico)

LCA Local Context Analysis (Análisis del Contexto Local)

MMR Maximum Marginal Relevance (Relevancia Marginal Máxima)

MMR-MD Maximum Marginal Relevance-MultiDocument (Relevancia Marginal Máxima Multi-documento)

MeSH Medical Subject Headings (Encabezados de Temas Médicos)

NCI National Cancer Institute (Instituto Nacional de Cáncer de los Estados Unidos)

NLM National Library of Medicine (Biblioteca Nacional de Medicina de los Estados Unidos)

PPR Personalized PageRank (PageRank Personalizado)

PPR-w2w Personalized PageRank *word to word* (PageRank Personalizado *palabra por palabra*)

PLN Procesamiento de Lenguaje Natural

ROUGE Recall-Oriented Understudy for Gisting Evaluation (Evaluación Basada en la Cobertura)

RST Rhetorical Structure Theory (Teoría de la Estructura Retórica)

SCU Summarization Content Units (Unidades de Resumen de Contenido)

SEE Summary Evaluation Environment (Entorno de Evaluación de Resúmenes)

SNOMED-CT Systematized Nomenclature of Medicine-Clinical Terms (Nomenclatura Sistematizada de Medicina-Términos Clínicos)

TF Term Frequency (Frecuencia del Término)

TAC Text Analysis Conference (Conferencia sobre Análisis Textual)

TSIN Text Semantic Interaction Network (Red de Interacción Semántica Textual)

UMLS Unified Medical Language System (Sistema de Lenguaje Médico Unificado)

WSD Word Sense Disambiguation (Desambiguación de Significados)

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1. Motivación

En la era en la que vivimos, la creación, distribución y manipulación de la información forman parte fundamental de las actividades sociales, culturales y económicas cotidianas. La cantidad de documentos electrónicos accesible desde cualquier lugar y cualquier dispositivo crece de manera exponencial, pero el tiempo disponible para procesarlos sigue siendo un recurso valioso y limitado. Este estado de exceso o sobrecarga de información, también conocido como *infoxication*, se ha ido agravando a medida que los avances tecnológicos conseguidos en las últimas décadas han fomentado una dinámica de crecimiento en la información generada y almacenada, al tiempo que han facilitado su distribución.

El exceso de información se presenta como una amenaza a la formación de conocimiento, entendido éste, desde una visión científico-técnica, como el resultado de procesar la información para convertirla en acciones efectivas encaminadas a resolver un problema o tomar una decisión (el conocido como *conocimiento accionable*). Así, el término sobrecarga de información fue acuñado por el economista Ackoff en 1967, al defender que uno de los problemas críticos de los sistemas de información era su incapacidad para evitar al directivo una sobrecarga de información irrelevante. Más recientemente, numerosos estudios han subrayado cómo el exceso de información puede hacer nuestro trabajo menos productivo, a la vez que producir estrés y ansiedad (Klingberg, 2009).

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, arbitrariamente ordenados, o sin orden, y que pre-existe al hombre y es infinita. Como en la biblioteca de Babel, en la web la sobrecarga de información es tal que su localización, organización y comprensión se ven muy limitadas. Las revistas científicas, los periódicos, las informaciones y estados financieros de las empresas, e incluso los libros, son accedidos cada vez más exclusivamente a través de Internet. Es por ello que la inmensa mayoría de la investigación en procesamiento de información está dirigida a este medio, caracterizado por su continuo dinamismo y crecimiento, y por la naturaleza heterogénea de la información que contiene (texto, imágenes, vídeo, sonido, etc.).

La *Generación Automática de Resúmenes (GAR)* trata, en cierta medida, de paliar los efectos negativos de la sobrecarga de información sobre la capacidad de los usuarios para obtener aquella que realmente les interesa y transformarla en conocimiento accionable. Supongamos, por ejemplo, que realizamos una búsqueda en Internet utilizando una serie de palabras como criterios. Al instante, el buscador nos devuelve cierto número de páginas que satisfacen tales criterios. Con elevada probabilidad, y en función de las características del buscador y de nuestro acierto a la hora de seleccionar los criterios de búsqueda, la mayoría de las páginas guardarán poca o ninguna relación con la información que deseamos encontrar. Pero aún suponiendo que logremos aislar los documentos que realmente se ajustan a nuestros intereses, lo más probable es que, dentro de estos documentos, un gran porcentaje de la información que contienen sea irrelevante; más aún, redundante. Sin duda, disponer de resúmenes de estos documentos puede ser de gran utilidad: por un lado, dado un único documento, es posible producir un resumen del mismo en el que se obvie o limite la información irrelevante; por otro lado, dado un conjunto de documentos que versen sobre un mismo tema, se puede generar un único resumen que condense la información común a todos ellos y que, además, incluya otra información de los documentos individuales que pudiera ser de interés para el lector.

Los resúmenes generados automáticamente pueden utilizarse como sustitutos de los documentos originales o simplemente como referencia en la selección de documentos para una lectura más profunda. La primera de estas aplicaciones la encontramos, por ejemplo, en la generación de resúmenes sobre noticias periodísticas, donde es posible encontrar sistemas comple-

tamente operativos como *NewsBlaster*[1], de la Universidad de Columbia. NewsBlaster recopila noticias de distintos periódicos digitales, agrupa aquellas que conciernen al mismo suceso, y realiza un resumen de todas ellas para mostrar al usuario una única noticia sobre el suceso. De este modo, el usuario ya no necesita consultar distintos periódicos para conocer en detalle el acontecimiento o las diferentes opiniones y puntos de vista de los distintos medios. Más aún, si se combina la generación de resúmenes con un sistema de personalización de noticias, el usuario podrá acceder únicamente a aquellas que le interesen según indique su modelo de usuario (Díaz y Gervás, 2005). En cuanto a la segunda aplicación se refiere, es decir, utilizar los resúmenes como indicadores de la relevancia de un documento, su utilidad es más que evidente en el ámbito académico y de la investigación, donde la labor documental conlleva leer y procesar un ingente número de publicaciones que a menudo resultan de escaso interés, o incluso parcialmente redundantes con respecto a otras publicaciones procesadas con anterioridad. Habida cuenta del escaso tiempo del que a menudo disponen los investigadores, el resumen surge como una ayuda inestimable.

Para terminar, conviene señalar que la generación automática de resúmenes resulta también de gran utilidad como paso intermedio en otras tareas de procesamiento de lenguaje natural. Así, por ejemplo, se ha demostrado que el uso de resúmenes en lugar de los documentos originales en tareas de recuperación y categorización de información produce un ahorro de tiempo sin pérdida significativa de efectividad (Mani et al., 2001; Saggion, Lloret, y Palomar, 2010).

1.2. La Generación Automática de Resúmenes

1.2.1. Definición del Problema

La generación automática de resúmenes consiste en la creación de una versión reducida de uno o varios documentos por parte de un programa de ordenador, de tal forma que el resumen producido condense la información importante del texto de entrada. Se trata de una disciplina con más de medio siglo de vida que, sin embargo, ha despertado el interés de la comunidad científica en las dos últimas décadas, motivada por el auge de Internet.

[1]Columbia NewsBlaster. http://newsblaster.cs.columbia.edu/. Consultada el 1 de noviembre de 2010

Para Mani (2001), en la elaboración de un resumen se han de tener en cuenta tres tipos de factores de contexto. En primer lugar, se deben considerar las propias características del documento que se desea resumir; es decir, su estructura, tamaño, o la especificidad del contenido tratado. En segundo lugar, se ha de considerar la aplicación o uso que se pretende dar al resumen. Así, si su propósito es servir para seleccionar documentos interesantes para una lectura posterior, sin duda la longitud del resumen será menor que si se espera utilizarlo como sustituto del documento original. Por último, el formato en que se presenta el resumen al usuario es también, en ciertos contextos, importante. Así, por ejemplo, un sistema web de noticias que, además del resumen, presente los enlaces a las distintas noticias de las que procede la información, proporciona, sin duda, un valor añadido. Todos estos elementos fueron previamente identificados por Sparck-Jones (1999), quien los denominó, respectivamente, factores de entrada, propósito y salida, y que serán analizados con más detalle en la Sección 2.1.

Los factores que intervienen en la elaboración del resumen determinan distintos tipos de resúmenes. La Sección 2.1 presenta una clasificación exhaustiva de los mismos. Baste aclarar aquí que, dada la complejidad y número de factores a considerar y el estado de la técnica actual, no resulta factible diseñar un sistema capaz de realizar resúmenes adecuados a cualquier escenario. Es por ello que la práctica totalidad de la investigación realizada hasta el momento, o bien se circunscribe a tipos de documentos específicos, o bien pretende abarcar documentos de cualquier tipo a costa de reducir la calidad de los resúmenes generados. Resulta evidente que no es lo mismo resumir un artículo científico en biomedicina que una noticia periodística. La primera diferencia la encontramos en la longitud y, por tanto, en la concisión. Así, la noticia suele ser mucho más concisa y, en general, presentará menos información redundante. La segunda y principal diferencia radica en las características del lenguaje utilizado: mientras que el lenguaje biomédico emplea una terminología muy especializada, el lenguaje periodístico hace uso de un vocabulario mucho más amplio, plagado de frases hechas y metáforas. La estructura también difiere de un tipo de documento a otro: mientras que los artículos científicos suelen estar estructurados en secciones más o menos estándar que, en cierto modo, predeterminan la naturaleza de la información incluida en cada una de ellas, las noticias periodísticas generalmente no presentan ningún tipo de subdivisión. Sin embargo, en estas

últimas es posible encontrar una estructura implícita, comúnmente denominada *estructura piramidal*, según la cuál la información más importante generalmente aparece en las primeras oraciones.

Más allá de todo lo anterior, la generación automática de resúmenes es, indudablemente, una tarea compleja, debido a la gran cantidad de otras tareas que implícitamente conlleva.

- En primer lugar, al elaborar un resumen textual es necesario delimitar los distintos temas tratados en el texto de entrada (*Detección de temas*) para, posteriormente, discernir cuáles de ellos constituyen el objeto de interés del lector.

- En segundo lugar, para identificar estos temas y determinar su importancia, es necesario resolver la ambigüedad del texto y asociar cada término ambiguo con su significado correcto en función del contexto en el que se utiliza (*Desambiguación léxica*). Ilustraremos esta afirmación con un ejemplo. Supongamos que se desea resumir un conjunto de noticias que tratan sobre la solvencia de los bancos y cajas de ahorros españolas. Puesto que el término "banco" en castellano presenta diferentes acepciones, es necesario determinar que la acepción correcta es aquella que lo define como una entidad crediticia. Sólo así el sistema de resúmenes podrá decidir que, a priori, la información relacionada con los conceptos "banco", "caja de ahorros" o "entidad financiera" es importante. De otro modo, si la acepción seleccionada fuera, por ejemplo, la que lo define como un conjunto de peces, el sistema determinaría que toda la información en torno al concepto "banco" no guarda ninguna relación con el resto del documento y, por lo tanto, no se incluiría en el resumen.

- En tercer lugar, es necesario resolver las referencias anafóricas y pronominales presentes en el texto, así como identificar a los respectivos referentes (*Resolución de referencias*). Una referencia es una expresión que denota un elemento o individuo, mientras que el referente es el propio elemento referido. Tanto para la correcta delimitación de la información relevante como para la adecuada presentación del resumen, es importante determinar las diferentes formas en que un mismo elemento es referido en el documento original. De este modo, por ejemplo, supongamos que deseamos resumir un artículo científico sobre la

prescripción farmacéutica en el tratamiento del resfriado común, que presenta las dos siguientes oraciones:

1. *Existen muchos virus implicados en la aparición del* ***resfriado común****.*
2. *El tratamiento de* ***la enfermedad*** *dependerá del virus que la causa.*

Si el sistema ya sabe que la información relacionada con el concepto "resfriado común" es relevante, para determinar que la información en la segunda oración también lo es deberá saber que "la enfermedad" no es más que otra manera de referirse al "resfriado común". Pero el problema va más allá: si el sistema determina que la segunda oración es relevante pero no la primera, y decide incluirla en el resumen, el resultado será un resumen incoherente, ya que, sin la primera oración, el lector no sabrá cuál es el referente de "la enfermedad".

- En cuarto lugar, y en estrecha relación con el problema anterior, es necesario identificar posibles acrónimos y abreviaturas (*Resolución de acrónimos*) y resolverlos para conocer las versiones expandidas de los mismos.

- En quinto lugar, y especialmente cuando se desea redactar resúmenes con una elevada tasa de compresión (es decir, resúmenes muy breves con respecto a la longitud del documento original), es necesario realizar un proceso de simplificación y combinación de oraciones con el fin de reducir la información irrelevante a su mínima expresión (*Simplificación de oraciones* y *Fusión o concatenación de oraciones*).

- Finalmente, si se ha de realizar un único resumen de un conjunto de documentos sobre un mismo tema, es necesario detectar aquella información que se repite a lo largo de los distintos documentos, con el objetivo de no incluirla repetida en el resumen (*Detección de redundancia*).

Sin embargo, y a pesar de todos los problemas planteados, la autora de este documento no conoce ningún trabajo en el que se aborden todos ellos. Es posible encontrar en la literatura ejemplos de sistemas que realizan resolución de referencias como paso previo a la selección de oraciones para

el resumen (Steinberger et al., 2007), y que demuestran que el uso de esta técnica se traduce en una mejora de la calidad de los resúmenes generados. También existen trabajos en los que las oraciones seleccionadas se simplifican y condensan para reducir la longitud del resumen final (Barzilay y McKeown, 2005; Filippova y Strube, 2008). Por otra parte, la detección de redundancia en sistemas de resúmenes multi-documento es una práctica bastante habitual (Zhao, Wu, y Huang, 2009; Plaza, Lloret, y Aker, 2010). Sin embargo, no se conocen trabajos, salvo los desarrollados como parte de esta tesis doctoral, en los que se analicen los efectos de la desambiguación y de la resolución de acrónimos y abreviaturas sobre la generación de resúmenes.

1.2.2. Enfoques Actuales

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. Un resumen por extracción es aquel que se compone íntegramente por material (típicamente oraciones) presente en el texto de entrada. Por el contrario, un resumen por abstracción puede incluir contenidos que no están presentes, al menos explícitamente, en el documento original.

Estudios recientes han demostrado que, en contra de la creencia inicial, los seres humanos generalmente realizan resúmenes mediante la selección de oraciones del documento original y, sólo en aquellos casos en los que se exige una compresión muy elevada del texto, utilizan técnicas de abstracción y reescritura (Banko y Vanderwende, 2004). Partiendo de esta misma hipótesis, Jing (2002) presenta una descomposición de los resúmenes realizados por personas con el objetivo de determinar cómo se generan estos resúmenes. Como resultado, se identificaron seis tipos de operaciones: simplificación de oraciones, combinación de oraciones, transformación sintáctica, parafraseado léxico, generalización y especialización. Sin embargo, los experimentos realizados sobre un corpus de 300 resúmenes mostraron que el 81 % de las oraciones en los resúmenes se correspondían íntegramente con oraciones presentes en los documentos originales.

Ya sea por este motivo, o por las dificultades que entraña la generación de resúmenes por abstracción, lo cierto es que durante los últimos años el interés y el esfuerzo de la comunidad científica se ha ido progresivamente trasladando hacia las técnicas de generación de resúmenes mediante extrac-

ción, en detrimento de las técnicas de abstracción.

Dentro de las técnicas de generación de resúmenes por extracción es posible distinguir, a su vez, muy diversos enfoques atendiendo a la profundidad del análisis realizado. Así, en sus inicios, la mayor parte de la investigación hizo uso de técnicas superficiales y heurísticas sencillas para identificar los segmentos relevantes en el documento a resumir, como por ejemplo, la posición de las distintas oraciones en el documento, la frecuencia de los términos que lo componen o el uso de ciertas expresiones o frases indicativas de la importancia de la oración en el documento. Posteriormente, y sobre todo durante la última década, el interés se ha centrado en enfoques que realizan un cada vez más sofisticado análisis del lenguaje natural para identificar el contenido relevante del documento. Para ello, analizan las relaciones entre palabras o la estructura del discurso. Con estos enfoques, se pretende solucionar los problemas de coherencia y cohesión inherentes a los enfoques superficiales, a la vez que se consigue una identificación más precisa del contenido destacado del documento. Como se verá en el capítulo dedicado al estudio del trabajo previo, ejemplos de estos enfoques hay muchos, y las técnicas utilizadas son muy variadas. Dentro de este último enfoque, que podríamos denominar discursivo, merecen especial atención, por su estrecha relación con este trabajo, los métodos basados en grafos. Típicamente, estos enfoques representan el texto como una red compleja en la que los nodos representan cada una de las unidades textuales en las que se divide el texto y las aristas representan algún tipo de relación entre estas unidades, generalmente de naturaleza léxica o sintáctica. La idea subyacente en este tipo de enfoques es la emergencia en la red de grupos de unidades que guardan estrecha relación entre sí y que determinan la información relevante del documento. La Sección 2.2 presenta un estudio exhaustivo de las características y la evolución de todos estos enfoques.

1.2.3. La Evaluación de Resúmenes Automáticos

La Sección 2.3 se dedica íntegramente al estudio de las estrategias y métricas más comúnmente utilizadas en la evaluación de resúmenes automáticos. Baste pues con recalcar en esta introducción que tal evaluación está llena de complicaciones. En primer lugar, evaluar un resumen significa medir su calidad en relación a dos tipos generales de características deseables: su contenido informativo y su calidad gramatical o legibilidad. Si bien la pri-

mera de estas características puede ser evaluada, con mayor o menor éxito, de manera automática, a día de hoy la evaluación de la legibilidad exige la participación de seres humanos. Si tenemos en cuenta que, para que la evaluación de un sistema sea estadísticamente significativa, el número de resúmenes a evaluar asciende a decenas, o incluso a cientos, y que, además, es deseable la participación de al menos tres jueces para compensar las posibles y frecuentes divergencias en sus juicios, es fácil darse cuenta de que no siempre es posible realizar este tipo de evaluaciones.

Pero incluso la evaluación del contenido informativo de los resúmenes requiere cierta participación humana. La manera más habitual de llevar a cabo esta evaluación consiste en la comparación del resumen generado automáticamente con respecto a uno o varios resúmenes realizados por personas. Cuanto más contenido comparten entre sí, mayor se presupone la calidad informativa del resumen automático. Una de las principales limitaciones de este tipo de evaluación es, por tanto, la necesidad de estos resúmenes manuales con los que comparar, cuya elaboración requiere mucho tiempo y esfuerzo. No obstante disponemos, gracias a las conferencias *DUC (Document Understanding Conferences)*[2] y *TAC (Text Analysis Conferences)*[3], de distintas colecciones de evaluación de resúmenes de noticias periodísticas. Además, los artículos científicos a menudo vienen acompañados de un *abstract* o resumen realizado por su propio autor que puede ser utilizado como modelo para la evaluación. Finalmente, diversos autores han elaborado colecciones similares en otros dominios, como libros de texto (Kazantseva y Szpakowicz, 2010) o páginas web turísticas (Aker y Gaizauskas, 2009).

1.2.4. Nuevas Aplicaciones de la Generación Automática de Resúmenes

La generación automática de resúmenes, tal y como fue inicialmente concebida, ha evolucionado progresivamente y dado lugar a un amplio espectro de tareas y aplicaciones. En la Sección 2.4 se estudian en detalle algunas de las más difundidas, como son la generación de resúmenes multi-documento, resúmenes adaptados a una consulta de usuario y resúmenes multilingüe.

De todas estas variantes de la tarea original, nos interesa especialmente

[2]Document Understanding Conferences (DUC). http://duc.nist.gov/. Consultada el 1 de noviembre de 2010

[3]Text Analysis Conferences (TAC). http://www.nist.gov/tac/. Consultada el 1 de noviembre de 2010

la generación de resúmenes multi-documento. La tarea de generar resúmenes a partir de múltiples fuentes es más compleja que la generación de resúmenes de un solo documento, y plantea retos adicionales. El principal de ellos es la detección de redundancia, o lo que es lo mismo, del contenido que se repite en los diferentes documentos a resumir. Aunque a priori pueda parecer sencillo, en absoluto lo es, ya que este contenido puede venir expresado de muy diferentes formas, con lo que no es suficiente detectar posibles coincidencias entre palabras u oraciones para afirmar que dos unidades textuales significan lo mismo.

1.3. Propuesta y Objetivos

La generación automática de resúmenes es una tarea compleja, dada la amplitud de cuestiones a tener en cuenta a la hora de producir un resumen de calidad. Existen muchas y muy diversas aproximaciones al problema que nos ocupa. La mayoría de ellas parten de una representación del documento que consiste únicamente en información que puede extraerse directamente del documento a resumir (generalmente, las palabras o sintagmas que componen el documento), ignorando los beneficios que el uso de representaciones más ricas y complejas puede aportar. Este enfoque presenta algunos problemas importantes derivados de no considerar la estructura semántica del documento y de las relaciones existentes entre los términos que lo componen (por ejemplo, sinonimia, hiperonimia, homonimia, co-ocurrencia o asociación semántica). Para ilustrar algunos de estos problemas, consideremos las siguientes oraciones:

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

2. ***Brain vascular 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 un significado común. El uso de fuentes de conocimiento externo para capturar el significado y las

relaciones entre los términos de ambas oraciones puede ayudar, por ejemplo, a determinar que "pregnancy" y "gestation" son conceptos sinónimos. Por el contrario, los enfoques tradicionales basados en términos y en meras representaciones sintácticas difícilmente podrán capturar esta sinonimia, lo que afectará a la calidad de los resúmenes generados. Pensemos si no en un enfoque simple de frecuencia de términos según el cual, la palabra "pregnancy", por su elevada frecuencia en el documento, se considera importante o representativa del tema principal del documento. Un enfoque de este tipo tenderá a seleccionar aquellas oraciones que contengan dicha palabra pero, erróneamente, considerará a las oraciones que contengan la palabra "gestation" poco importantes.

Más aún, supongamos las siguientes dos oraciones:

1. ***Pneumococcal pneumonia*** *is a lung infection caused by Streptococcus pneumoniae.*

2. ***Mycoplasma pneumonia*** *is another type of atypical pneumonia.*

Con un enfoque léxico o sintáctico nunca podremos saber que "pneumococcal pneumonia" y "mycoplasma pneumonia" son dos tipos de neumonía cuya única diferencia subyace en el tipo de bacteria que las origina. Por el contrario, el uso de una adecuada fuente de conocimiento biomédico permite, a través de las distintas relaciones entre los dos conceptos (por ejemplo, a través de la existencia del hiperónimo común "pneumonia"), determinar esta similitud entre ambos.

Todo lo expuesto justifica la realización de este trabajo, en el que se pretende adoptar un enfoque fundamentalmente semántico y lingüísticamente motivado, que haga un uso intensivo de conocimiento del dominio y de los recursos disponibles, para la realización automática de resúmenes por extracción.

Otro problema importante que, en opinión de la autora, comparten la mayoría de los sistemas actuales, es la falta de un proceso previo de desambiguación léxica. Aún aquellos que utilizan recursos de conocimiento externo, ante la presencia de términos ambiguos, generalmente se limitan a seleccionar aleatoriamente una de sus posibles acepciones o, a lo sumo, escoger la acepción más frecuente. Esto limita, como ya se ha comentado, la calidad de los resúmenes, puesto que si se escogen acepciones erróneas para algunos de los conceptos ambiguos del documento, éstos posiblemente serán interpreta-

dos como información inconexa y no relacionada con el resto de conceptos del documento. Por todo ello, el sistema desarrollado en esta tesis doctoral permite aplicar distintos algoritmos de desambiguación léxica sobre el documento o documentos a resumir.

Por otro lado, si bien disponer de una base de conocimiento sobre el dominio de los textos a resumir se ha demostrado que mejora la calidad de los resúmenes generados, presenta la desventaja de hacer el sistema aplicable únicamente a documentos del dominio considerado. Para solventar esta limitación, en este trabajo se propone un método genérico para la generación de resúmenes dependientes del dominio, aunque fácilmente configurable para trabajar con documentos de diferentes dominios sin más que disponer de una base de conocimiento del mismo, y de un algoritmo o estrategia que permita desambiguar posibles términos ambiguos y asignarles la acepción correcta dentro de las contempladas en la base de conocimiento.

El algoritmo que en este trabajo se propone ha sido diseñado para su uso tanto en generación de resúmenes mono-documento como multi-documento. No obstante, el interés se centra, fundamentalmente, en la generación de resúmenes mono-documento. En opinión de la autora, y a pesar de que en los últimos años la comunidad científica se ha decantado, de manera prioritaria, en el desarrollo de sistemas multi-documento, el estado actual de la investigación en generación de resúmenes mono-documento se encuentra lejos de ser satisfactorio, como prueba el hecho de que la utilización de este tipo de sistemas en entornos reales es prácticamente inexistente. Tal y como demuestran estudios recientes (Nenkova, 2005) sobre el desempeño de los sistemas presentados a las tareas competitivas de las conferencias DUC y TAC, el problema sigue abierto. Es más, los resultados demuestran que, sorprendentemente, los sistemas mono-documento presentan, en general, peores resultados que los sistemas multi-documento. La explicación parece estar en el hecho de que, cuando se realiza un resumen a partir de múltiples documentos sobre un mismo tema, la información que se repite a lo largo de los distintos documentos sirve de pista acerca de lo que es importante en los mismos; mientras que cuando se realiza el resumen a partir de un único documento, no se dispone de este tipo de indicios.

A continuación se procede a enumerar los objetivos concretos perseguidos en esta tesis doctoral.

1. Estudiar la generación automática de resúmenes como un problema

consistente en detectar los distintos temas tratados en un texto y determinar cuáles de estos temas constituyen información relevante dentro del mismo. En estrecha relación con lo anterior, se propondrán distintas heurísticas encaminadas a construir distintos tipos de resúmenes en función de la cobertura de información deseada. Se estudiará el problema desde un punto de vista semántico; es decir, dirigido por el contenido conceptual del texto a resumir.

2. Estudiar la estructura de la red formada por los conceptos y relaciones semánticas implícitos en el documento a resumir, así como las distintas posibilidades que dicha red proporciona a la hora de delimitar la información tratada en el texto. Para ello, se ha adoptado un enfoque basado en la representación del documento en forma de grafo, utilizando los conceptos o significados asociados a sus términos, y extendidos con distintas relaciones semánticas. A diferencia de otros trabajos que se centran en la agrupación de oraciones para determinar los temas comunes en múltiples documentos, y en la identificación de las oraciones centrales de cada grupo o *cluster*, en este trabajo el algoritmo de agrupamiento es aplicado a la identificación de conjuntos de conceptos estrechamente relacionados, que delimitan los distintos temas que se abordan en el texto, y cuya presencia en las distintas oraciones determina su grado de relevancia.

3. Implementar un sistema genérico que permita elaborar resúmenes de textos de diversos dominios, a la vez que permita utilizar conocimiento específico del dominio en cuestión para mejorar la calidad de los resúmenes generados. Dicho sistema habrá de ser fácilmente configurable y adaptable para trabajar sobre nuevos tipos de documentos sin más que modificar la base de conocimiento del dominio. Deberá, además, permitir la generación de resúmenes mono-documento y multi-documento.

4. Estudiar el efecto de la ambigüedad léxica sobre la generación automática de resúmenes, mediante la incorporación, dentro del sistema desarrollado, de distintos mecanismos de desambiguación. Valorar, de manera cuantitativa, los beneficios derivados de la resolución de ambigüedades como un paso previo a la generación del resumen.

5. Configurar y utilizar el sistema implementado para generar distintos tipos de resúmenes, tanto mono-documento como multi-documento, de

textos de distintos dominios. En concreto, se exponen en este trabajo el proceso desarrollado y los resultados alcanzados en la generación de resúmenes mono-documento de artículos científicos en biomedicina, la generación de resúmenes mono-documento de noticias periodísticas y la generación de resúmenes multi-documento de páginas web sobre destinos turísticos.

6. Evaluar los resúmenes generados utilizando las métricas comúnmente utilizadas en la tarea que nos ocupa, de cara a comparar dichos resúmenes con los elaborados por los seres humanos y con los generados por otros sistemas automáticos evaluados bajo las mismas condiciones experimentales.

1.4. Estructura del Documento

A continuación se describe brevemente cada uno de los capítulos que conforman este trabajo.

Capítulo 1. Introducción. Este capítulo presenta una revisión de la motivación y la problemática de la generación automática de resúmenes. Se enumeran, además, los objetivos concretos de este trabajo y se anticipa la estructura del mismo.

Capítulo 2. Trabajo Previo. Este capítulo expone 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. Asimismo, describe y analiza los distintos problemas que plantea la evaluación de resúmenes automáticos, a la vez que presenta las métricas de evaluación más utilizadas.

Capítulo 3. Herramientas y Recursos. Este capítulo describe las distintas herramientas y recursos que, no habiendo sido desarrollados expresamente para este proyecto, han sido utilizados en el mismo.

Capítulo 4. Uso de Grafos Semánticos para la Generación Automática de Resúmenes. Este capítulo presenta las distintas etapas en que se subdivide el método propuesto para la realización de resúmenes por extracción.

Capítulo 5. Caso de Estudio: Resúmenes Mono-documento de Artículos Científicos de Biomedicina. Este capítulo describe el proceso realizado para configurar el método descrito en el capítulo anterior para generar resúmenes mono-documento de artículos científicos en el dominio biomédico.

Capítulo 6. Caso de Estudio: Resúmenes Mono-documento de Noticias Periodísticas. Este capítulo describe el trabajo realizado para adaptar el sistema implementado para generar resúmenes mono-documento de noticias periodísticas.

Capítulo 7. Caso de Estudio: Resúmenes Multi-documento de Páginas Web de Información Turística. Este capítulo describe el trabajo realizado para especializar el método propuesto para generar resúmenes multi-documento de páginas web con información sobre destinos de interés turístico.

Capítulo 8. Evaluación. Este capítulo presenta la evaluación realizada con el objetivo de determinar la capacidad del sistema para realizar su cometido en los distintos dominios considerados en los casos de estudio.

Capítulo 9. Conclusiones y Trabajo Futuro. Este capítulo recopila las conclusiones extraídas del trabajo realizado, a la vez que propone distintas líneas de trabajo futuro.

Capítulo 2

Trabajo Previo

El presente capítulo tiene como objetivo presentar al lector la tarea de generar automáticamente resúmenes de texto. Para ello, se parte de la definición de resumen y de los factores que intervienen en su elaboración. A continuación, la Sección 2.2 presenta una revisión de los trabajos más destacados en generación de resúmenes, analizando la evolución de las técnicas empleadas desde los orígenes de esta disciplina hasta el momento actual, y destacando las fortalezas y debilidades de cada una de ellas. La Sección 2.3 analiza los distintos problemas que plantea la evaluación de resúmenes automáticos, a la vez que presenta las métricas de evaluación más utilizadas. Para concluir, en la Sección 2.4 se describen algunas de las más recientes aplicaciones o variantes de la tarea original de generar resúmenes automáticos.

2.1. El Resumen

Según Sparck-Jones (1999), un resumen consiste en la transformación de un texto mediante la reducción de su contenido, ya sea por selección o por generalización de lo que se considera 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 cumpla el resumen producido. Es decir, el contexto condiciona tanto el proceso como el resultado, siendo 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** determinan las características del texto a resumir, y se dividen 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 se refiere tanto a la organización explícita y generalmente marcada en el texto (diferentes secciones o apartados) como a la organización que se encuentra implícita en el discurso. La escala indica el 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 *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 éstos se relacionan entre sí.

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

- La **situación** pretende distinguir 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 discernir entre aquellos que se utilizan para ayudar al usuario a decidir si el texto es de su interés, aquellos que pretenden sustituir al documento original y aquellos cuyo objetivo es ofrecer una vista preliminar del texto antes de su lectura.

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

- El **material o extensión** determina el grado en que el resumen debe capturar el contenido presente en la fuente. Puede cubrir todo el contenido del texto original, o estar diseñado para cubrir únicamente un determinado tipo de información. A estos últimos se les denomina *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**, es posible distinguir 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. Con frecuencia, 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, y *resúmenes agregativos*, que incluyen información adicional no presente en el documento original con el objetivo de completar o matizar dicha información.

Partiendo del estudio de los factores anteriores, y dependiendo de las características del resumen consideradas, se pueden definir diferentes taxonomías de resúmenes. A continuación, y sin ánimo de exhaustividad, se presentan algunas de las más aceptadas.

- Una clasificación frecuente es aquella que distingue entre *resúmenes genéricos*, *resúmenes adaptados al usuario* y *resúmenes adaptados a una consulta*. Mientras que los primeros recogen los temas principales del documento y van destinados a un amplio y heterogéneo grupo de personas, los segundos se confeccionan teniendo en cuenta las preferencias y/o características del usuario a la hora de seleccionar los contenidos del resumen y la forma de presentación. Por su parte, los resúmenes adaptados a una consulta se construyen como respuesta a la consulta realizada por un usuario sobre un sistema de recuperación de información.

- 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 *mono-documento* y *multi-documento*. Estaremos ante un resumen mono-documento si en su elaboración todo el material procede de una única fuente o documento. Por el contrario, estaremos ante un resumen multi-documento si su contenido procede de distintos documentos que tratan sobre un mismo tema o suceso.

- 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.) presente en el texto de entrada, y aquellos que incluyen contenidos que no están presentes, al menos de manera explícita, en el documento original. Atendiendo pues a este criterio, se puede hablar de *resúmenes generados por extracción* y *resúmenes generados por abstracción*.

2.2. Técnicas de Generación Automática de Resúmenes

A la hora de clasificar el amplio abanico de técnicas utilizadas en generación automática de resúmenes se pueden adoptar principalmente dos enfoques (Mani, 2001):

- Distinguir entre técnicas que generan resúmenes mediante extracción y técnicas que generan resúmenes mediante abstracción, siendo por tanto el factor discriminante la existencia o no de un proceso de reescritura del resumen, utilizando técnicas de generación de lenguaje, a partir de una representación intermedia de la información contenida en el documento a resumir.

- Distinguir, en función de la profundidad del análisis acometido y del conocimiento empleado, entre *enfoques superficiales*, *enfoques basados en la estructura del discurso* y *enfoques en profundidad*.

En este trabajo se ha preferido adoptar la segunda clasificación, pues, en opinión de la autora, resulta difícil y confuso lograr un tratamiento homogéneo de la diversidad de enfoques utilizados en los sistemas de generación de resúmenes por extracción, que difieren considerablemente en el nivel del análisis realizado. Así pues, en el resto de la sección se estudiará un conjunto representativo de las técnicas de generación de resúmenes desde el nacimiento de la disciplina hasta la actualidad, distinguiendo en función de la profundidad del análisis lingüístico y semántico realizado. Añadiremos, no obstante, una cuarta categoría a las tres citadas anteriormente: los *enfoques basados en grafos*. Si bien pueden ser considerados un subconjunto de los enfoques discursivos, se ha optado por dedicarles una categoría exclusiva por dos razones: en primer lugar, porque claramente este trabajo se encuadra dentro de este grupo de técnicas y, por tanto, merecen un estudio independiente y exhaustivo; y en segundo lugar, porque la naturaleza de las relaciones utilizadas, así como las hipótesis y teorías subyacentes, generalmente difieren de las utilizadas en las técnicas basadas en la estructura del discurso.

Por otro lado, la práctica totalidad de los trabajos encuadrados en alguna de las dos primeras categorías (enfoques superficiales y enfoques discursivos) realizan resúmenes mediante extracción, mientras que sólo algunos de los denominados enfoques profundos utilizan técnicas de abstracción y reescritura del texto. Como ya se adelantó al inicio de este capítulo, las **técnicas de extracción** generan resúmenes compuestos íntegramente por material del documento original. Para ello, en una primera etapa o *fase de análisis*, se limitan a la extracción de segmentos clave del texto; posteriormente, 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. Por su parte, las **técnicas de abstracción** generan resúmenes que incluyen contenidos que no están presentes explícitamente en el texto de entrada. Durante la *fase de análisis*, construyen una representación semántica del texto fuente, mediante la identificación de conceptos genéricos y relaciones entre ellos, generalmente haciendo uso de alguna plantilla o esquema que 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 para reescribir el texto que conformará el resumen final. Su principal inconveniente es que se trata de técnicas únicamente aplicables a dominios muy acotados.

En función de los factores que intervienen en la construcción del resumen, revisados en la Sección 2.1, el paradigma más adecuado puede variar. A continuación, se realiza un repaso de las técnicas de generación de resúmenes que en la actualidad gozan de mayor aceptación, siguiendo la clasificación acordada. Para cada una de ellas, se citarán los trabajos más relevantes y los autores que más han contribuido a su desarrollo.

2.2.1. Enfoques Superficiales

Los primeros trabajos en generación de resúmenes datan de las décadas de los cincuenta y los sesenta, y vienen de la mano de dos autores, Luhn y Edmundson, de influencia decisiva en el desarrollo posterior de la disciplina (Luhn, 1958; Edmundson, 1969). Durante las siguientes décadas, y hasta finales de los noventa, el interés por la misma fue escaso. Sin embargo, y muy especialmente en los últimos años, la investigación en el área ha crecido significativamente.

Las primeras apuestas de esta segunda etapa abordan el problema mediante el uso de técnicas superficiales para identificar los segmentos relevantes en el documento fuente. El tamaño de estos segmentos o unidades varía de unas técnicas a otras. Aunque la mayoría trabajan a nivel de oración, algunas utilizan unidades más pequeñas, como sintagmas nominales o proposiciones, mientras que otras utilizan unidades mayores, como párrafos. Sin embargo, la arquitectura general difiere poco o nada de unos enfoques a otros. Hahn y Mani (2000) definen dos fases genéricas en la construcción del resumen: *análisis* y *síntesis*. La mayor parte del trabajo se realiza durante la fase de análisis, si bien dicho análisis continúa siendo 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 evalúan un conjunto de características para cada unidad, se normalizan y se suman. Durante la fase de síntesis, se extraen las unidades mejor puntuadas y se construye el resumen mediante la simple concatenación de las mismas. Algunos trabajos, además, realizan eliminación de redundancia y algún otro procesamiento para mejorar la coherencia del resumen. La

Figura 2.1 muestra la arquitectura genérica de un sistema "tradicional" de generación de resúmenes.

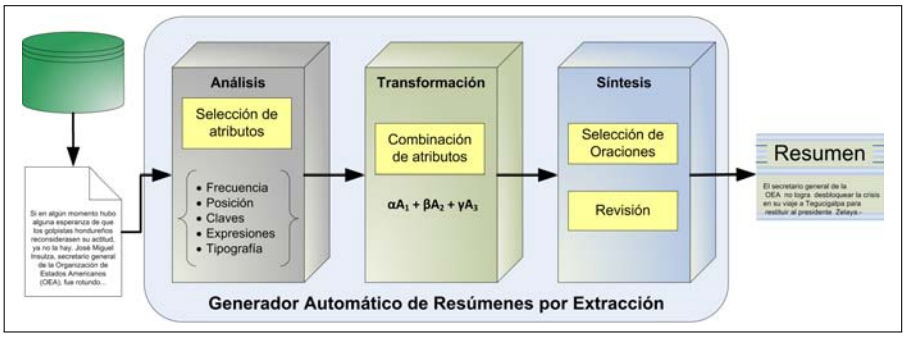


Figura 2.1: Arquitectura general de un sistema tradicional de generación de resúmenes por extracción

Por tanto, la principal diferencia entre unas técnicas y otras radica en las características ponderadas para calcular las puntuaciones a las oraciones. A continuación, se analizan en detalle las principales heurísticas utilizadas en la selección de las oraciones para el resumen (Paice, 1980).

2.2.1.1. Frecuencia de Palabras

Se trata de un método muy simple e intuitivo que consiste en buscar ocurrencias de palabras y expresiones que se refieran al tema o temas centrales del documento. El método más utilizado parte del texto completo, del que se eliminan 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 un cierto umbral son eliminadas de la lista. A continuación, 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 oración.

Los sistemas propuestos por Luhn (1958), Edmundson (1969), Kupiec *et al.* (1995), Hovy y Lin (1999) o Teufel y Moens (1997), entre otros, utilizan diferentes medidas para el cálculo de las frecuencias de los términos: la frecuencia del término (*term frequency, tf*) y la frecuencia del término multiplicada por la inversa de la frecuencia del documento (*term frequency* $\times$ *inverse document frequency, tf* $\times$ *idf*) (Sparck-Jones, 1972). Con respecto a la forma de puntuar la oración, algunos enfoques puntúan la aparición de grupos de palabras, 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 (Edmundson, 1969; Kupiec, Pedersen, y Chen, 1995; Teufel y Moens, 1997).

A pesar de la simplicidad y la antigüedad de esta heurística, a día de hoy sigue gozando de gran popularidad, si bien los trabajos actuales la utilizan en combinación con otras técnicas más sofisticadas encaminadas a mejorar la coherencia y/o la cohesión de los resúmenes generados. En este sentido, cabe mencionar el trabajo de Lloret *et al.* (2008), que presenta un sistema tradicional de frecuencia de términos mejorado con un módulo de implicación textual (*textual entailment*). Citando a Herrera *et al.* (2005), el termino "implicación textual" se utiliza para indicar la situación en la que la semántica de un texto en lenguaje natural se puede inferir de la semántica de otro texto. Lloret *et al.* (2008) aplican esta técnica como paso previo a la generación del resumen, con el objetivo de eliminar la información redundante. De entre las oraciones no redundantes, se seleccionan aquellas mejor ponderadas según el enfoque clásico de frecuencia de términos. Otro ejemplo de trabajo reciente lo encontramos en Tseng (2009). En él se propone un sistema de generación de resúmenes de noticias periodísticas especialmente diseñado para su uso en teléfonos móviles. Bajo la premisa de que los resúmenes destinados a estos dispositivos requieren un elevado ratio de compresión, el sistema selecciona las cinco oraciones más importantes de una noticia en función de las frecuencias de sus términos. A continuación, cada una de estas oraciones se combina con el titular de la noticia (que a priori se considera importante y que, por tanto, ha de formar parte del resumen) para generar resúmenes candidatos. Para finalmente decantarse por uno de los cinco resúmenes candidatos, se utiliza una función que pondera distintas características como la posición de la oración en la noticia o la similitud de sus primeras palabras con aquellas del titular.

2.2.1.2. Estructura del Documento

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. A continuación, se puntúan las oraciones en función de la presencia de estas palabras claves.

El primer trabajo en el que se aplica esta heurística lo encontramos en

(Edmundson, 1969). En él se asignan pesos a las palabras significativas del título, subtítulo y encabezados, y se calcula la puntuación final de la oración como la suma de los pesos de sus términos que se encuentran presentes en alguna de estas secciones "importantes". Teufel y Moens (1997), por su parte, puntúan cada oración dividiendo el número de palabras que también forman parte del título entre el total de términos de la oración.

De nuevo, se trata de una heurística muy utilizada en los trabajos más recientes. En concreto, es muy frecuente en sistemas de generación de resúmenes de conversaciones mantenidas a través de correos electrónicos o "hilos" (Carenini, Ng., y Zhou, 2008; Wan y McKeown, 2004). En ellos, es de suponer que la información codificada en el asunto del primer mensaje resuma por sí misma la información tratada a lo largo de la conversación, y por ello se utiliza para extraer palabras claves en función de las cuales evaluar la importancia de las distintas oraciones de la conversación. Bawakid *et al.* (2008) formulan una hipótesis similar en su sistema de resúmenes guiados por consultas. Su propuesta incluye el cálculo de la similitud con respecto al título de las distintas oraciones que componen el documento a resumir, como una de las características de la función de pesos ideada para calcular la importancia relativa de dichas oraciones. Pero a diferencia de los enfoques anteriores, la similitud entre las oraciones y el título se calcula desde una perspectiva semántica, como la suma ponderada de la similitud entre pares de palabras del título y las oraciones, utilizando la métrica de Jiang y Conrath (1997) como indicador de similitud semántica.

2.2.1.3. Localización

Baxendale (1958) fue el primero en observar que, dentro de un párrafo, la primera oración generalmente contiene la información más relacionada con el tema tratado en el párrafo. Partiendo de esta idea, posteriores trabajos han estudiado la estructura de los textos para determinar posiciones de las oraciones que indican, con una alta probabilidad, la presencia de información relevante (Kupiec, Pedersen, y Chen, 1995; Teufel y Moens, 1997), como pueden ser las oraciones bajo los encabezados de secciones como "Conclusión", las que ocupan los primeros y últimos párrafos del documento, etc. Sin embargo, las posiciones relevantes dependen en gran medida del tipo de documento considerado. Así, mientras que las noticias periodísticas suelen cumplir con relativa fidelidad la denominada *regla de la pirámide invertida*

(i.e. los datos de mayor interés se incluyen en primer lugar y, a continuación, se desarrollan aspectos secundarios), no ocurre así en otros tipos de documentos, incluso dentro del propio dominio periodístico, como la editorial o la crónica.

Al igual que ocurriera con las dos heurísticas ya estudiadas, la posición de las oraciones en el documento continúa siendo una heurística habitual en los sistemas actuales, si bien ha dejado de utilizarse como criterio único o predominante en la selección de oraciones para pasar a desempeñar un papel secundario, y simplemente matizar o completar las puntuaciones asignadas a las oraciones por otros criterios más sofisticados. Así, por ejemplo, Bossard *et al.* (2008) utilizan la posición de la oración como atributo en la función de pesos para el cálculo de la relevancia de las oraciones. En este trabajo, la puntuación asignada al criterio "localización" persigue reflejar cuan cerca se encuentra la oración del inicio del documento, y se calcula según la Ecuación 2.1, donde n_j indica la posición de la oración O_j en el documento.

$$Posición(O_j) = \sqrt[2]{\frac{1}{n_j}} \tag{2.1}$$

Por el contrario, otros muchos trabajos defienden que las oraciones situadas en los últimos párrafos del documento contienen información tanto o más importante que la presente en los primeros párrafos, pues a menudo constituyen una recopilación de lo expuesto a lo largo del documento (Bawakid y Oussalah, 2008).

2.2.1.4. Palabras o Expresiones Indicadoras

Algunas palabras o sintagmas de una oración, aunque no sean en sí mismas palabras clave, aportan ciertas pistas sobre si la oración trata con información relevante. Los trabajos de Edmundson (1969) y Rush *et al.* (1971) son los ejemplos más antiguos en la aplicación 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, indican contenido importante del texto, mientras que las palabras stigma indican contenido irrelevante. Rush (1971) utiliza un método muy similar en el que considera también expresiones cortas. Kupiec (1995) también utiliza grupos de palabras en lugar de palabras individuales (por ejemplo, "en conclusión").

Resulta fácil darse cuenta de que esta heurística depende en gran medida del dominio al que pertenecen los documentos a resumir, y que las palabras clave aprendidas o definidas para textos de un determinado dominio difícilmente podrán ser utilizadas satisfactoriamente en otro dominio diferente. Esta limitación es el principal motivo por el que el uso de esta heurística en los sistemas actuales de generación de resúmenes es muy limitado. Sin embargo, es posible encontrar algunos trabajos que comparten la misma idea o hipótesis subyacente (la existencia de palabras o grupos de palabras que definen lo que es importante para cierta categoría o tipo de documentos). Aker y Gaizauskas (2010), por ejemplo, defienden que los seres humanos construyen un modelo conceptual sobre lo que es importante respecto a cierto objeto, y que dicho modelo influye en sus decisiones a la hora de determinar la información importante que lo describe. Siguiendo esta premisa, definen manualmente un conjunto de categorías de información que se consideran importantes a la hora de describir determinados monumentos o lugares de interés turístico (e.g. al describir una iglesia o un puente, el lugar donde se localizan o el año en que fueron construidos constituyen dos categorías de información relevante). Una vez conocido el modelo conceptual, utilizan patrones de dependencias sintácticas para determinar qué oraciones en el documento a resumir contemplan la información relativa a cada una de las categorías definidas.

2.2.1.5. Uso de Aprendizaje Automático en la Selección de Oraciones

Los enfoques actuales utilizan técnicas más sofisticadas para estimar la importancia de las oraciones y decidir cuáles de ellas seleccionar para el resumen. En particular, durante los últimos años el uso de algoritmos de Aprendizaje Automático (AA) 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 de entrenamiento junto con sus resúmenes generados de forma manual, que permitan aprender automáticamente los pesos de los distintos atributos. Evidentemente, esta técnica es muy dependiente del corpus.

La Figura 2.2 muestra la arquitectura general de un sistema de este tipo. La aplicación de aprendizaje automático a la generación de resúmenes se aborda por primera vez en Kupiec *et al.* (1995), donde se utiliza un clasi-

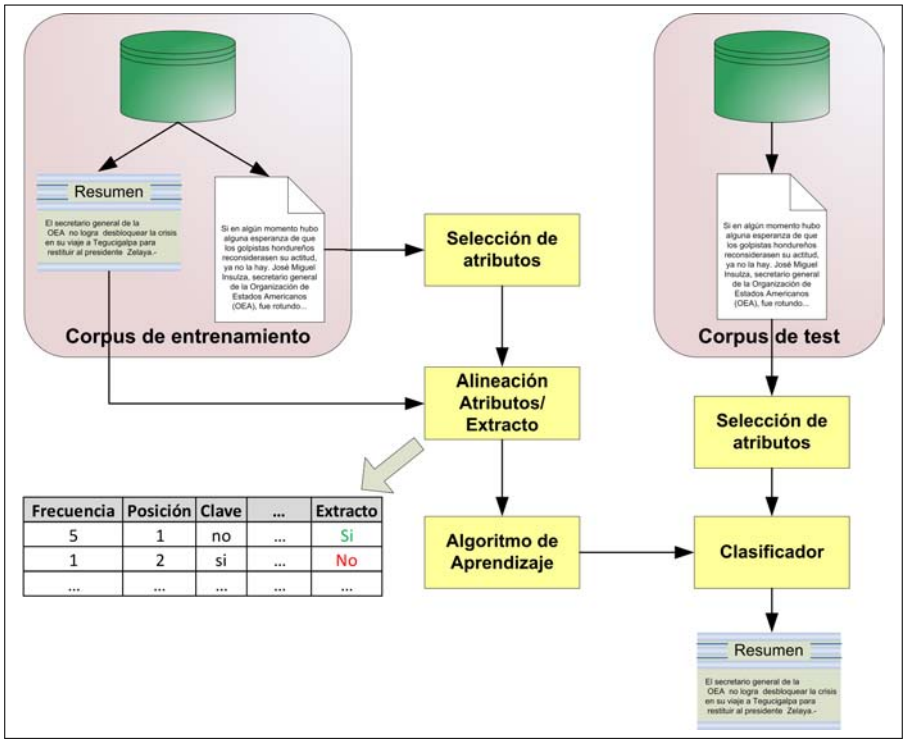


Figura 2.2: Arquitectura de un sistema de aprendizaje automático para la generación de resúmenes

ficador bayesiano para determinar la combinación óptima de los atributos 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. Para el entrenamiento, Kupiec *et al.* utilizan un corpus de 188 documentos, acompañados de sus respectivos resúmenes, y procedentes de 21 colecciones científicas. El algoritmo toma cada una de las oraciones y calcula la probabilidad con la que debería ser incluido en el resumen automático, en base a su similitud con respecto al resumen manual. Los resultados corroboran aquellos obtenidos por Edmundson (1969), y parecen indicar que la mejor característica es la posición de las oraciones.

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 central de cada documento. Por su parte, 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 del lenguaje, se denominan *relaciones retóricas*. En Neto *et al.* (2002) se 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.

Más reciente e innovador es el trabajo presentado en Metzler y Kanungo (2008), donde el objetivo es seleccionar oraciones para construir un resumen destinado a responder a la pregunta o consulta planteada por un usuario. Para ello, entrenan y comparan dos clasificadores: un modelo de regresión y un árbol de decisión. El conjunto de atributos utilizado consta de dos tipos o categorías de información: atributos relacionados o dependientes de la consulta, y atributos independientes de la consulta. Entre los primeros se encuentran el número de palabras de la oración que aparecen en la consulta o la probabilidad de que la consulta haya sido generada a partir de la oración. Entre los atributos que no dependen de la consulta, se encuentran la longitud y la posición de la oración.

Para terminar el recorrido por los sistemas que utilizan técnicas de aprendizaje automático, citaremos el trabajo de Binwahlan *et al.* (2009), en el que se analiza el efecto de la estructura de los atributos seleccionados sobre la calidad de los resúmenes generados, mediante optimización de enjambre de partículas (*particle swarm optimization*). Las características analizadas son la "centralidad" de la oración con respecto al resto de oraciones del documento, la similitud con el título y con la primera oración, y la presencia de alguna de las diez palabras más frecuentes del documento.

2.2.1.6. Ventajas e Inconvenientes de los Métodos Superficiales

Las técnicas estudiadas hasta este punto presentan la ventaja de su relativa sencillez y su bajo coste, ya que apenas hacen uso de conocimiento o de complejas técnicas de procesamiento lingüístico. Además, son relativamente independientes del dominio, lo que sin duda constituye un argumento a su favor. Sin embargo, no resultan apropiadas para todos los tipos de resúmenes. Si se trata, por ejemplo, de resumir textos muy extensos, el ratio de comprensión que se necesita es muy elevado, y resulta imposible de alcanzar

sin utilizar cierto grado de abstracción. Además, los resúmenes generados frecuentemente adolecen de falta de cohesión y coherencia.

Por **cohesión** se entiende la unión o conexión existente entre las cosas. En nuestro caso particular, se refiere a la característica de todo texto bien formado, consistente en que las diferentes oraciones se conectan entre sí mediante diversos procedimientos lingüísticos que permiten que cada oración sea interpretada en relación con las demás. En otras palabras, si al elaborar el resumen resulta seleccionada una oración en la que se referencia al contenido de una oración anterior que no ha sido incluida en el resumen, éste resultará incomprensible, o cuanto menos confuso, para el lector. La Figura 2.3 ilustra un problema de cohesión frecuente en los resúmenes generados por los enfoques superficiales. En el resumen generado, se observa cómo tanto la primera como la segunda oración se refieren a hechos y conceptos introducidos en las restantes oraciones no incluidas en el resumen.

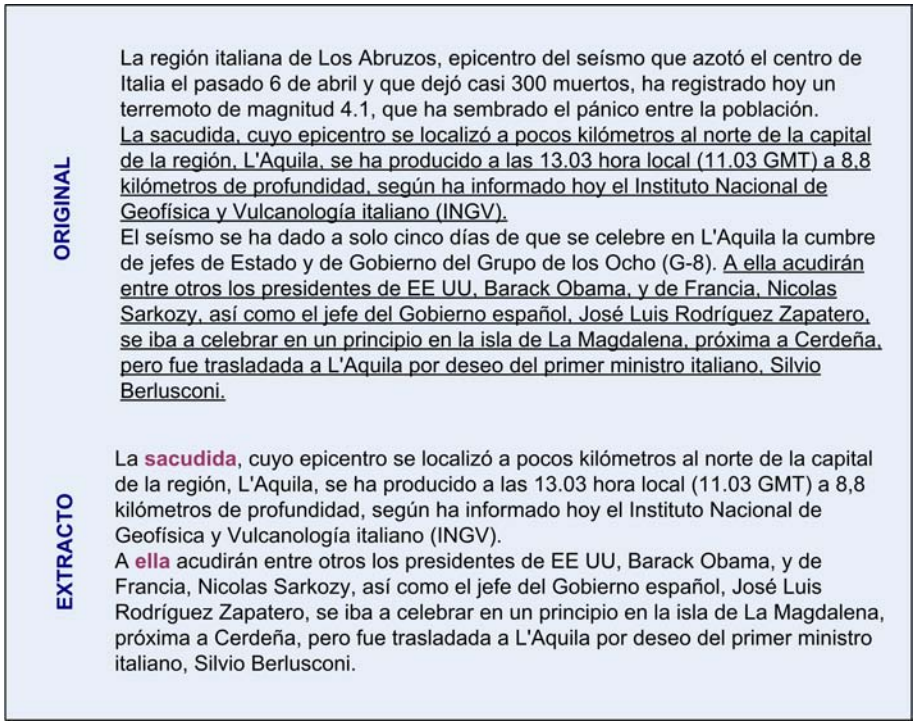


Figura 2.3: Problemas de cohesión en los resúmenes automáticos

La **coherencia** es una propiedad de los textos bien formados que permite concebirlos como entidades unitarias, de manera que las ideas secundarias aportan información relevante para llegar a la idea principal y comprender

el significado global del texto. Tiene que ver, pues, con la unidad temática del texto, con su estructura y con la organización lógica de las ideas (i.e. que las distintas partes mantengan relaciones de significado y que haya una adecuada progresión temática). De nuevo, podemos ver el problema a través de un ejemplo en la Figura 2.4. En el extracto generado, resulta imposible discernir a qué empresa (CNPC o YPF) se refiere el término "compañía".

Figura 2.4: Problemas de coherencia en los resúmenes automáticos

Sin embargo, y a pesar de estas limitaciones, distintos estudios empíricos justifican el uso de métodos de extracción. Kupiec (1995) realiza una evaluación con la que demuestra que aproximadamente el 80 % de las frases incluidas en los resúmenes manuales aparecen tal cual o con pequeñas modificaciones en el texto original. Por su parte, Morris *et al.* (1992) llevan 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 la información incluida 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 operaciones de generalización y abstracción.

2.2.2. Enfoques Basados 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 acerca 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. Este modelo es precisamente lo que las técnicas discursivas aspiran a capturar.

Dentro de los métodos basados en la estructura del discurso, es posible distinguir, a su vez, dos grupos de técnicas: las que analizan la cohesión del documento y las que se concentran en el análisis de su coherencia.

2.2.2.1. Análisis de la Cohesión

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, significados 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.

A modo de ejemplo, el trabajo de Hearst (1997) compara bloques de texto adyacentes, en función del solapamiento de vocabulario, para identificar 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 tema del texto; aunque una definición más completa debería incluir una enumeración de las relaciones entre palabras que se consideran a la hora de construir dichas cadenas. En Barzilay y Elhadad (1997), se presenta una solución que, sin hacer uso de una interpretación semántica compleja, produce un resumen identificando en el texto fuente secuencias de términos agrupados mediante distintas relaciones de cohesión textual: repetición, sinonimia, hiperonimia, antonimia y holonimia. Para establecer las relaciones, se utiliza la base de conocimiento léxico WordNet, tratando el problema de la polisemia mediante la creación de cadenas alternativas para los distintos significados posibles y la elección

de la mejor cadena en función del número de relaciones y sus pesos. 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 el mismo significado en WordNet.

La aproximación más aceptada para la representación de la cohesión textual son los llamados *grafos de cohesión*. Como ya se ha mencionado, dentro de un documento, las palabras y oraciones se encuentran conectadas entre sí por medio de distintos tipos de relaciones. Estas relaciones se pueden representar en una estructura de grafo, en el que los vértices son los distintos elementos textuales (típicamente oraciones) y los arcos representan las relaciones entre ellos. 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 representan 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 entre las palabras de los párrafos.

Son muchos los trabajos recientes que utilizan las técnicas descritas para capturar las relaciones implícitas entre las unidades textuales y mantener así la cohesión del resumen generado. En este sentido, cabe mencionar el trabajo de Reeve, Han y Brooks (2007), que supone una adaptación del método de las cadenas léxicas para generar resúmenes de documentos biomédicos. Para ello, en lugar de trabajar con términos, el texto se traduce a conceptos del UMLS (*Unified Medical Language System*, ver Sección 3.1.3) que posteriormente se encadenan entre sí, de tal forma que cada cadena constituye una lista de conceptos que pertenecen al mismo tipo semántico en UMLS. Cada cadena se puntúa multiplicando la frecuencia con la que su concepto más frecuente aparece en el documento por el número de conceptos que componen la cadena. Las puntuaciones obtenidas se utilizan para determinar la cadena de conceptos más "fuerte" (i.e. aquella que, con una mayor probabilidad, representa el tema principal del documento). González

y Fuentes (2009) de nuevo presentan una adaptación del método de las cadenas léxicas. Su aportación consiste en considerar nuevas relaciones para medir la cohesión interna de las cadenas: la relación "extra-fuerte", la relación "fuerte" y la relación "media-fuerte". A la hora de puntuar las cadenas, se tienen en cuenta distintas heurísticas, como su longitud, su posición de inicio en el documento o el tipo de relaciones que enlazan a las palabras que la componen. En función de su puntuación, las cadenas se clasifican en uno de los tres tipos siguientes: "fuerte", "media" y "débil". Para construir el resumen, las oraciones se seleccionan utilizando las relaciones fuertes pero, en caso de no ser suficientes para alcanzar el ratio de compresión deseado, se acude al resto de relaciones.

Una modalidad de sistemas en los que el concepto de cohesión textual juega un papel destacado son aquellos que generan *resúmenes guiados o conducidos por eventos*. Estos sistemas seleccionan las oraciones para el resumen en función de los eventos que se describen en ellas y de las distintas relaciones que existen entre dichos eventos. No obstante, el tipo de relaciones consideradas varía de unos enfoques a otros. Así, por ejemplo, Liu *et al.* (2007) utilizan cinco tipos de relaciones entre verbos, extraídas de la ontología VerbOcean (Chklovski y Pantel, 2004), que se corresponden con los eventos descritos en un conjunto de documentos. Por el contrario, en Li *et al.* (2006), las relaciones entre eventos se definen como una función de la similitud semántica entre ellos y de la frecuencia en que ocurren unidos.

Finalmente, cabe encuadrar en esta categoría a todos aquellos sistemas que, si bien utilizan técnicas superficiales para determinar la importancia relativa de las oraciones del documento (e.g. palabras clave o frecuencia de términos), incluyen un módulo destinado a identificar y resolver las posibles anáforas o referencias pronominales presentes en el mismo. Tal es el caso del trabajo presentado por Steinberg *et al.* (2007), donde se propone un método para la resolución de anáforas basado en la teoría del *Análisis de la Semántica Latente* (*Latent Semantic Analysis, LSA*) y se aplica para mejorar la calidad de los resúmenes generados por un sistema sencillo basado en las frecuencias de los términos del documento.

2.2.2.2. Análisis de la Coherencia

Halliday (1985), Mann y Thompson (1988) y Van Dijk (1988) coinciden en afirmar que la coherencia textual representa la estructura general o superes-

tructura de un texto, visto éste como un conjunto de oraciones, y en términos de las relaciones de alto nivel que se establecen entre ellas. Han sido muchas las teorías propuestas para el análisis de la estructura argumentativa de un texto: la *Teoría de la Estructura Retórica* (Mann y Thompson, 1988), las *Gramáticas Discursivas* (Longacre, 1979), las *Macro-estructuras* (Van Dijk, 1988) o las *Relaciones de Coherencia* (Hobbs, 1985). En nuestra exposición, no obstante, nos centraremos en la teoría de Mann y Thompson, por ser la de mayor difusión y aplicación en generación de resúmenes.

La Teoría de la Estructura Retórica *(Rhetorical 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 de computación lingüística, y 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 es la definición del concepto de relación retórica para referirse a un tipo de relación asimétrica que se establece entre dos segmentos de texto a los que se denomina, respectivamente, *núcleo* y *satélite*. El núcleo contiene información que es "central" en el documento, mientras que el satélite aporta información que completa o complementa al núcleo. Las relaciones en la RST (en concreto, en el trabajo original se presentan 25 relaciones diferentes) 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; y destacan por su importancia las relaciones de circunstancia, motivación, propósito y solución. Una vez definidas las relaciones y los segmentos, el texto se representa como un árbol, en el que los nodos internos representan las relaciones y los nodos hoja representan los segmentos. La estructura arbórea definida permite calcular la importancia de cada uno de los segmentos textuales, partiendo de los nodos hoja y propagando los resultados hasta la raíz. La Figura 2.5 muestra, a modo de ejemplo, el árbol retórico construido para un texto.

Los principios de esta teoría han sido aplicados en numerosos sistemas de generación de resúmenes, muchos de los cuales no son sino meras variantes de la teoría original. Así, por ejemplo, Marcu (1999; 2000) se inspira en esta teoría para construir un árbol que representa la estructura retórica del texto, y posteriormente utiliza este árbol para calcular la relevancia de los

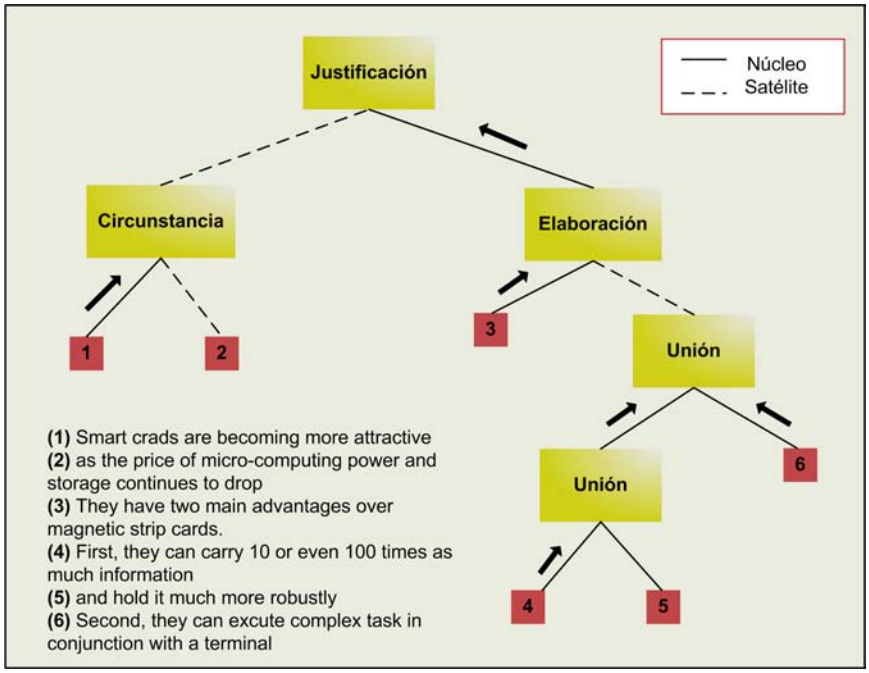


Figura 2.5: Ejemplo de árbol de estructura retórica (Marcu, 1997)

términos que actúan como nodos del árbol y que permiten la composición de resúmenes a distintos niveles de detalle. Radev *et al.* (2000; 2002) desarrollan una teoría similar, la llamada *Teoría de la Estructura Inter-Documento* (*Cross-Document Structure Theory, CST*), cuya principal diferencia respecto a la RST subyace en las relaciones consideradas, que se establecen entre distintos documentos, en lugar de dentro de un mismo documento. Está pues diseñada para su aplicación en generación de resúmenes multi-documento. Barzilay y Lapata (2008) proponen un novedoso enfoque para representar y medir la coherencia textual. Mediante la representación del discurso a través de una cuadrícula o parrilla de entidades, las autoras defienden que es posible capturar patrones de transición entre oraciones, lo que a su vez les permite, utilizando técnicas de aprendizaje automático, aprender las propiedades que debe cumplir un texto coherente. El enfoque propuesto es aplicado a distintas tareas, entre ellas, la ordenación de oraciones y la evaluación de resúmenes automáticos.

Concluimos el repaso de las técnicas basadas en la coherencia textual con el trabajo de Ouyang *et al.* (2009) sobre generación de resúmenes multi-documento guiados por una consulta. La principal aportación de este trabajo

es la creación de una representación jerárquica de las palabras presentes en el conjunto de documentos, bajo la hipótesis de que esta relación general-específico puede explotarse con el fin de mejorar la diversidad y la coherencia de los resúmenes generados. Una vez construida la jerarquía, el siguiente paso consiste en determinar qué palabras de este árbol son las importantes, utilizando la noción de Subsunción de Información (*Information Subsumption, IS*), y que resultan ser aquellas que, sin llegar a un nivel excesivo de generalización, cuentan con un mayor número de descendientes en la jerarquía. La selección final de las oraciones para el resumen se realiza según su cobertura de estas palabras que previamente se han reconocido como las más importantes en el conjunto de documentos.

2.2.3. Enfoques Basados en Grafos

El propósito de esta sección 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.

Las técnicas de modelado de texto basadas en grafos se asientan, en su mayoría, 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 Königsberg, 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. Para 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 (smallworld 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* o nodos concentradores) con un gran número de conexiones, mientras que el resto de nodos están muy poco conectados entre sí.

Durante los últimos años, se han multiplicado los trabajos en los que los principios de la teoría de grafos se aplican para 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 léxica, 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. Así, por ejemplo, en Lin (1998) y Lee (1999), el problema de la desambiguación léxica 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 *et al.* (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. 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. Agirre y Soroa (2009) utilizan WordNet y una variante del algoritmo PageRank para resolver la ambigüedad léxica de las palabras de un texto de forma no supervisada. Para ello, primero crean un grafo que representa la jerarquía completa de WordNet. Posteriormente, las palabras ambiguas en el documento se añaden como nodos a este grafo, estableciendo enlaces dirigidos desde ellas hasta cada uno de sus posibles significados o *senses* en WordNet. Tras asignar pesos a los nodos, se aplica PageRank para propagar dicha información a través del grafo. Finalmente, se escoge como significado de cada palabra ambigua aquel que resulte representado por el nodo con mayor peso.

En cuanto a generación automática de resúmenes se refiere, en las secciones anteriores ya han sido revisados 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 basadas 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. En esta sección, nos centraremos en el estudio de diferentes métodos que utilizan el concepto de *centralidad (centrality)* para capturar las oraciones "centrales" en un documento o conjunto 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, relaciones que a su vez pueden ser de naturaleza léxica, sintáctica o semántica.

Muchos de los métodos para el cálculo de la centralidad están basados en *PageRank* (Brin y Page, 1998), el algoritmo 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 a 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.

En esta línea, Erkan y Radev (2004b) presentan *LexRank*, uno de los métodos más aceptados para calcular la centralidad en un grafo, aplicado a la generación automática de resúmenes multi-documento. LexRank construye un grafo para el conjunto de documentos a resumir en el que existe un vértice por cada oración del mismo. Para determinar los enlaces entre los vértices, las oraciones se representan por sus vectores de frecuencias ($tf \times idf$), y se calcula la similitud léxica entre ellos utilizando la métrica del coseno, obteniendo así una matriz de similitudes. Aquellos pares de oraciones que presenten una similitud superior a un determinado umbral se enlazan entre sí en el grafo. Partiendo de la hipótesis de que las oraciones que son similares a muchas otras son las más importantes en relación al tema central del

documento, la extracción de oraciones relevantes consiste en identificar las oraciones que actúan como centroides en el grafo. 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 interpreta como 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. A esta variante del algoritmo original, los autores la denominan *LexPageRank* (Erkan y Radev, 2004a).

Un algoritmo similar es *TextRank* (Mihalcea y Tarau, 2004), utilizado para la generación de resúmenes mono-documento, aunque también ha sido aplicado a otras tareas de procesamiento de lenguaje, como la extracción de palabras clave. TextRank 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. 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. En cualquier caso, ambos algoritmos presentan resultados prometedores sin necesidad de realizar un análisis en profundidad del texto.

Un trabajo interesante, y en el que se inspira esta investigación, se presenta en Yoo *et al.* (2007). En él se propone un método para la agrupación de documentos con contenidos similares, concebido como un paso previo a la generación de resúmenes multi-documento en el dominio biomédico. Las tres primeras fases del algoritmo realizan el agrupamiento de los documentos, mientras que en las tres últimas se genera un resumen para cada grupo de documentos. A continuación, se describen todas estas fases:

1. **Representación de la colección de documentos como grafos de descriptores MeSH**. Partiendo de una colección de documen-

tos biomédicos, algunos de los cuales tratan sobre un mismo tema, el primer paso consiste en la traducción de cada uno de los documentos a una estructura de árbol, en la que los nodos corresponden a los descriptores de la terminología MeSH (ver Sección 3.1.2) identificados en el documento y sus hiperónimos, y las aristas corresponden a las relaciones *is a* existentes entre descriptores. El árbol se completa añadiendo nuevas aristas que representan las relaciones de co-ocurrencia de los descriptores entre los documentos del corpus. Finalmente, una vez obtenido el grafo que representa a cada documento, los grafos se fusionan para obtener una representación única de toda la colección.

2. **Agrupamiento del corpus de documentos**. En esta segunda fase, se aplica sobre la representación gráfica de la colección de documentos un algoritmo de agrupamiento, generando como resultado distintos modelos de agrupación de documentos. Cada modelo captura las principales relaciones existentes entre conjuntos de documentos, en términos de los descriptores clave que contienen.

3. **Asignación de documentos a los modelos**. A continuación, cada uno de los documentos de la colección se asigna a uno de los modelos de documentos generados en la etapa anterior. La elección del modelo al que se asigna un documento se realiza en función del número de descriptores en común y de la centralidad de estos descriptores en el modelo.

4. **Representación de las oraciones como grafos**. Al igual que en la primera etapa, y para cada conjunto de documentos, cada oración se representa como un grafo. Para obtener este grafo, y a diferencia de como ocurriera con la representación de los documentos, los descriptores se extienden utilizando las relaciones del modelo de agrupamiento correspondiente, en lugar de utilizar la jerarquía completa de MeSH.

5. **Construcción de la red de interacción semántica textual (*TSIN, Text Semantic Interaction Network*)**. En esta etapa, y para cada conjunto de documentos, las relaciones entre oraciones se representan utilizando una red de interacción semántica, en la que los vértices son oraciones y las aristas indican similitud semántica entre ellas, calculada como la distancia de edición (*edit distance*) entre sus representaciones gráficas.

6. **Selección de contenidos para el resumen**. Finalmente, se seleccionan las oraciones significativas para el resumen de cada conjunto de documentos, que son aquellas representadas por los nodos centrales de la red de interacción semántica.

En Wan *et al.* (2007) se propone un algoritmo basado en grafos capaz de resolver simultáneamente dos tareas diferentes: extracción de palabras claves y generación de resúmenes, bajo la hipótesis de que ambas tareas persiguen un objetivo similar y que, además, pueden reforzarse mutuamente. Para ello, definen tres tipos de relaciones:

- Relaciones oración-oración, calculadas en función de la similitud semántica de sus respectivos contenidos.

- Relaciones palabra-palabra, que expresan distintos tipos de relaciones homogéneas entre palabras.

- Relaciones palabra-oración, que refleja la importancia de la palabra en la oración.

A continuación, construyen tres grafos distintos que reflejan, respectivamente, los tres tipos de relaciones anteriores, e iterativamente, calculan la importancia o prestigio de las palabras y las oraciones en dichos grafos. Finalmente, las oraciones con mayor prestigio se seleccionan para el resumen, mientras que las palabras con mayor prestigio se combinan para producir palabras clave.

Filippova *et al.* (2009) presentan un sistema de resúmenes multi-documento que, partiendo de una colección de noticias financieras, extrae las oraciones que contienen información, no sólo relevante sino también novedosa, correspondiente a la compañía sobre la que versan las noticias. Para ello, construyen un grafo de la colección de noticias en la que cada nodo representa una oración, y cada arista representa la similitud entre las oraciones enlazadas en términos de la métrica del coseno calculada sobre sus vectores $tf \times idf$. Aplicando PageRank sobre el grafo generado, extraen las oraciones relevantes, para, en una etapa posterior, seleccionar de entre estas aquellas que expresan información novedosa sobre la compañía u organización en cuestión.

2.2.4. Enfoques en Profundidad

Como ya se adelantara, las técnicas o enfoques en profundidad generalmente realizan resúmenes mediante abstracción. Abstraer implica realizar inferencias sobre el contenido del texto, e incluso hacer referencia a conceptos previos o a un conocimiento que se presupone. De este modo, es posible conseguir un mayor grado de compresión en el resumen, lo que resulta especialmente interesante a la hora de resumir documentos muy largos o para realizar resúmenes multi-documento. Es posible distinguir tres etapas en el proceso de generación de un resumen por abstracción:

1. Construcción de una representación semántica de las oraciones del documento.
2. Realización de operaciones de selección, agregación y generalización sobre estas representaciones.
3. Traducción del resultado a lenguaje natural.

Dentro de las investigaciones que utilizan técnicas de abstracción se pueden distinguir dos líneas claramente diferenciadas: aquellas que utilizan **extracción de información** y las que utilizan **compresión**. En el resto de la sección se presentan brevemente los principios sobre los que se asientan ambos enfoques, así como los trabajos de mayor transcendencia en cada uno de ellos. El lector podrá observar cómo la mayoría de estos trabajos datan de las décadas de los 80 y los 90. Ello es debido a que, durante los últimos años, el interés y el esfuerzo de la comunidad científica se ha ido progresivamente trasladando hacia las técnicas de generación de resúmenes mediante extracción, en detrimento de las técnicas de abstracción. El motivo fundamental radica en el todavía insuficiente estado de desarrollo de las técnicas de representación de conocimiento y de generación de lenguaje natural, que reduce la aplicabilidad de estos enfoques a dominios muy restringidos.

2.2.4.1. Abstracción Utilizando Técnicas de Extracción de Información

Los enfoques basados en extracción de información recorren el texto en busca de un conjunto de información predefinida para incluir en el resumen. Por ello, a pesar de tratarse de enfoques capaces de producir resúmenes de alta calidad, su validez se restringe únicamente a dominios muy concretos.

Una técnica muy popular consiste en la utilización de *plantillas* (*templates*), que recogen la información que se considera relevante para una cierta categoría 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 lo encontramos en el sistema *FRUMP* (DeJong, 1982), aplicado sobre artículos periodísticos de cincuenta dominios diferentes, y que define una plantilla o guión distinto para cada tipo de artículo. Por su parte, Radev y McKeown (1998) utilizan plantillas para obtener resúmenes multi-documento de artículos periodísticos sobre ataques terroristas.

Algunos enfoques combinan la extracción de plantillas con técnicas de análisis estadístico. Paice y Jones (1993) combinan técnicas de indexado con técnicas de abstracción, y utilizan estructuras semánticas para organizar el contenido de documentos de investigación en el dominio de los cultivos agrícolas (Figura 2.6). Este tipo de documentos se caracterizan por ser muy estructurados, y en ellos se pueden observar una organización, estilismo y semántica relativamente constante. Trabajos posteriores utilizan métodos más sofisticados para rellenar las plantillas, en conjunción con análisis estadístico y aprendizaje automático de patrones a partir de colecciones de textos anotados (Riloff, 1996).

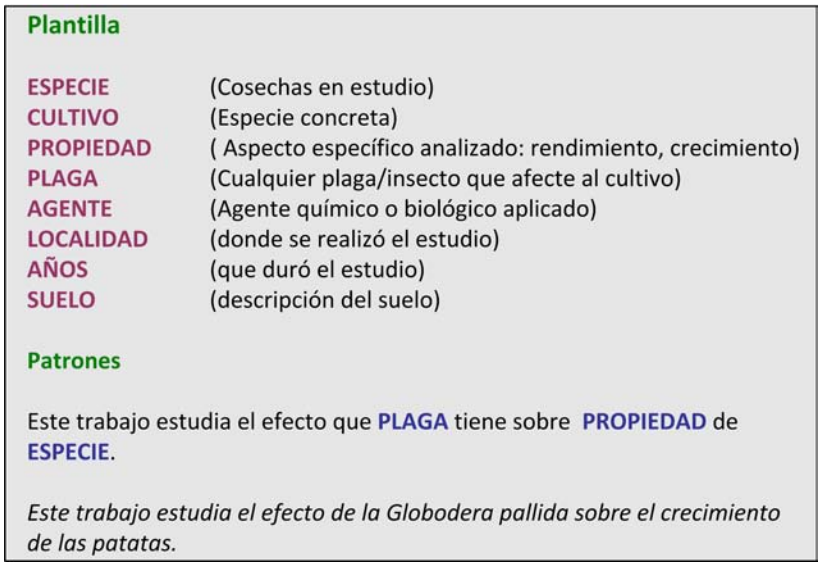


Figura 2.6: Ejemplo de plantilla de extracción (Paice y Jones, 1993)

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), por ejemplo, utilizan en su sistema *TOPIC* los conceptos de relevancia y generalización para crear una estructura jerárquica o grafo del texto. En este grafo, los nodos hoja se corresponden con los 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 a la medida del dominio de los textos, algunos enfoques hacen uso de tesauros de propósito general y uso público como WordNet.

Hovy y Lin (1999) también utilizan WordNet, 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.

2.2.4.2. Abstracción Utilizando Técnicas de Compresión de Texto

Los enfoques basados en compresión abordan el problema desde el punto de vista de la generación de lenguaje, y realizan operaciones de selección, agregación y generalización para 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 palabras, éstas son individual o colectivamente seleccionadas, agregadas o generalizadas, para producir resúmenes. El sistema SUSY (Fum, Gmda, y Tasso, 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 treinta conceptos. Cada oración se representa 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 previamente establecido. Es lo que se denomina *Suposición de Conceptos Altamente Referenciados* (*Highly Referenced Concept Assumption, HRCA*). Witbrock y Mittal (1999) extraen del documento original un conjunto de palabras que luego ordenan formando oraciones utilizando un modelo de lenguaje basado en bi-gramas. 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 et al., 1998; 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.

2.3. Evaluación de Resúmenes Automáticos

La presente sección pretende introducir al lector una cuestión de interés primordial en el desarrollo de sistemas automáticos de resúmenes, como es la evaluación de los resúmenes generados, entendida ésta como la medida de la calidad del sistema y de su potencialidad de uso en un entorno real.

La sección comienza con una breve reflexión sobre la necesidad de esta evaluación, y los problemas y dificultades que plantea. Seguidamente, se presenta una clasificación de los métodos o formas genéricas de evaluar un resumen, ya sea automático o manual, en función de las distintas perspectivas u ópticas bajo las que es posible valorar su calidad. La sección finaliza con la descripción de las métricas de mayor difusión.

2.3.1. Introducción a la Evaluación de Resúmenes

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 pretendido del resumen.

2.3.2. Clasificación de los Métodos de Evaluación de Resúmenes Automáticos

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 de evaluación intrínseca** se basan en el análisis directo del resumen producido para juzgar su calidad. En este tipo de evaluación, se pueden tener en cuenta tanto criterios gramaticales, como la cohesión y coherencia del texto generado, como criterios de cobertura informativa, que analizan la presencia u omisión en el resumen de los temas principales del texto de entrada. Para evaluar el grado de cobertura, lo habitual es recurrir a resúmenes realizados manualmente por expertos y compararlos con los

generados de manera automática, lo que conlleva, como ya hemos adelantado, un elevado coste en tiempo y recursos, más aún teniendo en cuenta que la evaluación del sistema ha de realizarse sobre una colección de textos suficientemente extensa como para que los resultados obtenidos sean estadísticamente significativos. Para evaluar la corrección gramatical también es frecuente solicitar a jueces humanos una valoración de los resúmenes, durante la cual pueden producirse discrepancias entre los jueces que invaliden la evaluación.

Los **métodos de evaluación extrínseca** estudian el resumen en el contexto de la tarea para la que ha sido generado. La mayoría pretenden estimar la repercusión de utilizar resúmenes automáticos en lugar de los textos completos. Evalúan, por ejemplo, si un resumen automático permite clasificar correctamente un documento o responder a ciertas preguntas sobre el texto de entrada con la misma precisión que si se utilizase el documento original. 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.

2.3.3. Métricas de Evaluación de Resúmenes Automáticos

A continuación se describen algunas de las métricas más comúnmente utilizadas en la evaluación de sistemas de generación de resúmenes. De entre la relativamente amplia variedad de métricas propuestas, resulta interesante distinguir, por un lado, las que persiguen evaluar la cobertura informativa del resumen de las que, además, aspiran a valorar su legibilidad y corrección gramatical; y por otro lado, aquellas que han sido diseñadas para su cálculo automático de las que requieren la participación de jueces humanos.

2.3.3.1. Precisión y Cobertura

Una primera aproximación al problema consiste en utilizar métricas tradicionales de recuperación de información para medir la calidad de la cobertura informativa de los resúmenes automáticos, en comparación con otros redactados manualmente (Salton y McGill, 1983). 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. Como principales ventajas, presentan

su facilidad, rapidez de cómputo, y la no intervención de seres humanos.

Considérense las posibles combinaciones respecto a la coincidencia de oraciones entre el resumen generado automáticamente y el redactado de manera manual. Si una oración se selecciona en ambos resúmenes o en ninguno de ellos, estaremos ante un *verdadero positivo (true positive, TP)* o un *verdadero negativo (true negative, TN)*. Por el contrario, si únicamente se selecciona en el resumen manual o en el automático, estaremos, respectivamente, ante un *falso negativo (false negative, FN)* o un *falso positivo (false positive, FP)*. A partir de esta clasificación, podemos definir las siguientes métricas clásicas de recuperación de información:

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

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

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

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

- La **medida-F** *(F-Score)* es 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.

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4}$$

2.3.3.2. Índice de Utilidad

Otro método de evaluación que se centra exclusivamente en la cobertura informativa del resumen es el denominado *Índice de Utilidad* (Radev, Jing, y Budzikowska, 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. Su cálculo se puede realizar de manera automática, aunque precisa de la asignación manual de los pesos que determinan la importancia de las distintas oraciones.

2.3.3.3. Similitud de Contenidos

Donaway *et al.* (2000) proponen como métrica para evaluar la calidad informativa de un resumen la denominada *Similitud de Contenidos*, que puede ser utilizada tanto para evaluar resúmenes generados por extracción como resúmenes generados por abstracción. Una de las medidas definidas para calcular dicha similitud es la *Prueba de Vocabulario (Vocabulary Test)*, donde se aplican métodos tradicionales de recuperación de información para calcular la distancia entre las representaciones vectoriales de los resúmenes automático y manual utilizando la métrica del coseno. De nuevo, su cálculo puede ser automatizado, aunque generalmente los resúmenes de referencia habrán de ser elaborados manualmente.

2.3.3.4. SEE: Summary Evaluation Environment

En el año 2000, y promovida por el *National Institute of Standards and Technology (NIST*, comienza su andadura la serie de conferencias *DUC (Document Understanding Conferences)*[1]. Desde entonces, y hasta 2007, año en que pasan a formar parte de las conferencias *TAC (Text Analysis Conference)*[2], se han celebrado ininterrumpidamente cada año. Surgen como una iniciativa para el desarrollo de un marco común para la evaluación (y consecuente mejora) de los sistemas de resúmenes automáticos, y se han convertido en el principal foro de evaluación de este tipo de sistemas.

En cada una de sus ediciones 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. La organización prepara los conjuntos de documentos, que suelen consistir

[1] Document Understanding Conferences (DUC). http://duc.nist.gov/. Consultada el 1 de noviembre de 2010

[2] Text Analysis Conference (TAC). http://www.nist.gov/tac/. Consultada el 1 de noviembre de 2010

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 que debían comparar los resultados de los distintos 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)* (Lin y Hovy, 2002b). El método de evaluación subyacente consiste en dividir los textos a comparar en "unidades de discurso", que el revisor debe 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, se utiliza la cobertura (*coverage*) como medida de la adecuación del resumen al modelo, definida como sigue (Ecuación 2.5):

$$C = \frac{MU\_que\_coinciden \times E}{Total\_MU\_modelo} \tag{2.5}$$

Donde *MU* representa a cada unidad de discurso, y E indica el ratio de completitud o de coincidencia entre las unidades, y puede variar de 1 a 0: 1 para *all* (todo), $\frac{3}{4}$ para *most* (en su mayoría), $\frac{1}{2}$ para *some* (algo), $\frac{1}{4}$ para *hardly any* (casi nada) y 0 para *none* (nada) (Lin y Hovy, 2002a). La Figura 2.7 muestra el sistema software desarrollado para realizar la evaluación siguiendo la metodología SEE.

2.3.3.5. ROUGE: Recall-Oriented Understudy for Gisting Evaluation

A partir de la edición de 2004 de las *Document Understanding Conferences*, la evaluación del contenido informativo de los resúmenes se realiza de modo automático. En DUC 2004 comienza a utilizarse un sistema, *ROUGE* (Lin, 2004b), que posteriormente se convertiría en el más utilizado para la evaluación de resúmenes generados tanto por extracción como por abstracción. ROUGE son las siglas de *Recall-Oriented Understudy for Gisting Evaluation*, y comprende un amplio abanico de métricas que evalúan la cobertura de contenidos mediante la comparación del resumen automático con otro

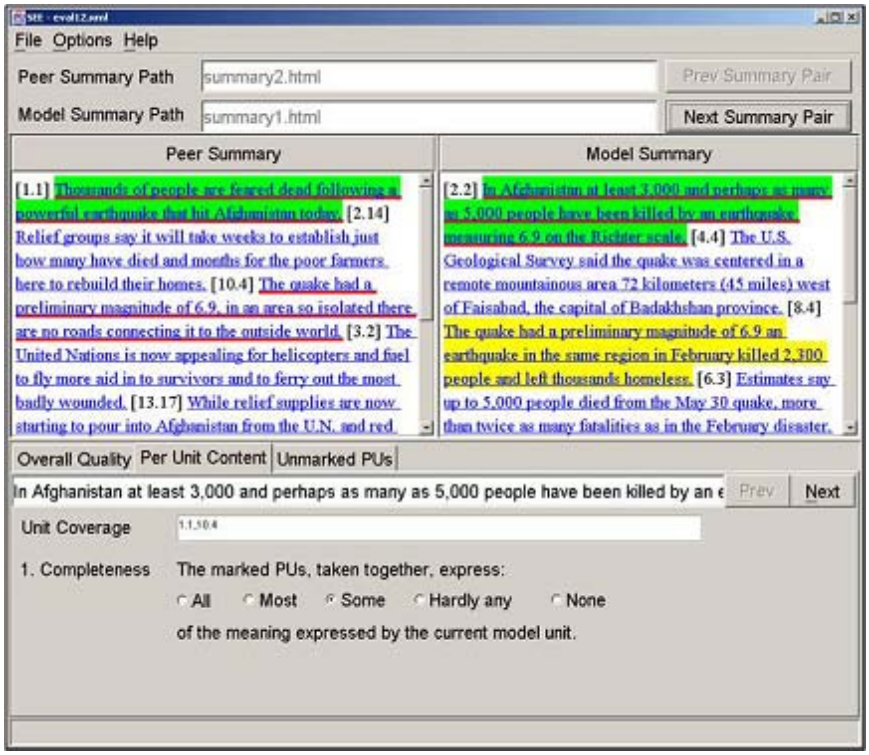


Figura 2.7: Interfaz de usuario de SEE (Lin y Hovy, 2002)

u otros considerados ideales, y mediante el cálculo de las unidades coincidentes entre ambos tipos de resúmenes. Entre las medidas calculadas por la herramienta destacan, por su correlación demostrada con juicios humanos, las siguientes:

- **ROUGE-N**, que contabiliza el número de secuencias de palabras (*n-gramas*) que coinciden entre un resumen candidato y uno o más modelos, por lo que se pueden calcular las medidas ROUGE-1, -2, -3, etc.
- **ROUGE-L**, que emplea la longitud de las secuencias más largas que coinciden en el candidato y en el modelo.
- **ROUGE-W**, una versión ponderada de ROUGE-L que, además de la longitud de la secuencia, valora la ausencia de "huecos" en la misma.
- **ROUGE-SN**, que tiene en cuenta bi-gramas que no necesariamente han de aparecer consecutivos en el texto, sino que pueden presentar hasta un máximo de N términos entre ellos.

A pesar de las críticas recibidas, Lin y Hovy (2003) demostraron, 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.

2.3.3.6. El Método Pirámide

En la edición de 2005 de la conferencia DUC, y con carácter opcional, los participantes tenían la posibilidad de someter sus resúmenes a una evaluación manual según el método *Pyramid* (Passonneau et al., 2005; Nenkova, Passonneau, y McKeown, 2007). Posteriormente, este método ha seguido utilizándose, junto con ROUGE, en las subsiguientes ediciones de las conferencias DUC y, en la actualidad, en las conferencias TAC. 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. 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 SCUs, cuya altura será igual al número n de resúmenes de referencia considerados. A cada SCU de una capa T_i se le asigna un peso W_i que depende del número de resúmenes en los que aparece, de manera que las SCU de mayor importancia se sitúan en la cúspide de la pirámide. Si D_i es el número de SCU de un resumen que aparecen en el nivel T_i, entonces el peso del resumen D se calcula utilizando la Ecuación 2.6.

$$D = \sum_{i=1}^{n} i \times D_i \tag{2.6}$$

De este modo, el mejor resumen será aquel que contenga más SCU de los niveles superiores. La Figura 2.8 ilustra un posible ejemplo de evaluación utilizando una pirámide de cuatro niveles de altura, en la que se han definido dos SCU en el nivel superior, tres SCU en el segundo nivel y dos SCU en el tercero, y en la que los resúmenes evaluados están representados mediante elipses que rodean a los SCU coincidentes con los definidos en la pirámide.

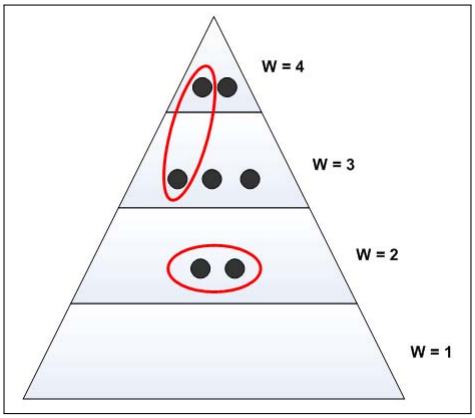


Figura 2.8: Evaluación de resúmenes con el método Pirámide

2.3.3.7. Evaluación de la Legibilidad

Con respecto a la legibilidad de los resúmenes, desde la conferencia DUC 2005 y hasta la actualidad, se evalúan manualmente los siguientes criterios lingüísticos (Dang, 2005):

- **Corrección gramatical**: El texto no debe presentar faltas ortográficas, errores de formato, fragmentos incoherentes, etc.
- **Redundancia**: El resumen no debe presentar información innecesaria o repetida. En especial, no debería repetir un nombre o sintagma nominal más de lo necesario, sino sustituirlo por el correspondiente pronombre siempre que sea posible.
- **Claridad referencial**: Debe ser sencillo identificar qué o quiénes son las entidades a las que hace referencia el texto.
- **Foco**: El resumen debe incluir únicamente información relacionada entre sí y con el tema principal del documento.
- **Estructura y coherencia**: El resumen debe estar bien estructurado y bien organizado, constituir un conjunto coherente de información.

La calidad de un resumen con respecto a cada uno de estos criterios recibe una puntuación en una escala de cinco posibles valores: *1-very poor* (muy pobre), *2-poor* (pobre), *3-acceptable* (aceptable), *4-good* (buena) y *5-very good* (muy buena). 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, por lo que suelen ir acompañadas de una evaluación del contenido informativo del resumen, generalmente, utilizando ROUGE.

2.4. Nuevas Aplicaciones de la Generación Automática de Resúmenes

La generación automática de resúmenes, tal y como fue inicialmente concebida (resúmenes indicativos de un único documento y en un único idioma) ha evolucionado progresivamente y dado lugar a un amplio espectro de tareas y aplicaciones. A continuación se comentan brevemente aquellas que, en opinión de la autora, merecen especial atención por su difusión y trascendencia en el momento de redactar este trabajo.

2.4.1. Resúmenes Multi-documento

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 o bien coincide, o bien es irrelevante o de carácter secundario. La consideración de este problema ha resultado en la extensión de las investigaciones en generación de resúmenes mono-documento a la generación de resúmenes a partir de colecciones de documentos relacionados temáticamente.

La tarea de generar resúmenes a partir de múltiples fuentes plantea retos adicionales. En primer lugar, la selección de los documentos que comparten una misma temática, y que contribuirán a la redacción de un mismo resumen, debe realizarse cuidadosamente, para evitar mezclar en el resumen información inconexa y no relacionada. En segundo lugar, el hecho de contener los documentos información común puede dar lugar a resúmenes redundantes. Por ello, la detección y eliminación de la redundancia es uno de los principales problemas a los que se enfrenta la generación automática de resúmenes multi-documento. 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.

Al igual que ocurriera en la generación de resúmenes mono-documento, los sistemas multi-documento se pueden clasificar en función del uso de conocimiento y de la profundidad del análisis realizado. La mayoría de los sistemas son, una vez más, sistemas de extracción de oraciones. Las técnicas utilizadas para seleccionar estas oraciones son también similares, si no las mismas, que en el caso de los sistemas mono-documento. Frecuentemente incorporan distintos módulos para tratar de solventar, con mayor o menor éxito, algunos de los problemas comentados, en especial, la agrupación de documentos de contenido similar y la detección y eliminación de la redundancia.

En cuanto al primer problema se refiere, la agrupación de documentos similares, entre las contribuciones más destacadas se encuentran el enfoque de reformulación de MULTIGEN (Hatzivassiloglou et al., 2001), el uso de Webclopedia en NEATS (Lin y Hovy, 2001) y la aproximación basada en grafos de MEAD (Radev, BlairGoldensohn, y Zhang, 2001).

En relación al segundo problema, la eliminación de la redundancia, un enfoque comúnmente utilizado consiste en medir la similitud entre pares de oraciones y utilizar alguna técnica de agrupamiento para identificar temas *themes* de información común (Mckeown et al., 1999; Radev, Jing, y Budzikowska, 2000; Marcu y Gerver, 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 (Hendrickx et al., 2009). Una medida muy popular es la *Relevancia Marginal Máxima (Maximum Marginal Relevance, MMR)*, utilizada en los trabajos de Carbonell, Geng y Goldstein (1997) y Carbonell y Goldstein (1998). En el contexto multi-documento, y de un resumen orientado a la consulta de un usuario, tratan de evitar el problema de contemplar 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 Multi-documento (Maximum Marginal Relevance-*

MultiDocument, MMR-MD). En Radev (2000) se introduce la noción de *Subsunción de Información entre Oraciones (Cross-Sentence Informational Subsumption, CSIS)*, 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, como son la equivalencia semántica, equivalencia informativa, igualdad literal e inclusión informativa. En Lloret *et al.* (2008), el problema de la redundancia se resuelve utilizando técnicas de implicación textual (Herrera, Peñas, y Verdejo, 2005), de modo que, si se determina que una oración O_1 implica otra O_2, ambas oraciones se consideran equivalentes desde el punto de vista de la información que presentan, y en consecuencia, sólo se incluye en el resumen aquella que mejor puntuada resulte por el método de selección de oraciones. En Zhao *et al.* (2009), el problema se aborda mediante filtrado de contenidos, utilizando lógica borrosa para determinar el grado de similitud entre las oraciones candidatas a añadir en el resumen y las oraciones que ya forman parte de él.

2.4.2. Resúmenes Adaptados al Usuario o a una Consulta

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 quienes van dirigidos.

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, utilizando la Relevancia Marginal Máxima para ordenar las oraciones en función de su similitud con los términos introducidos en la consulta y evitar la inclusión de oraciones redundantes. Por su parte, Sanderson (1998) selecciona el pasaje más relevante según la consulta, utilizando la técnica del *Análisis del Contexto Local (Local Context*

Analysis, LCA). 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 otros muchos trabajos, la expansión de la consulta se realiza utilizando *WordNet* (Maña, de Buenaga, y Gómez, 1999; Amini y Gallinari, 2002; Zhao, Wu, y Huang, 2009). Finalmente, también es posible utilizar técnicas de aprendizaje para confeccionar resúmenes personalizados. Un trabajo en esta dirección puede encontrarse en Lin (1999), donde las características que se exploran son el número de palabras de la consulta y el número de palabras frecuentes que aparecen en la oración. Otro trabajo que utiliza técnicas de aprendizaje automático es el de Wei *et al.* (2009), esta vez empleando dos tipos de características o atributos (dependientes e independientes de la consulta) para entrenar dos clasificadores que se retroalimentan en la etapa de selección de oraciones para el resumen.

2.4.3. Resúmenes Multilingües

La generación automática de resúmenes multilingüe es un área de investigación nacida en los últimos años, y cuyo objetivo es adaptar las técnicas tradicionales de extracción para contemplar documentos redactados en distintos idiomas. Algunos sistemas como SUMMARIST (Hovy y Lin, 1999) generan resúmenes multi-documento y multilingües extrayendo las oraciones de documentos en distintos idiomas, y traduciendo el resumen resultado a uno de ellos o a un idioma distinto. Por el contrario, otros sistemas como NewsBlaster (Blair-Goldensohn et al., 2004) o los presentados por Bouayad-Agha *et al.* (2009) y Saggion (2008), realizan la traducción antes de realizar la extracción de las oraciones; es decir, realizan un paso previo para traducir a un idioma común todos los documentos a resumir. No obstante, la mayoría de estos sistemas presentan un resultado pobre en cuanto a legibilidad y a calidad gramatical se refiere, debido principalmente al software utilizado para realizar la traducción automática.

Capítulo 3

Herramientas y Recursos

El propósito de este capítulo es presentar las diferentes herramientas y recursos lingüísticos que, no habiendo sido desarrollados expresamente para este proyecto, han sido utilizados en el mismo.

La primera sección está dedicada al estudio de algunas de las ontologías y terminologías, tanto de propósito general como del dominio biomédico, más utilizadas en tareas de procesamiento de lenguaje natural.

La segunda sección describe el propósito y funcionamiento de la arquitectura para ingeniería textual *GATE*, así como el interés de su utilización en tareas de procesamiento del lenguaje.

La tercera sección introduce *MetaMap*, una aplicación que permite identificar conceptos biomédicos del Metatesauro de *UMLS* inmersos en textos no estructurados.

La cuarta sección presenta *Personalized PageRank*, un algoritmo no supervisado de desambiguación léxica basado en grafos.

Finalmente, las dos últimas secciones presentan las herramientas *WordNet::Similarity* y *WordNet:SenseRelate*, cuyos propósitos son, respectivamente, calcular la similitud semántica entre conceptos de *WordNet* y desambiguar el significado de los mismos.

3.1. Ontologías, Terminologías y Léxicos

El término *ontología* se define en 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) la define como una base de datos que describe los conceptos del mundo o de algún dominio, sus propiedades y las relaciones que se establecen entre ellos.

Según Cabré (1995), la *terminología* es, desde la óptica de una disciplina científico-técnica, el conjunto de las unidades de expresión y comunicación que permiten transferir el pensamiento especializado. En este sentido, el trabajo terminológico no se limita a recopilar las denominaciones de una determinada área con una finalidad informativa o descriptiva, sino que persigue además el objetivo de fijar unas unidades terminológicas como formas normalizadas, en aras de la consecución de una comunicación profesional precisa, moderna y unívoca.

Por su parte, y apropiándonos de la definición de Boguraev y Pustejovsky (1996), el término *base de conocimiento léxico* hace referencia a grandes repositorios de información léxica, que incorporan, más allá de la descripción estática de las palabras, propiedades y valores asociados a las mismas, restricciones sobre el comportamiento del mundo, diferentes interpretaciones de los términos en función del contexto o generalizaciones lingüísticas.

A continuación se describen brevemente algunas de las ontologías, terminologías y bases de conocimiento léxico más utilizadas en los sistemas de procesamiento de información. Las tres primeras se corresponden con terminologías especializadas en biomedicina que promueven una manera estándar de nombrar los conceptos del dominio (Bodenreider, Mitchell, y McCray, 2003). Sin lugar a dudas, los recursos lingüísticos más utilizados en tareas de procesamiento del lenguaje biomédico son *SNOMED-CT*, *MeSH* y *UMLS*; y es por ello que los hemos escogido como posibles candidatos a utilizar en el caso de estudio de generación de resúmenes de artículos científicos en biomedicina. Las tres últimas se corresponden con recursos de carácter general, y por tanto, son apropiadas para el procesamiento de textos que comunican conceptos universales y no específicos de una determinada disciplina. En concreto, se estudian la base de datos léxica *WordNet*, el tesauro *Roget* y la ontología *Cyc* como fuentes de conocimiento alternativas para los casos de estudio de generación de resúmenes de noticias periodísticas y de páginas web turísticas.

3.1.1. SNOMED-CT

SNOMED-CT[1] son las siglas de *Systematized Nomenclature of Medicine Clinical Terms*. Se trata de una extensa terminología médica desarrollada por el *College of American Pathologists (CAP)*, y mantenida por la *International Health Terminology Standards Development Organisation (IHTSDO)*. Disponible en varios idiomas, proporciona un lenguaje común que facilita la indexación, el almacenamiento, la recuperación y la agregación de datos médicos. SNOMED-CT presenta los siguientes componentes básicos:

- **Conceptos**: representan una unidad mínima de significado.
- **Jerarquías**: los conceptos en SNOMED-CT están organizados jerárquicamente. El concepto *SNOMED CT* es el nodo raíz. Bajo este, existen 19 jerarquías de primer nivel. Conforme se desciende en la jerarquía, los conceptos se hacen progresivamente más específicos. La Tabla 3.1 muestra un extracto de esta jerarquía.

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)

Tabla 3.1: Extracto de la jerarquía de conceptos de SNOMED-CT

- **Relaciones**: que enlazan conceptos entre sí. Existen dos tipos de relaciones: relaciones *"es un"*, que conectan conceptos en una jerarquía; y relaciones *"de atributos"*, que enlazan conceptos entre jerarquías (i.e. *"debido a"* o *"agente causal"*).
- **Descripciones**: términos o nombres asociados a un concepto que aportan una mayor flexibilidad a la hora de expresarlos.

[1]SNOMED International. SNOMED-CT. http://www.snomed.org/snomedct. Consultada el 1 de noviembre de 2010

La versión 2010 en inglés contiene más de 308.000 conceptos, 758.000 descripciones y 823.000 relaciones semánticas.

3.1.2. Medical Subject Headings

Medical Subject Headings (MeSH)[2] es un tesauro desarrollado por la *National Library of Medicine (NLM)* 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 especialidad. 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 conceptos como *tobillo*. En la versión 2009 de MeSH, se cuentan 25.186 descriptores, además de 180.000 conceptos suplementarios recogidos en un tesauro separado. Se incluyen también más de 160.000 términos de ayuda para localizar el descriptor más apropiado. Los árboles de descriptores no constituyen una clasificación exhaustiva de las distintas materias, sino que están diseñados para ayudar en las búsquedas en la base de datos *MEDLINE*; además de como guía para las personas encargadas de asignar categorías a documentos. A modo de ejemplo, la Figura 3.1 muestra la información asociada al concepto *Encephalitis* en MeSH.

3.1.3. Unified Medical Language System

Unified Medical Language System (UMLS)[3], desarrollado por la *National Library of Medicine*, es un sistema que garantiza referencias cruzadas entre más de cien vocabularios y clasificaciones en distintos idiomas, incluyendo MeSH y SNOMED-CT. UMLS presenta tres fuentes de conocimiento: el *Metatesauro*, el *Léxico Especializado* y la *Red Semántica*.

- El **Léxico Especializado**[4], únicamente disponible en lengua inglesa,

[2]National Library of Medicine. MeSH. http://www.nlm.nih.gov/mesh/. Consultada el 1 de noviembre de 2010

[3]National Library of Medicine. UMLS. http://www.nlm.nih.gov/research/umls. Consultada el 1 de noviembre de 2010

[4]National Library of Medicine. UMLS Specialist Lexicon fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. Consultada el 1 de noviembre de 2010

MeSH Heading	Encephalitis
Tree Number	C02.182.500
Tree Number	C10.228.140.430
Tree Number	C10.228.228.210.150
Tree Number	C10.228.228.245
Annotation	GEN; coord with specific organism /infection heading (IM) or other cause (IM); viral encephalitis = ENCEPHALITIS, VIRAL unless ENCEPHALITIS, ARBOVIRUS but see note there; ENCEPHALOMYELITIS & specifics & SUBACUTE SCLEROSING PANENCEPHALITIS (see note there) are also available; DF: ENCEPH
Scope Note	Inflammation of the BRAIN due to infection, autoimmune processes, toxins, and other conditions. Viral infections (see ENCEPHALITIS, VIRAL) are a relatively frequent cause of this condition.
Entry Term	Brain Inflammation
Entry Term	Encephalitis, Infectious
...	...
See Also	Encephalomyelitis
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 RA RH RI RT SU TH UR US VE VI
Entry Version	ENCEPH
Date of Entry	19990101
Unique ID	D004660

Figura 3.1: Información asociada al concepto *Encephalitis* en MeSH. Fuente: http://www.nlm.nih.gov/mesh/. Consultada el 1 de noviembre de 2010

contiene en su versión 2009AA unos 108.000 informes léxicos y más de 186.000 cadenas de términos. Cada entrada presenta información sintáctica, morfológica y ortográfica, incluyendo la categoría sintáctica (verbo, sustantivo, adjetivo, adverbio, pronombre, etc.), las inflexiones de género y número, las conjugaciones de los verbos, los comparativos y superlativos de los adjetivos y adverbios, e incluso posibles patrones de complementariedad (objetos y otros argumentos que pueden acompañar a los verbos, nombres y adjetivos). La Tabla 3.2 muestra, a modo de ejemplo, la información asociada al término *anaesthetic* en el Léxico Especializado.

- El **Metatesauro**[5] es una base de datos multilingüe y multipropósito que contiene información sobre conceptos biomédicos y relacionados

[5]National Library of Medicine. UMLS Metathesaurus fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. Consultada el 1 de noviembre de 2010

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

Tabla 3.2: Término *anaesthetic* en el Léxico Especializado

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 de literatura biomédica y en la investigación clínica.

El Metatesauro está organizado por conceptos. Un concepto es una unidad de significado. Cada concepto en el Metatesauro queda definido por un identificador de concepto unívoco (*Concept Unique Identifier, CUI*) y un texto también unívoco (*String Unique Identifier, SUI*), aunque puede presentar más de un nombre (*Atom Unique Identifier, AUI*). Los conceptos, sus distintos nombres y sus vocabularios de origen se almacenan en la tabla *MRCONSO*. A modo de ejemplo, la Tabla 3.3 muestra la información almacenada para el concepto *C0001175:AIDS* (SIDA).

CUI=C0001175	**AUI**=A2878223
language=ENG	**source**=SNOMEDCT
status=S	**string_type_source**=PT
LUI=L0001842	**code**=62479008
string_type=PF	**string**=AIDS
SUI=S0011877	**restriction_level**=4
preference=N	**suppress**=N

Tabla 3.3: Entrada en la tabla *MRCONSO* del Metatesauro para el concepto *C0001175:AIDS*

El propósito del Metatesauro es enlazar nombres alternativos y vistas de un mismo concepto, así como identificar relaciones útiles entre diferentes conceptos. En concreto, estas relaciones pueden ser de dos tipos: entre conceptos dentro de un mismo vocabulario o entre con-

ceptos de diferentes vocabularios. Todas las relaciones se encuentran almacenadas en la tabla *MRREL*, a excepción de las relaciones de co-ocurrencia, que se encuentran en *MRCOC*, y las de equivalencia entre conceptos de distintos vocabularios, que se encuentran en *MRMAP* y *MRSMAP*. Ejemplos de estas relaciones son *CHD* ("hijo"), *PAR* ("padre"), *QB* ("puede ser calificado por"), *RQ* ("relacionado y posible sinónimo") y *RO* ("relacionado con"). Así, por ejemplo, los conceptos *C0009443:Common Cold* (resfriado común) y *C0027442:Nasopharynx* (nasofaringe) están conectados a través de la relación *RO*. No obstante, muchas de las relaciones almacenadas en la tabla *MRREL* se encuentran replicadas en otras tablas. Así, por ejemplo, la tabla *MRHIER* lista las jerarquías a las que pertenece un concepto. Permite saber, por ejemplo, que el concepto *C0035243:Respiratory Tract Infections* es un padre del concepto *C0009443:Common Cold*.

Todos los conceptos del Metatesauro están, a su vez, asignados al menos a un tipo de la Red Semántica a través de la tabla *MRSTY*. Así, por ejemplo, el concepto *C0009443:Common Cold* pertenece al tipo semántico *Disease or Syndrome*.

- La **Red Semántica**[6] consiste en un conjunto de 132 categorías o *tipos semánticos* que garantizan una clasificación consistente de todos los conceptos representados en el Metatesauro, así como un conjunto de relaciones entre tales tipos semánticos. Los 53 enlaces entre los tipos semánticos establecen la estructura de la red y representan las relaciones más importantes en el dominio biomédico. La relación principal es la hiperonimia (*is a*), que establece la jerarquía entre los tipos de la red. Existe otro grupo de relaciones, agrupadas en cinco categorías principales: *physically related to* (relación física), *spatially related to* (relación espacial), *temporally related to* (relación temporal), *functionally related to* (relación funcional) y *conceptually related to* (relación conceptual). La tabla *SRSTR* describe la estructura de la Red Semántica, ya que almacena las relaciones existentes entre tipos, incluidas tanto las relaciones jerárquicas como las no jerárquicas. Así, por ejemplo, los tipos semánticos *Disease or Syndrome* (Enfermedad o síndro-

[6] National Library of Medicine. UMLS Semantic Network fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html. Consultada el 1 de noviembre de 2010

me) y *Pathologic Function* (Función patológica) están conectados a través de la relación *es un* en esta tabla.

UMLS dispone de una herramienta, **MetamorphoSys**[7], pensada para facilitar la configuración y personalización del Metatesauro. En general, un usuario puede estar interesado en crear subconjuntos personalizados del Metatesauro por dos motivos:

- Para excluir vocabularios que no necesita en su aplicación. De esta forma, se reduce significativamente su tamaño y se facilita su manejo.
- Para modificar el formato de los datos de salida y aplicarles distintos filtros.

El resultado de la ejecución de MetamorphoSys es un conjunto de archivos de extensión *ORF (Original Release Format)* o *RRF (Rich Release Format)* que contienen los subconjuntos seleccionados del Metatesauro, el Léxico Especializado y la Red Semántica, junto con un script para cargarlos en una base de datos *Oracle* o *Mysql*. La Figura 3.2 muestra la interfaz de usuario de MetamorphoSys para la selección de vocabularios de UMLS.

3.1.4. Roget's Thesaurus

El tesauro *Roget* fue escrito en 1805 por el cirujano inglés Peter Mark Roget y publicado por primera vez en 1852. A diferencia de los diccionarios tradicionales, la organización no es alfabética, sino que clasifica el conocimiento como una jerarquía de nueve niveles, desde las *clases* en la raíz, hasta las *palabras* en las hojas. Las seis clases en las que se organiza el conocimiento son *Words expressing abstract relations* (Palabras que expresan relaciones abstractas), *Words relating to space* (Palabras relacionadas con el espacio), *Words relating to matter* (Palabras relacionadas con la materia), *Words relating to the intellectual faculties* (Palabras relacionadas con las facultades intelectuales), *Words relating to the voluntary powers* (Palabras relacionadas con la voluntad)y *Words relating to the sentiment and moral* (Palabras relacionadas con los sentimientos y la moral). A su vez, por ejemplo, la clase *Words expressing abstract relations* se subdivide en las secciones *Being in*

[7]National Library of Medicine. UMLS MetamorphoSys fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlsmetamorph.html. Consultada el 1 de noviembre de 2010

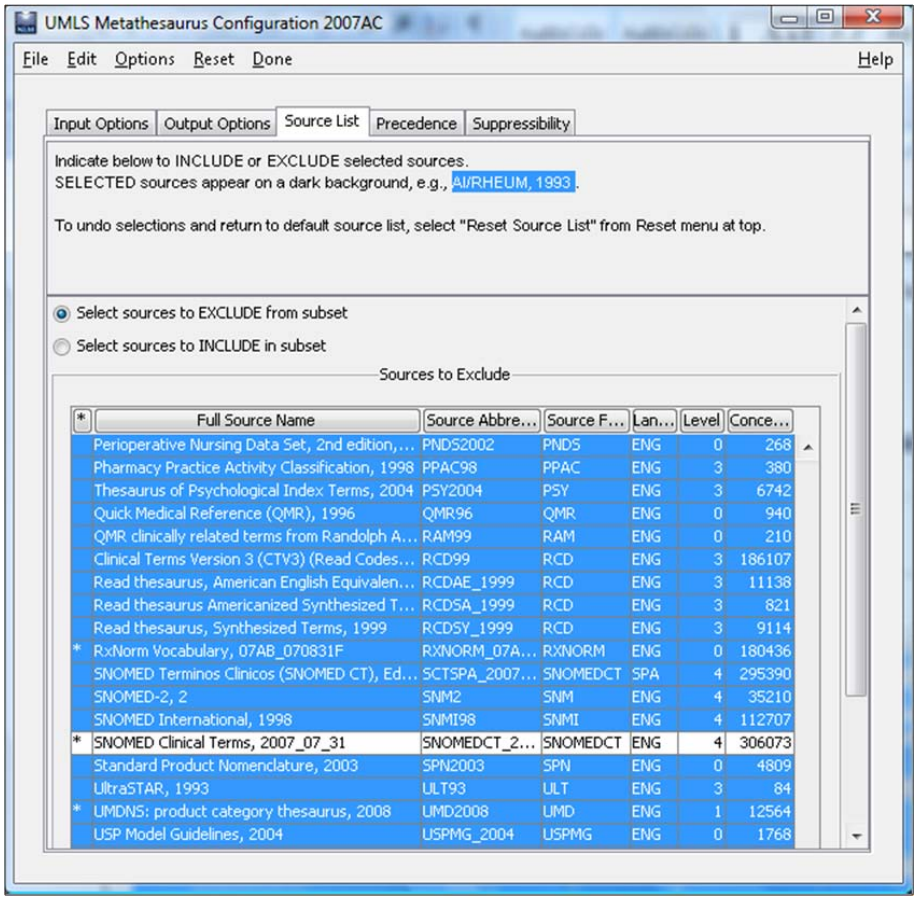
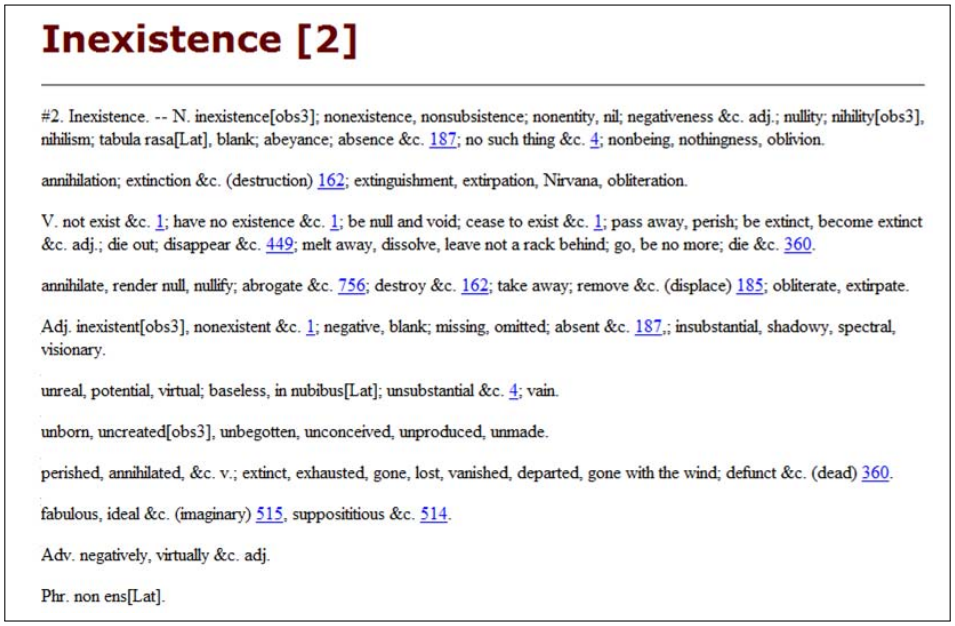


Figura 3.2: Selección de fuentes en MetamorphoSys

the abstract (el Ser desde una perspectiva abstracta), *Being in the concrete* (el Ser desde una perspectiva concreta), *Formal existence* (Existencia formal) y *Modal existence* (Existencia modal). Por su parte, la sección *Being in the abstract* presenta los encabezados *Existence* (Existencia) y *Nonexistence* (Inexistencia). A partir de este tercer nivel comienzan a aparecer conjuntos específicos de palabras, entre las que se intercalan palabras clave en cursiva precedidas por números, que sirven como referencias cruzadas a otros conjuntos del tesauro (Figura 3.3).

Así pues, el tesauro Roget no etiqueta explícitamente las relaciones entre términos, sino que los agrupa con relaciones implícitas. De esta forma, *become* (convertirse en) se relaciona con *conversion* (conversión) y con *developing* (producción); mientras que *production* (producción) se relaciona, entre otros, con *destruction* (destrucción).



Inexistence [2]

#2. Inexistence. -- N. inexistence[obs3]; nonexistence, nonsubsistence; nonentity, nil; negativeness &c. adj.; nullity; nihility[obs3], nihilism; tabula rasa[Lat], blank; abeyance; absence &c. 187; no such thing &c. 4; nonbeing, nothingness, oblivion.

annihilation; extinction &c. (destruction) 162; extinguishment, extirpation, Nirvana, obliteration.

V. not exist &c. 1; have no existence &c. 1; be null and void; cease to exist &c. 1; pass away, perish; be extinct, become extinct &c. adj.; die out; disappear &c. 449; melt away, dissolve, leave not a rack behind; go, be no more; die &c. 360.

annihilate, render null, nullify; abrogate &c. 756; destroy &c. 162; take away; remove &c. (displace) 185; obliterate, extirpate.

Adj. inexistent[obs3], nonexistent &c. 1; negative, blank; missing, omitted; absent &c. 187.; insubstantial, shadowy, spectral, visionary.

unreal, potential, virtual; baseless, in nubibus[Lat]; unsubstantial &c. 4; vain.

unborn, uncreated[obs3], unbegotten, unconceived, unproduced, unmade.

perished, annihilated, &c. v.; extinct, exhausted, gone, lost, vanished, departed, gone with the wind; defunct &c. (dead) 360.

fabulous, ideal &c. (imaginary) 515, supposititious &c. 514.

Adv. negatively, virtually &c. adj.

Phr. non ens[Lat].

Figura 3.3: Cabecera *Existence* en Roget's Thesaurus. Fuente: http://poets.notredame.ac.jp/Roget/contents.html. Consultada el 1 de noviembre de 2010

De las 15.000 entradas del manuscrito original se ha pasado a las más de 250.000 en su décima edición, que data de 1992; y en la actualidad se encuentra disponible en numerosas fuentes en Internet. En concreto, la versión de 1911 digitalizada se puede obtener a través de la página del Proyecto Gutenberg[8], pero también puede consultarse de forma interactiva desde la web del proyecto ARTFL[9].

Tanto la versión de 1911 como sus versiones posteriores han sido utilizadas frecuentemente en tareas de procesamiento de lenguaje. Cassidy (2000), por ejemplo, lo utilizó para construir la red semántica FACTORUM; Jarmasz y Szpakowicz (2003) lo emplean para automatizar la construcción de cadenas léxicas, basándose en el trabajo de Morris y Hirst (1991), donde dichas cadenas se construían manualmente utilizando el tesauro Roget. Yarowsky (1992) concibe las categorías del tesauro como aproximaciones de clases conceptuales, y las utiliza para construir modelos estadísticos para desambiguación léxica.

[8]The Gutenberg Project. http://www.gutenberg.org/ebooks/22. Consultada el 1 de noviembre de 2010

[9]The ARTFL Project. Roget's Thesaurus Search Form. http://machaut.uchicago.edu/rogets. Consultada el 1 de noviembre de 2010

3.1.5. The Cyc Knowledge Base

La ontología *Cyc*[10] es un repositorio de propósito general en desarrollo desde 1984, cuyo objetivo es almacenar tanto conocimiento especializado como de "sentido común". Su versión de código abierto, *OpenCyc 2.0*, consta de más de 300.000 conceptos y aproximadamente 3.500.000 afirmaciones explícitas o axiomas sobre la realidad (Lenat, 1995). El principal objetivo de esta base de conocimiento es abarcar todos los objetos y acciones de la realidad del ser humano que, por resultar obvias o evidentes, difícilmente se encuentran publicados en los libros de texto o enciclopedias; como por ejemplo, que se ha de estar despierto para comer, o que no se puede recordar algo si aún no ha sucedido. Codificar este conocimiento de sentido común implica, por lo tanto, trabajar con relaciones de tiempo, causa, espacio, sustancia, intención, contradicción, incertidumbre, creencias, emociones, planes, etc.

El conocimiento en Cyc se representa en *CycL*, un lenguaje de alto nivel basado en lógica de predicados. La representación de los significados de las palabras en Cyc se consigue fundamentalmente de manera declarativa, utilizando un vocabulario específico para la representación de las palabras, sus propiedades (i.e. categoría gramatical) y de predicados que definen la correspondencia y relaciones entre las palabras en lenguaje natural y los conceptos en Cyc. Por ejemplo, el término *#$TemporalStuffType* se utiliza para definir la persistencia de los objetos en el tiempo, mientras que el término *#$isa* describe la propiedad "es un tipo de" entre distintos objetos. La Figura 3.4 muestra un extracto de la entrada *Road Vehicle* (Vehículo de carretera) en OpenCyc.

Cyc no es una ontología de significados, sino que los conceptos que se representan son únicamente aquellos necesarios para permitir el razonamiento de sentido común, por lo que no existe una traducción comprensible entre los conceptos de Cyc y las palabras del lenguaje natural.

Muchos trabajos han utilizado Cyc para facilitar distintas tareas de procesamiento de lenguaje. En Curtis *et al.* (2005) se describe un sistema de preguntas y respuestas. En Witbrock *et al.* (2003) se presenta un sistema interactivo de adquisición de conocimiento. En ambos trabajos destaca el papel predominante de la relación *#$isa* en el razonamiento. Por ejemplo, en respuesta a una palabra ambigua como *turkey*, el sistema pregunta al usuario si se refiere al término turkey cuando éste se trata de un pájaro, un

[10]Open Cyc. http://www.opencyc.org/. Consultada el 1 de noviembre de 2010

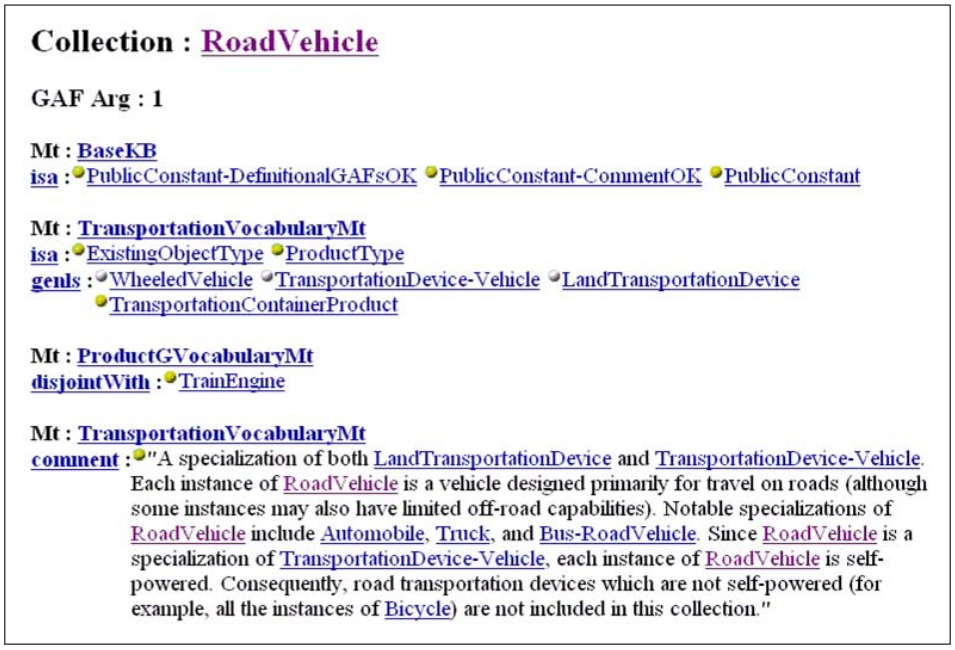


Figura 3.4: Información en OpenCyc para el concepto *Road Vehicle*.
Fuente: http://www.opencyc.org/. Consultada el 1 de noviembre de 2010

tipo de carne o un país. Para terminar, cabe mencionar el trabajo de Curtis *et al.* (2006), en el que se presenta una aplicación de Cyc a la tarea de desambiguación léxica que hace uso de relaciones más allá de la hiperonimia, y de propiedades de alto nivel de los objetos implicados.

3.1.6. WordNet

WordNet[11] (Miller et al., 1990; Miller et al., 1998) es una base de datos léxica para el inglés, desarrollada en el marco de un proyecto iniciado en 1985, y en el que participaron conjuntamente distintas organizaciones gubernamentales y privadas norteamericanas, entre ellas, la Universidad de Princeton y el Departamento de Investigación Naval. En la actualidad, es mantenida por el Laboratorio de Ciencias Cognitivas de la Universidad de Princeton. WordNet persigue dos objetivos principales: por un lado, construir una combinación de diccionario y tesauro que sea intuitivo y fácil de utilizar; y por otro lado, dar soporte en tareas de análisis textual y procesamiento del lenguaje natural.

La diferencia fundamental de WordNet respecto a otros sistemas con

[11] WordNet. A lexical database for English. http://wordnet.princeton.edu/wordnet/. Consultada el 1 de noviembre de 2010

propósitos similares radica en la organización del léxico en torno a cinco categorías: nombres, verbos, adjetivos, adverbios y elementos funcionales. Para la representación de los conceptos, WordNet utiliza los denominados *synonym sets* o *synsets*, que pueden verse como grupos de elementos de datos semánticamente equivalentes. Al contrario de lo que ocurre con los diccionarios de sinónimos o tesauros tradicionales, un synset no tiene una palabra que actúa como identificador del conjunto. El significado del synset lo aportan pequeñas definiciones (*glosses*), que en ocasiones pueden ser ejemplos de oraciones que matizan el significado del concepto (Tabla 3.4). Además, WordNet proporciona para cada palabra lo que se ha denominado *cuenta polisémica*, y que no es más que una medida del grado en que la palabra se utiliza con cada uno de sus significados. De este modo, si una palabra presenta un valor muy alto para un determinado synset, se puede inferir que se trata su acepción más habitual.

(20) pretty – (pleasing by delicacy or grace; not imposing; “pretty girl”; “pretty song”; “pretty room”)
(3) pretty – ((used ironically) unexpectedly bad; “a pretty mess”; “a pretty kettle of fish”)

Tabla 3.4: Definiciones o *glosses* en WordNet para el sustantivo *pretty*

Además de la relación léxica de sinonimia, WordNet ofrece otras relaciones semánticas como la antonimia, hiperonimia, hiponimia, meronimia y relaciones morfológicas, que se expresan como punteros entre synsets. No obstante, las relaciones se organizan de manera distinta para cada una de las cinco categorías sintácticas en las que se estructura WordNet, aunque todas ellas presentan la relación básica de sinonimia. A continuación, se profundizará en las relaciones correspondientes a aquellas categorías gramaticales que se han considerado, a priori, de interés para el propósito de este trabajo: sustantivos, verbos y adjetivos.

- **Sustantivos**. WordNet presenta aproximadamente 57.000 formas nominales organizadas en unos 48.000 synsets, organizados en una jerarquía a través de relaciones de generalización (hiperonimia) entre synsets, creándose un sistema de herencia léxica en el que cada palabra hereda los rasgos distintivos de su ancestro. Así, por ejemplo, el primer significado de la palabra *rose* presenta la jerarquía de hiperónimos mostrada en la Figura 3.5, cada uno de los cuales viene

acompañado de su definición o gloss.

Sense 1
rose, rosebush – (any of many shrubs of the genus Rosa that bear roses)
shrub, bush – (a low woody perennial plant usually having several major...)
woody plant, ligneous plant – (a plant having hard lignified tissues or ...)
vascular plant, tracheophyte – (green plant having a vascular ...)
plant, flora, plant life – (a living organism lacking the power ...)
organism, being – (a living thing that has (or can develop) the ...)
living thing, animate thing – (a living (or once living) entity)
object, physical object – (a tangible and visible entity; an ...)
physical entity – (an entity that has physical existence)
entity – (that which is perceived or known or inferred to ...)

Tabla 3.5: Hiperónimos del lexema *rose* en WordNet

Además de la relación de hiperonimia, los sustantivos en WordNet se relacionan a través de las siguientes relaciones entre sus synsets:

- **Hiponimia (*hyponym*)**: Y es un hipónimo de X si todo Y es un X. Por ejemplo, *trucha* es un hipónimo de *pescado*, puesto que trucha es un tipo de pescado.
- **Holonimia (*holonym*)**: Y es un holónimo de X si X es parte de Y. Por ejemplo, *ventana* es un holónimo de *coche*, puesto que la ventana es una parte del coche.
- **Meronimia (*meronym*)**: Y es un merónimo de X si Y es parte de X. Por lo tanto, *coche* es un merónimo de *ventana*.
- **Términos coordinados (*coordinate terms*)**: X e Y son términos coordinados si comparten un mismo hiperónimo. Por ejemplo, *edificio* y *puente* son términos coordinados, ya que comparten el hiperónimo *construcción*.

- **Verbos**. WordNet presenta más de 21.000 formas verbales, organizadas en aproximadamente 8.400 synsets, organizados en 15 dominios semánticos diferentes (Figura 3.6).

  A diferencia de lo que ocurriera con los sustantivos, que se organizaban en torno a un sistema de herencia léxica, la estructuración de los verbos en WordNet se apoya en el concepto de *implicación léxica (lexical entailment)*. Este principio se refiere a la relación existente entre dos verbos, V_1 y V_2, cuando la oración *alguien/algo* V_1 implica lógicamente

verbs of body care and functions	verbs of creation
verbs of change	verbs of emotion
verbs of cognition	verbs of motion
verbs of communication	verbs of perception
verbs of competition	verbs of possession
verbs of consumption	verbs of social interaction
verbs of contact	weather verbs

Tabla 3.6: Grupos genéricos de verbos en WordNet

a la oración *alguien/algo* V_2. Por ejemplo, *roncar* implica *dormir*, ya que la oración *Está roncando* implica a la oración *Está durmiendo*. Nótese que se trata de una relación unilateral; es decir, si el verbo V_1 implica el verbo V_2, no puede ocurrir que V_2 implique V_1, salvo que ambos verbos sean sinónimos. Además de esta relación de implicación léxica o **troponimia**, los verbos en WordNet se relacionan a través de las siguientes relaciones semánticas:

- **Hiperonimia (*hyperonym*)**: el verbo V_1 es un hiperónimo del verbo V_2 si la actividad representada por V_2 es un tipo de V_1. Por ejemplo, *viajar* es un hiperónimo de *moverse*.
- **Términos coordinados (*coordinate terms*)**: el verbo V_1 es un término coordinado de V_2 si ambos verbos comparten un hiperónimo. Por ejemplo, *balbucear* y *susurrar* son términos coordinados, ya que comparten el hiperónimo *hablar*.
- **Oposición**: un ejemplo de relación de oposición sería la existente entre verbos como *atar* y *desatar*, o *aparecer* y *desaparecer*.
- **Causalidad**: la relación de causa se establece entre dos verbos que actúan, respectivamente, como causativo y resultado. Un ejemplo de esta relación lo encontraríamos en la pareja de verbos *enseñar* y *aprender*.

- **Adjetivos**: WordNet contiene aproximadamente 19.000 adjetivos, organizados en torno a unos 10.000 synsets y divididos en dos clases principales: descriptivos y relacionales. Se consideran **adjetivos descriptivos** aquellos que otorgan a los sustantivos valores de atributos bipolares, y por tanto, se organizan en base a relaciones de antonimia y sinonimia. Ejemplos claros de este tipo de adjetivos serían *bonito* o *divertido*. Por el contrario, los **adjetivos relacionales** significan algo

parecido a "relativo a" o "asociado con" el sustantivo al que se refieren, siendo la antonimia la relación básica entre estos adjetivos. Un ejemplo de este segundo tipo de adjetivos sería *dental* cuando acompaña al sustantivo *higiene*.

3.2. GATE

GATE (Generic Architecture for Text Engineering)[12] es una conocida infraestructura para el desarrollo de software de procesamiento de lenguaje natural desarrollada por la Universidad de Sheffield en 1995, y en continua evolución desde entonces. Surge con el objetivo de facilitar el trabajo de científicos y desarrolladores especificando e implementando una arquitectura para la construcción de aplicaciones de ingeniería lingüística, y proporcionando un entorno gráfico para el desarrollo de los distintos componentes que generalmente se necesitan 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, corpora u ontologías.
- **Processing Resources (PRs)**, que representan entidades que son, en su mayoría, algoritmos como analizadores, generadores, etc.
- **Visual Resources (VRs)**, que representan la visualización y edición de los componentes de la interfaz gráfica.

El conjunto de recursos integrados en GATE recibe el nombre de *CREOLE (Collection of REusable Objects for Language Engineering)*. Todos los recursos se encuentran empaquetados en archivos 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 una interfaz para Java. 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.

[12]GATE. http://gate.ac.uk/. Consultada el 1 de noviembre de 2010

Uno de los componentes más populares de GATE es *ANNIE* (*A Nearly-New Information Extraction System*). 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 3.5.

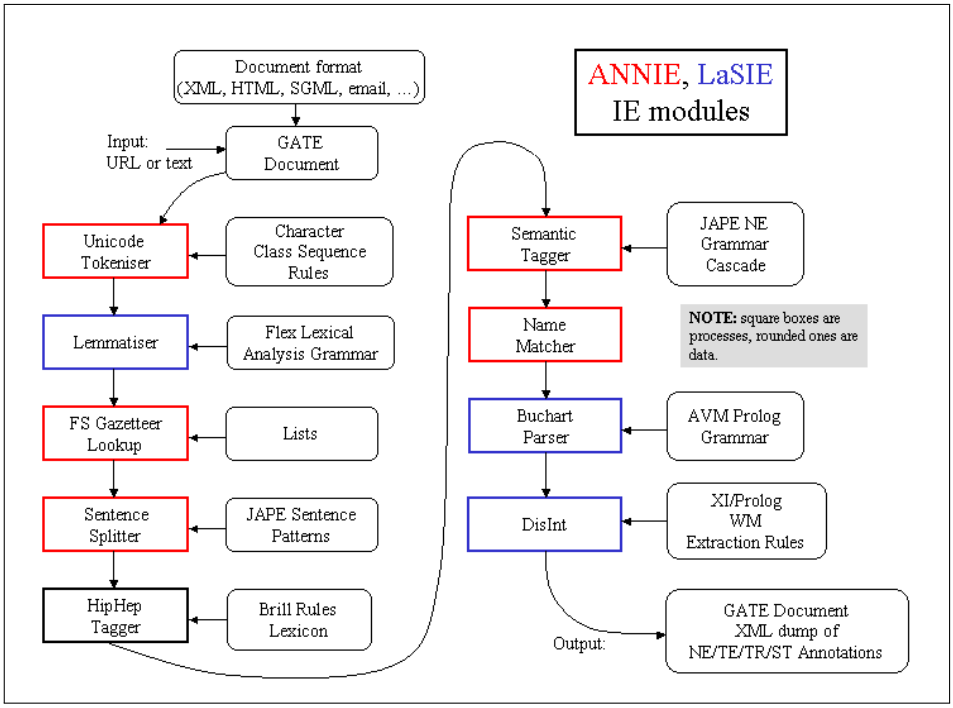


Figura 3.5: Componentes de ANNIE.
Fuente: http://gate.ac.uk/g8/page/print/2/sale/tao/. Consultada el 1 de noviembre de 2010

A continuación se proporciona una breve explicación de los módulos de ANNIE utilizados en este trabajo.

- **Tokenizer**. Realiza la división del texto en unidades simples (*tokens*) tales como números, símbolos, signos de puntuación, espacios y palabras de distintos tipos.

- **Gazetteer**. 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.

- **Sentence Splitter**. Consiste en un conjunto de transductores de estados finitos que permiten segmentar el texto en oraciones. Utiliza la lista de abreviaturas del *gazetteer* para distinguir los puntos 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 le asocia una etiqueta *Split*.

- **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. 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.

3.3. MetaMap

MetaMap[13] es un programa desarrollado por la *National Library of Medicine* de los Estados Unidos para la traducción de textos biomédicos a conceptos del Metatesauro 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 la identificación de los conceptos.

Aunque fuera inicialmente concebido como soporte en la recuperación de material bibliográfico de MEDLINE, lo cierto es que la necesidad de determinar automáticamente los conceptos inmersos en un texto 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 (Rindfleisch, Tanabe, y Weinstein, 2000).

[13] National Library of Medicine. MetaMap. http://mmtx.nlm.nih.gov/. Consultada el 1 de noviembre de 2010

- Identificación de terminología anatómica en textos médicos (Sneiderman, Rindflesch, y Bean, 1998).

- Categorización de textos biomédicos (Perea et al., 2008).

La traducción de los términos de un documento a conceptos del Metatesauro de UMLS 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 Metatesauro: *Coronary*, *Heart* y *Disease*, en lugar de indexar con el concepto único *Coronary Heart Disease*.

- Si se realiza la traducción exacta 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 tienen en cuenta coincidencias parciales o complejas. Por ejemplo, *obstructive sleep apnea* se corresponde con el término *Obstructive apnea* en el Metatesauro. Por su parte, *intensive care medicine* se corresponde con *Intensive Care* y *Medicine*.

- Por último, incluso encontrándose el término exacto en el Metatesauro, éste puede ser ambiguo y tener distintos conceptos asociados. Por ejemplo, el Metatesauro contiene dos conceptos para el término *ventilation*, uno relacionado con el flujo de aire en los edificios, y otro relacionado con la respiración. MetaMap implementa un algoritmo de desambiguación (opción *-y*) que favorece a aquellos conceptos semánticamente consistentes con el resto de conceptos en su contexto, a través de la información contenida en sus tipos semánticos (Humphrey et al., 2006). No obstante, la ambigüedad en el Metatesauro sigue siendo, a día de hoy, el talón de aquiles de UMLS. Sus autores reconocen que la

inclusión de este algoritmo de desambiguación sólo se ha traducido en ligeras mejoras, y que la resolución de esta ambigüedad es una de las principales líneas de trabajo tanto a corto como a largo plazo (Aronson y Lang, 2010).

Un estudio (Divita, Tse, y Roth, 2004) realizado para comparar la identificación de conceptos UMLS en texto libre realizada por MetaMap con la realizada por expertos en terminología médica concluye que MetaMap presenta una alta cobertura (93.3 %) y una precisión razonable (55.2 %); y que la mayoría de las divergencias se deben a la existencia de conceptos no contemplados en el Metatesauro. Hay que precisar que dicho estudio data del año 2004, y que, desde entonces, la cobertura de UMLS ha mejorado considerablemente. Desafortunadamente, no se conoce la existencia de estudios más recientes.

3.3.1. Funcionamiento del Algoritmo

El programa MetaMap acepta como entrada el texto para el que se desean conocer los conceptos de UMLS asociados. Seguidamente, para cada oración, ejecuta los siguientes pasos:

1. Se analiza 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 de inflexión y derivación, abreviaturas, acrónimos y sinónimos.

3. Se obtienen del Metatesauro los posibles candidatos, que serán aquellos términos que contienen alguna de las variantes.

4. Se utiliza una función de evaluación para calcular una medida de la fuerza o probabilidad de cada candidato, y se ordenan 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.

6. En caso de especificar que se desea utilizar la opción *-y* para desambiguar estos "mejores candidatos", se aplica el algoritmo de desambiguación descrito en (Humphrey et al., 2006) para determinar, de entre todos ellos, el candidato correcto de acuerdo con su contexto. A continuación se muestra un ejemplo del conjunto de candidatos extraídos para el sintagma *heart attack trial* (Tabla 3.7). 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 posible para el término *trial*, y *Heart Attack*, con una puntuación de 734.

Phrase: "Heart Attack Trial"
Meta Candidates (8):
827 C0008976:Trial (Clinical Trial) [Research Activity]
734 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
660 C0018787:Heart [Body Part, Organ, or Organ Component]
660 C0277793:Attack, NOS (Onset of illness) [Finding]
660 C0699795:Attack (Attack device) [Medical Device]
660 C1261512:attack (Attack behavior) [Social Behavior]
660 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
660 C1304680:Attack (Observation of attack) [Finding]
Meta Mapping (901):
734 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
827 C0008976:Trial (Clinical Trials) [Research Activity]

Tabla 3.7: Conceptos candidatos para el sintagma *heart attack trial* y selección final de los "mejores" candidatos

3.4. Personalized PageRank

El objetivo de la desambiguación léxica (*WSD*, por sus siglas en inglés) es determinar el significado correcto de una palabra polisémica de acuerdo con el contexto en el que se utiliza. Se trata de toda una disciplina dentro del campo del procesamiento de lenguaje natural habitualmente considerada como una tarea intermedia, pero indispensable, para el desarrollo de aplicaciones finales de ingeniería lingüística (Wilks y Stevenson, 1996).

Actualmente existen dos categorías principales para la clasificación de los métodos empleados en desambiguación léxica: *métodos supervisados* y *métodos no supervisados*. Aunque los primeros generalmente producen mejores resultados (Agirre y Edmonds, 2006), su principal inconveniente radica

en la necesidad de disponer de grandes colecciones de ejemplos etiquetados manualmente, cuya elaboración resulta excesivamente costosa. Por su parte, los métodos no supervisados, y en concreto aquellos conocidos como *métodos basados en conocimiento*, constituyen una buena alternativa que no requiere de datos previamente etiquetados, sino que explotan el conocimiento lingüístico presente en recursos externos (e.g. WordNet, diccionarios como el Logman o el Collins, etc.) para desambiguar las palabras.

Personalized PageRank[14] (Agirre y Soroa, 2009) es un algoritmo de desambiguación léxica basada en conocimiento desarrollado en la Universidad del País Vasco, que utiliza una base de conocimiento léxico (*Lexical Knowledge Base, LKB*) para determinar el significado adecuado de un término de acuerdo con su contexto. Como su propio nombre sugiere, utiliza el algoritmo *PageRank* (Brin y Page, 1998), diseñado para determinar la relevancia de las páginas web indexadas por un motor de búsqueda. PageRank asigna diferentes pesos a los nodos de un grafo a través del análisis de su estructura, dando preferencia a aquellos nodos enlazados a su vez por otros nodos con un peso elevado. Es decir, no sólo importa el número de enlaces que alcanzan un determinado nodo, sino también la importancia de los nodos de los que parten dichos enlaces.

Personalized PageRank modifica el algoritmo original para, utilizando WordNet como base de conocimiento léxico, crear un grafo que representa la jerarquía completa de WordNet. Una vez construido el grafo de la base de conocimiento, cuando se desea desambiguar una palabra con múltiples significados posibles en un texto, se añade un nodo a este grafo que representa a la palabra ambigua, y se enlaza con los nodos que representan todos y cada uno de sus posibles significados en WordNet. A continuación, se asignan pesos a estos nodos que se propagan a través de la red, eligiéndose el significado con mayor peso como significado desambiguado de la palabra objetivo. Este algoritmo, al que llamaremos *PPR estándar* o simplemente *PPR*, resulta ser muy eficiente, ya que permite desambiguar todas las palabras de un documento simultáneamente, pero puede ocasionar problemas si dos o más de los significados posibles de una palabra ambigua se relacionan entre sí en WordNet, ya que PageRank asignará los pesos a estos significados en lugar de transferirlos a las palabras relacionadas. Para evitar esta situa-

[14] UKB: Graph Based Word Sense Disambiguation and Similarity. http://ixa2.si.ehu.es/ukb/. Consultada el 1 de noviembre de 2010

ción, Agirre y Soroa (2009) presentan una segunda variante del algoritmo original, a la que denominan *PPR palabra-por-palabra (PPR word-to-word)*, en la que se crea un grafo por cada palabra ambigua, y no se asigna peso alguno a la palabra a desambiguar. De este modo, toda la información utilizada para ponderar los posibles significados se obtiene del resto de palabras del documento. Este algoritmo es más exacto que la versión estándar, pero menos eficiente, debido al elevado número grafos que se han de crear. Los autores demuestran que este algoritmo produce mejores resultados que otros algoritmos de desambiguación no supervisados.

Para su utilización, Personalized PageRank requiere tres recursos de entrada:

- El texto a desambiguar, en el que se ha de indicar la palabra o palabras ambiguas junto con el contexto de cada una que se desea considerar para realizar la desambiguación (generalmente, la oración en la que aparecen).
- La *base de conocimiento léxico*, que en el caso del trabajo presentado por Agirre y Soroa (2009) se corresponde con distintas versiones de WordNet.
- Un *diccionario* o lista de palabras (típicamente lemas), cada una de ellas enlazadas con al menos un concepto de la base de datos léxica.

3.5. WordNet::Similarity

WordNet::Similarity[15] (Pedersen, Patwardhan, y Michelizzi, 2004) es un paquete software desarrollado en Perl que implementa distintas métricas de similitud y relación semántica basadas en la estructura y el contenido de WordNet.

Las métricas de similitud semántica típicamente trabajan sobre pares sustantivo-sustantivo y verbo-verbo, pues sólo estas figuras sintácticas se pueden organizar en jerarquías *is a*, y cuantifican cómo de similar es un concepto (o *synset*) A a otro B. Por ejemplo, una métrica de esta categoría determinaría que un *gato* es más parecido a un *perro* que a una *mesa*, puesto que *gato* y *perro* comparten el ancestro *carnívoro* en la jerarquía de

[15] WordNet::Similarity. http://search.cpan.org/dist/WordNet-Similarity/. Consultada el 1 de noviembre de 2010

sustantivos de WordNet. Las medidas de similitud semántica se dividen, a su vez, en *medidas basadas en el contenido informativo* y *medidas basadas en la longitud del camino*. Por su parte, las métricas de relación semántica no se limitan a estas jerarquías, y por tanto se pueden aplicar sobre cualquier elemento del discurso. Tienen en cuenta otras relaciones presentes en WordNet, además de la hiperonimia, como las relaciones *has-part* (tiene parte), *is-made-of* (está hecho de) y *is-attribute-of* (es atributo de). Así, por ejemplo, un *capítulo* es parte de un *texto*, y *niñez* es un atributo de *niño*. A continuación, se describen brevemente las métricas implementadas para cada una de las categorías mencionadas.

1. **Similitud basada en el contenido informativo**. Se trata de una medida de la especificidad de un concepto, y se apoya en la idea del *Least Common Subsumer* (LCS), o del ancestro común más específico entre dos conceptos. Dentro de este subgrupo, en WordNet::Similarity se encuentran las siguientes métricas.

   - **Resnik (res)**: calcula la similitud como el contenido informativo (CI) del LCS (Resnik, 1995).
   - **Lin (lin)**: extiende la métrica de Resnik para calcular la similitud utilizando la Ecuación 3.1 (Lin, 1998).

$$Similitud = \frac{CI(LCS)}{CI(synset_1) + CI(synset_2)} \tag{3.1}$$

   - **Jiang & Conrath (jcn)**: calcula la similitud utilizando la Ecuación 3.2 (Jiang y Conrath, 1997).

$$Similitud = \frac{1}{CI(synset_1) + CI(synset_2) - 2 \times CI(LCS)} \tag{3.2}$$

2. **Similitud basada en la longitud del camino**. Se basa en el cálculo de la distancia entre dos conceptos, medida en términos del número de nodos o aristas que los separan y, en ciertos casos, de la profundidad de ambos conceptos en la jerarquía.

   - **Path Length**: calcula la similitud como la inversa del número de nodos presentes en el camino mínimo entre ambos synsets.

- **Wu & Palmer (wup)**: considera la profundidad de los dos synsets en la taxonomía de WordNet, así como la profundidad del LCS, y las relaciona mediante la Ecuación 3.3 (Wu y Palmer, 1994).

$$Similitud = \frac{2 \times depth(LCS)}{depth(synset_1) + depth(synset_2)} \tag{3.3}$$

- **Leacock & Chodorow (lch)**: calcula la similitud utilizando la Ecuación 3.4 (Leacock y Chodorow, 1998), donde *length* es la longitud del camino mínimo entre ambos synsets y D es la profundidad máxima de la taxonomía.

$$Similitud = -\log \frac{length}{2 \times D} \tag{3.4}$$

3. **Similitud basada en la relación semántica**. El concepto de relación semántica es mucho más amplio que el de similitud. Una medida de relación semántica cuantifica la fuerza de la relación entre los significados de dos palabras.

   - **Extended Gloss Overlaps (Adapted Lesk)**: determina la relación entre dos conceptos a partir del número de superposiciones de palabras entre sus *glosses* o definiciones en WordNet y de los enlaces directos entre ellas en la red. La similitud final se calcula como la suma de los cuadrados de las longitudes (en número de palabras) de las superposiciones (Lesk, 1986).
   - **Context Vectors (vector)**: utiliza información sobre la co-ocurrencia de términos en un corpus para obtener una medida de la relación entre dos conceptos. Para ello, construye una matriz de co-ocurrencias para cada una de las palabras presentes en las definiciones de los conceptos, y representa cada acepción con un vector que es la media de estos vectores de co-ocurrencia. La similitud entre dos conceptos se determina como el coseno del ángulo entre sus vectores (Patwardhan, 2003).
   - **Hirst & St-Onge (hso)**: se basa en la longitud del camino, y considera que las relaciones entre conceptos en WordNet tienen una determinada dirección. Por ejemplo, la relación *is a* es ascendente, mientras que la relación *has-part* es horizontal. Para

calcular la similitud entre dos conceptos, intenta encontrar un camino entre ellos que cumpla dos condiciones: no ser demasiado largo y no cambiar de dirección con demasiada frecuencia (Hirst y St Onge, 1998).

WordNet::Similarity puede utilizarse desde su interfaz web[16] o bien a través de la línea de comandos, invocando al programa *similarity.pl* y especificando la métrica de similitud y los conceptos sobre los que se desea calcular. Por ejemplo, en la Tabla 3.8 se observa cómo la primera sentencia ejecuta la métrica *Lesk* para el primer significado de *car* y el primer significado de *bike*. La segunda sentencia ejecuta la misma métrica pero, al no especificar los significados de los conceptos, devuelve la mayor similitud encontrada entre los distintos significados de ambos conceptos. Finalmente, la tercera sentencia devuelve la similitud entre todos los posibles significados de *car* y *bike*.

```
> similarity.pl --type WordNet::Similarity::vector car#n#1 bike#n#1
car#n#1 bike#n#1 0.756395613030899
> similarity.pl --type WordNet::Similarity::vector car#n bike#n
car#n#1 bike#n#2 0.90618485736825
> similarity.pl --type WordNet::Similarity::vector -allsenses car#n bike#n
car#n#1 bike#n#2 0.90618485736825
car#n#2 bike#n#2 0.877783971247943
car#n#1 bike#n#1 0.756395613030899
car#n#2 bike#n#1 0.701289428083391
car#n#5 bike#n#2 0.621345209159249
car#n#5 bike#n#1 0.50292923868496
car#n#4 bike#n#2 0.460159722727335
car#n#4 bike#n#1 0.407797080123658
car#n#3 bike#n#2 0.395296214612913
car#n#3 bike#n#1 0.342766679044078
```

Tabla 3.8: Ejecución de WordNet::Similarity desde la línea de comandos

3.6. WordNet::SenseRelate

WordNet::SenseRelate[17] es un paquete Perl que aborda el problema de la desambiguación léxica, permitiendo identificar los significados adecuados de

[16] WordNet::Similarity. http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi. Consultada el 1 de noviembre de 2010

[17] WordNet::SenseRelate. http://www.d.umn.edu/~tpederse/senserelate.html. Consultada el 1 de noviembre de 2010

las palabras ambiguas de un texto de entre sus acepciones candidatas en WordNet. WordNet::SenseRelate utiliza las medidas de relación y similitud semántica definidas en el paquete WordNet::Similarity (Sección 3.5), y presenta dos algoritmos de desambiguación:

- ***WordNet::SenseRelate::AllWords***: concebido para desambiguar todas y cada una de las palabras presentes en un texto.
- ***WordNet::SenseRelate::TargetWord***: pensado para asignar el sentido adecuado a una determinada palabra.

WordNet::SenseRelate::AllWords acepta como entrada el documento que se desea procesar y un listado de parámetros que establecen las distintas opciones de ejecución. A continuación, elimina todas aquellas palabras que no se encuentran en WordNet, y aquellas que aparecen en una lista de parada (por defecto o proporcionada por el usuario). Seguidamente, para cada una de las oraciones del texto, ejecuta el algoritmo de desambiguación *Maximum Relatedness Disambiguation*, descrito en Patwardhan *et al.* (2005).

WordNet::SenseRelate puede utilizarse desde la línea de comandos, invocando al programa *wsd.pl*. Como parámetros, se han de especificar los siguientes: el nombre del fichero que almacena el texto a desambiguar, el formato de dicho texto, la métrica de similitud que se desea utilizar, el tamaño (en número de palabras) de la ventana de contexto y, si se desea, una lista de palabras vacías. Como resultado, se obtiene un listado de los conceptos asociados a cada una de las oraciones del documento, así como su categoría gramatical. La Tabla 3.9 muestra el resultado de ejecutar WordNet::SenseRelate, utilizando el algoritmo Lesk como métrica de similitud semántica, para desambiguar el significado de todas las palabras de la oración *The red car is parked near the supermarket.*

```
> wsd.pl --type WordNet::Similarity::lesk --context sentencesFile
--format tagged --stoplist config/SRStopWord.txt
The red#n#4 car#n#1 be#v#1 parked#a#1 near#a#2 the supermarket#n#1
```

Tabla 3.9: Ejecución de WordNet::SenseRelate desde la línea de comandos para la oración *The red car is parked near the supermarket*

Capítulo 4

Uso de Grafos Semánticos para la Generación Automática de Resúmenes

En esta sección se presenta un método genérico para la realización de resúmenes de texto. El algoritmo diseñado permite generar resúmenes de documentos de distintos tipos, tanto en cuanto a su estructura como a su temática o dominio, sin más que disponer de una base de conocimiento (BC) que cubra satisfactoriamente el vocabulario de los documentos a resumir, y de un método para desambiguar el significado de los términos y traducirlos a conceptos de la base de conocimiento. Permite, además, generar resúmenes individuales a partir de múltiples documentos sobre un mismo tema, simplemente proporcionándole un método de detección y eliminación de redundancia.

El algoritmo de generación de resúmenes implementado se basa en la representación del documento como un grafo de conceptos del dominio y en la identificación de los conceptos centrales en este grafo como paso previo a la extracción de las oraciones para el resumen. Esta representación conceptual permite capturar las relaciones semánticas entre los elementos textuales y determinar con mayor precisión el tema principal del documento, lo que a su vez permite construir resúmenes de mayor calidad.

La Figura 4.1 muestra la arquitectura general del método propuesto para la generación de resúmenes mono-documento. En ella se distinguen siete etapas secuenciales. 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, y tras acometer un proceso de desambiguación léxica (WSD), serán representadas como grafos de conceptos utilizando la base de conocimiento del dominio. Seguidamente, los distintos grafos de las oraciones se combinan en un único grafo que representa el documento. A continuación, se aplica un algoritmo de agrupamiento o *clustering* basado en la detección de vértices concentradores o *hubs*, con el objetivo de identificar conjuntos de conceptos estrechamente relacionados que delimitan los distintos temas tratados en el texto. Finalmente, y en función de la similitud de las oraciones con respecto a cada uno de los temas, se seleccionan las oraciones que conforman el resumen. Cada una de estas etapas se describe detalladamente en las siguientes secciones.

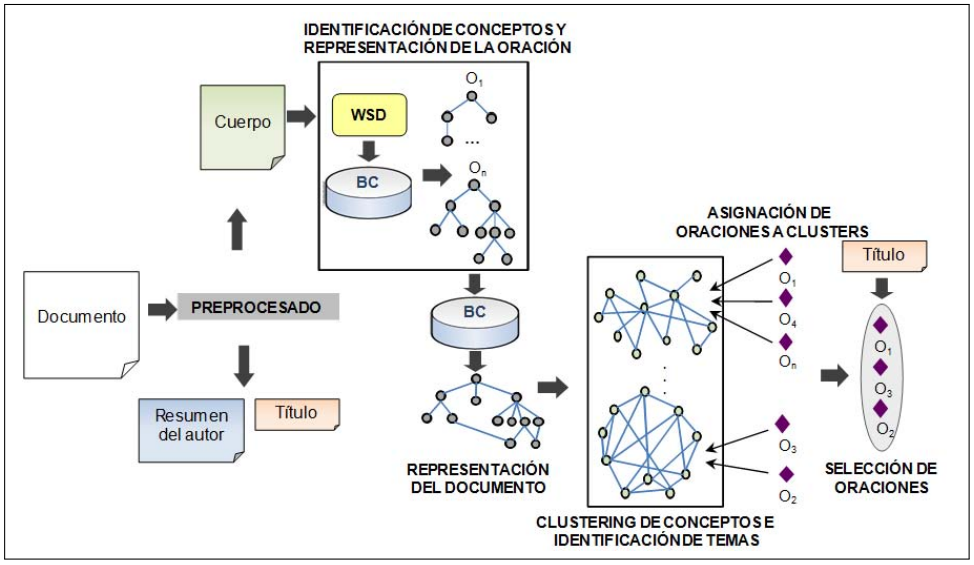


Figura 4.1: Arquitectura del método de generación automática de resúmenes

4.1. Etapa I: Pre-procesamiento

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. En concreto, en esta fase se realizan las siguientes acciones sobre el documento:

1. Se eliminan aquellas secciones que no intervienen en la generación del resumen. Puesto que tales secciones dependerán del dominio y del tipo

concreto del documento, el sistema permite especificar estas secciones "irrelevantes" en un archivo XML de configuración.

2. Se eliminan aquellos términos que, por tener un significado muy general, no serán de utilidad a la hora de discriminar entre oraciones relevantes e irrelevantes para el resumen. Tendrían esta consideración, entre otros, los términos pertenecientes a las categorías gramaticales de preposición, artículo o conjunción. Para ello, se utiliza una lista de parada (*stop list*) que deberá ser elaborada teniendo en consideración las singularidades del dominio de trabajo.

3. En el caso de que el documento incluya un título y un resumen o *abstract* elaborado por el propio autor, se extrae el contenido de ambas secciones, separándolas del cuerpo del documento.

4. Finalmente, el cuerpo del documento se divide en oraciones. Para ello, se utiliza la librería GATE (Sección 3.2), invocando a los componentes de ANNIE desde el API correspondiente, de manera transparente al usuario. Seguidamente se describe brevemente el proceso realizado:

   - Iniciar GATE y cargar los *plugins* necesarios; en este caso, únicamente ANNIE.
   - Crear un corpus que encapsule los documentos que se desean procesar.
   - Establecer los módulos o *resources* que se van a utilizar en el procesamiento, en el orden en el que han de ser aplicados. En nuestro caso, los recursos requeridos son *DefaultTokenizer*, *DefaultGazetteer* y *SentenceSplitter*.
   - Ejecutar los módulos anteriores sobre el corpus generado.
   - Como resultado de la ejecución se generará un documento XML por cada uno de los textos del corpus, en el que se habrán identificado las diferentes oraciones que los componen.

4.2. Etapa II: Traducción de las Oraciones a Conceptos del Dominio

El objetivo de esta etapa es traducir el léxico del documento a conceptos de la base de conocimiento del dominio. Para ello, es preciso disponer de

algún mecanismo de desambiguación (*Word Sense Disambiguation, WSD*) que permita discernir, de entre los posibles significados de un término, su significado preciso de acuerdo con el contexto en el que se halla inmerso (Navigli, 2009). En este sentido, tanto el algoritmo concreto de desambiguación como los recursos lingüísticos y semánticos a utilizar dependerán en gran medida del dominio al cual pertenece el texto que se desea desambiguar y de la base de conocimiento utilizada.

De este modo, cada oración del texto de entrada es analizada por el algoritmo de desambiguación y, como resultado, se obtiene de la base de conocimiento una lista de conceptos o significados. La Figura 4.2 ilustra el proceso descrito. El número de conceptos identificados dependerá del grado en que la base de conocimiento elegida cubra el vocabulario del dominio. Puede suceder, por tanto, que existan términos en la oración que no se correspondan con ningún concepto, que distintos términos se identifiquen con un mismo concepto, o incluso que varios términos se representen mediante un único concepto en la base de conocimiento.

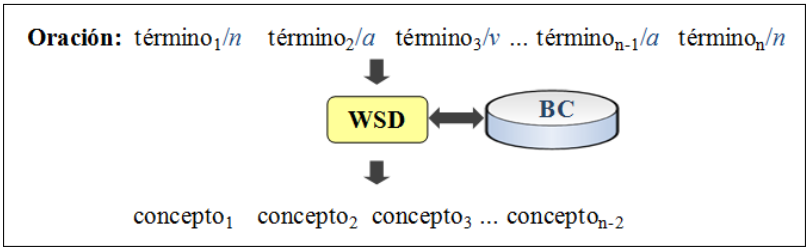


Figura 4.2: Traducción de una oración a conceptos de la base de conocimiento

Finalmente, dependiendo del dominio y con la información que proporcione la base de conocimiento correspondiente, puede interesar realizar un segundo filtrado que permita eliminar aquellos términos que, por su significado o categoría gramatical, se consideren excesivamente generales, y que, por tanto, no aportan ninguna información para la generación del resumen.

4.3. Etapa III: Representación de las Oraciones como Grafos de Conceptos

El objetivo de esta etapa es construir una representación en forma de grafo para cada oración del documento, de manera que capture su estructura semántica y las relaciones entre sus términos. Para ello, los conceptos descu-

biertos para cada oración en la etapa anterior se expanden con los conceptos de niveles superiores en la jerarquía de la base de conocimiento (*hiperónimos* o *super-categorías*). Por lo tanto, será requisito indispensable para el funcionamiento del algoritmo disponer de una base de conocimiento que contemple la relación de hiperonimia entre sus conceptos.

A continuación, y una vez expandidos todos los conceptos con sus hiperónimos, las distintas jerarquías se combinan en un único grafo que representa a la oración. En este grafo, cada vértice representa un concepto distinto (es decir, si dos términos de la oración se corresponden con el mismo concepto, sólo se crea un vértice en el grafo que representa a ambos términos), mientras que las aristas, temporalmente sin etiquetar, representan relaciones semánticas (hiperonimia, generalización o relaciones *es un*) entre dichos conceptos.

Finalmente, bajo la hipótesis de que los conceptos que se encuentran en los niveles superiores de la jerarquía representan información muy genérica, se eliminan del grafo los n niveles superiores. El valor de n deberá ser determinado empíricamente y dependerá del dominio de trabajo. La Figura 4.3 muestra el aspecto final que presentaría el grafo de una oración. En esta figura, los conceptos "ignorados" por ser excesivamente genéricos se muestran con una tonalidad más clara que el resto de conceptos. Se puede comprobar, además, que el grafo de una oración no ha de ser necesariamente conexo.

4.4. 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 las relaciones semánticas que se establecen entre los conceptos de todo el documento. Este grafo se puede extender con nuevas y más específicas relaciones (además de la relación de generalización o hiperonimia), para obtener una representación más completa del documento. En concreto, en el trabajo que aquí se presenta las relaciones semánticas pueden ser configuradas atendiendo al dominio de aplicación y a las posibilidades que ofrezca la base de conocimiento.

A continuación, se asigna un peso a cada una de las aristas del grafo del documento. Para el cálculo de estos pesos, el sistema permite utilizar distintas métricas de similitud entre conjuntos, a la vez que permite la definición

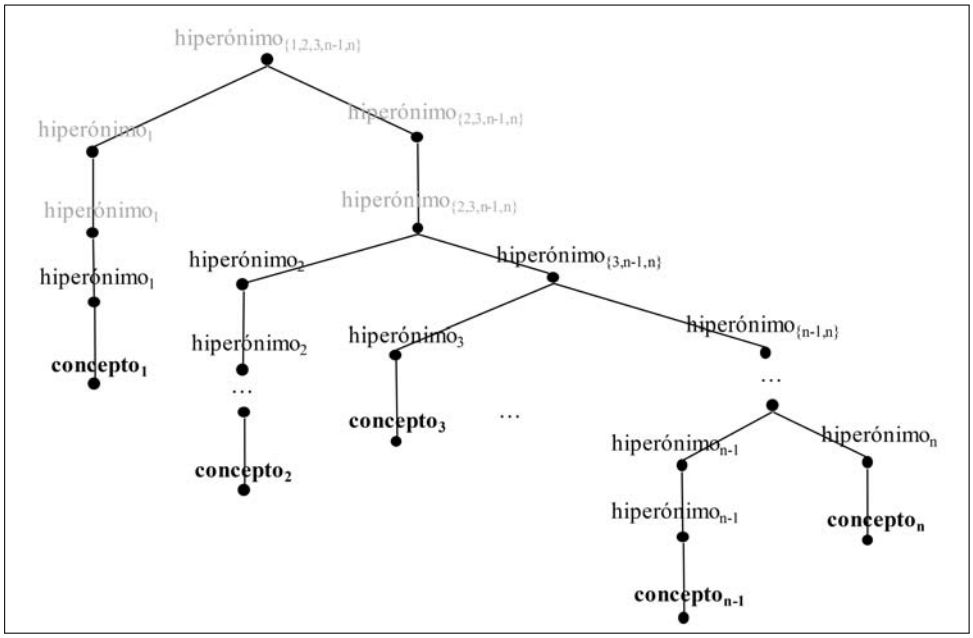


Figura 4.3: Grafo semántico de una oración

de nuevas métricas de una forma sencilla. No obstante, en este trabajo se han implementado y evaluado las siguientes dos métricas basadas, respectivamente, en el coeficiente de similitud de Jaccard y en el coeficiente de similitud de Dice-Sorensen. Nótese que ambas métricas premian a las relaciones que conectan conceptos específicos frente a las que conectan conceptos más generales. Ambos coeficientes se explican a continuación.

- El **coeficiente de similitud de Jaccard** (Jaccard, 1901). Según esta métrica, la similitud entre dos conjuntos se define como el tamaño de la intersección dividido entre el tamaño de la unión de tales conjuntos. Este índice proporciona siempre un valor entre 0 y 1. Aplicando esta definición al problema que nos ocupa, el peso de una arista que conecta dos conceptos no finales, A y B, se calcula utilizando la Ecuación 4.1, donde α representa el conjunto de todos los ancestros del concepto A, incluido el propio concepto, y β representa el conjunto de todos los ancestros del concepto B, incluido B. Para las aristas que conectan dos conceptos finales (i.e. nodos hoja), el peso asignado es siempre '1'.

$$peso(A, B) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \tag{4.1}$$

La Figura 4.4 muestra el grafo de un documento genérico donde las aristas han sido etiquetadas utilizando el índice de Jaccard. En este grafo, los distintos tipos de aristas simbolizan distintos tipos de relaciones semánticas entre los conceptos.

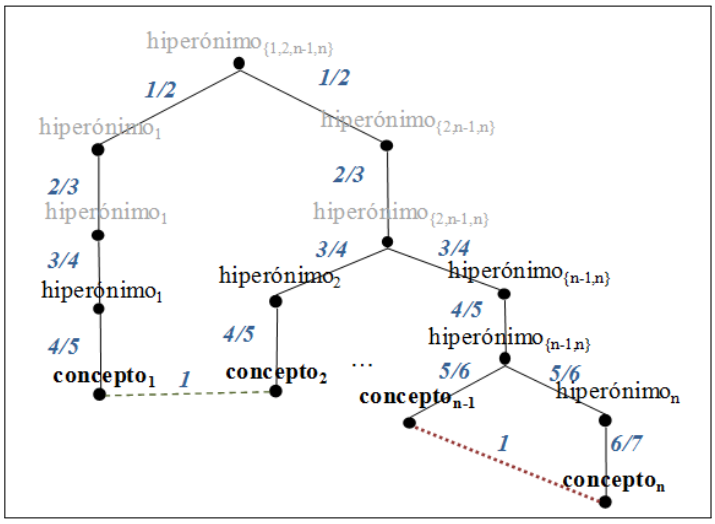


Figura 4.4: Grafo semántico de un documento cuyas aristas han sido etiquetadas utilizando el coeficiente de Jaccard

- El **coeficiente de similitud de Dice-Sorensen** (Legendre y Legendre, 1998) puede considerarse una variación del coeficiente de Jaccard que otorga doble importancia a los elementos que están presentes en ambos conjuntos. El resultado es también un valor entre 0 y 1. Siguiendo con la terminología utilizada para definir el coeficiente de Jaccard, la Ecuación 4.2 permite calcular el índice de Dice-Sorensen. De nuevo, esta definición sólo se aplica a las aristas que representan relaciones entre conceptos en los que al menos uno de ellos no está representado por un nodo hoja, mientras que las aristas que conectan dos nodos hoja reciben un peso de '1'.

$$peso(A, B) = \frac{2 \times |\alpha \cap \beta|}{2 \times |\alpha \cap \beta| + |\alpha - \beta|} \tag{4.2}$$

La Figura 4.5 muestra el grafo de un documento genérico donde las aristas han sido etiquetadas utilizando el índice de Dice-Sorensen.

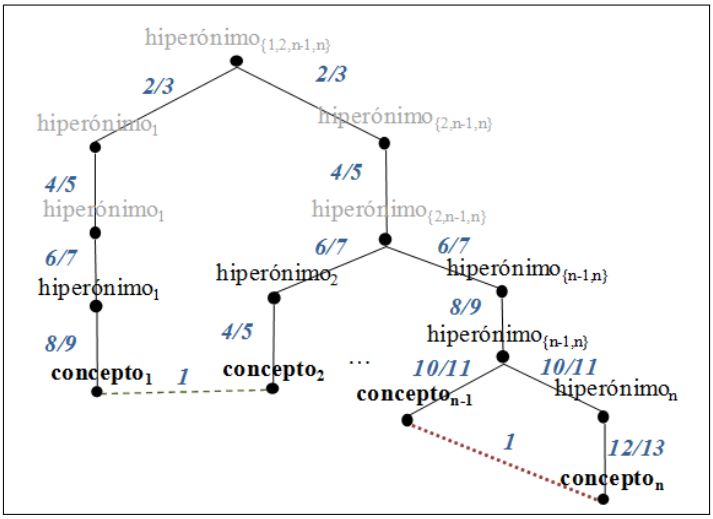


Figura 4.5: Grafo semántico de un documento cuyas aristas han sido etiquetadas utilizando el coeficiente de Dice-Sorensen

4.5. Etapa V: Clustering de Conceptos e Identificación de Temas

El propósito de esta etapa es agrupar los conceptos del grafo del documento, utilizando para ello un algoritmo de agrupamiento (*clustering*) basado en la conectividad (*degree-based method*), similar al propuesto en Erkan y Radev (2004b). Con ello se pretende identificar grupos de conceptos que mantengan una estrecha relación semántica entre sí; es decir, un significado similar o de algún modo relacionado. La hipótesis de partida es que cada uno de estos conjuntos debe representar un tema distinto del documento; y que, dentro de ellos, los conceptos centrales (*centroides*) aportan la información necesaria y suficiente sobre cada tema.

Ferrer-Cancho y Solé (2001) han demostrado que los grafos que representan la relación entre las palabras en los textos en inglés forman una *red libre de escala* (Barabási y Albert, 1999). Se parte, por lo tanto, de la hipótesis de que el grafo obtenido para el documento a resumir forma una red de este tipo. Una red libre de escala (*scale-free network*) es un tipo específico de red compleja en la que unos pocos nodos están altamente conectados entre sí (nodos *hub*), mientras que el grado de conexión de los restantes nodos es relativamente bajo. Esta propiedad surge como consecuencia de dos características que pueden observarse fácilmente en las redes reales: en pri-

mer 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 $P(K) = k^{-\gamma}$, 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.

Asumiendo pues que el grafo del documento se comporta como una red libre de escala, el algoritmo de agrupamiento comienza localizando en este grafo el conjunto de nodos más conectados. De este modo, 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 4.3.

$$salience(v_i) = \sum_{\substack{\forall e_j | \exists v_k \\ \wedge e_j conecta(v_i, v_k)}} peso(e_j) \tag{4.3}$$

Los n vértices de mayor *salience* reciben el nombre de *hub vertices* o vértices concentradores, y representan los nodos más conectados del grafo, tanto en relación al número de aristas como al peso de las mismas. El valor del parámetro n deberá ser determinado empíricamente, ya que depende del dominio y de la longitud de los documentos a resumir (pues ambos afectan a la estructura y a la conectividad del grafo). En general, será un valor entre el 2 % y el 20 % del número total de vértices del grafo.

A continuación, los *hub vertices* se agrupan formando *Hub Vertex Sets (HVS)* o conjuntos de vértices concentradores, que son grupos de vértices fuertemente conectados y que constituirán los centroides de los clusters a construir. Para ello, en una primera etapa, el algoritmo propuesto busca iterativamente para cada *hub vertex*, 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*, 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 (Ecuaciones 4.4 y 4.5).

$$Intra - conectividad(HVS_i) = \sum_{\substack{\forall e_j | \exists v,w \in HVS_i \\ \wedge e_j conecta(v,w)}} peso(e_j) \tag{4.4}$$

$$Inter - conectividad(HVS_i, HVS_j) = \sum_{\substack{\forall e_j | \exists v \in HVS_i, w \in HVS_j \\ \wedge e_k conecta(v,w)}} peso(e_k) \tag{4.5}$$

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 grupos de conceptos finales. La asignación se realiza calculando el grado de conexión del concepto a asignar con respecto a cada *HVS*, según la Ecuación (4.6), reajustando los *HVS* y los vértices asignados en un proceso iterativo.

$$conectividad(v, HVS_i) = \sum_{\substack{\forall e_j | \exists w \in HVS_i \\ \wedge e_j conecta(v,w)}} peso(e_j) \tag{4.6}$$

4.6. Etapa VI: Asignación de Oraciones a Clusters

Una vez han sido creados los grupos de conceptos, el siguiente paso consiste en 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 (e.g. la distancia de Levenshtein o distancia de edición (Levenshtein, 1966)) no resultan adecuadas. En su lugar, se utiliza un mecanismo de votos (Yoo, Hu, y Song, 2007), por el que cada vértice (v_k) de una oración (O_j) asigna a cada cluster (C_i) en el que se encuentra una puntuación ($w_{k,j}$) que será distinta dependiendo de si pertenece o no al *HVS*

de dicho cluster (Ecuación 4.7).

$$similitud(C_i, O_j) = \sum_{v_k | v_k \in O_j} w_{k,j} \tag{4.7}$$

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

Los valores de γ y δ en la Ecuación 4.7 se han establecido empíricamente a 1.0 y 0.5 respectivamente, lo que significa que se atribuye doble importancia a los conceptos que pertenecen a los *HVS* que a los restantes.

4.7. Etapa VII: Selección de Oraciones para el Resumen

El último paso del algoritmo consiste en extraer las N oraciones del texto original que constituirán el resumen, en función de su distancia semántica respecto a los distintos clusters, calculada de acuerdo con la Ecuación 4.7. Aunque el tamaño del resumen dependerá de las características del texto a resumir y del uso deseado del mismo, es una afirmación generalmente aceptada que la extensión idónea debería oscilar entre un 15 % y un 35 % del texto de referencia (Hovy, 2005). El sistema permite especificar este ratio tanto en porcentaje sobre el número de oraciones como en número de palabras.

En cualquier caso, y con independencia del número de oraciones a extraer, se han investigado tres heurísticas para la selección de estas oraciones, en función del tipo de resumen que se desee generar:

- **Heurística 1:** Bajo la hipótesis de que el cluster de mayor tamaño representa el tema principal del documento, y por tanto, es el único que debería tenerse en cuenta para la generación del resumen, se seleccionan de este cluster las N oraciones con las que presenta una mayor similitud. De esta forma, se pretende que el resumen sólo incluya información acerca del tema principal del documento.

- **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. Con esta heurística se pretende que el resumen incluya información acerca de todos los temas tratados en el documento, independientemente de su importancia relativa.

- **Heurística 3:** Para cada oración, se calcula una única puntuación, como la suma de sus similitudes respecto a cada uno de los clusters promediados por su tamaño (Ecuación 4.8). Finalmente, se seleccionan las N oraciones con mayor puntuación global. De esta forma, aunque casi toda la información del resumen se toma del tema principal del documento, también se permite incluir cierto grado de información secundaria o complementaria.

$$Puntuación(O_j) = \sum_{C_i} \frac{similitud(C_i, O_j)}{|C_i|} \tag{4.8}$$

Además de estos criterios, que podríamos llamar de *similitud grafo semántica (GrSem)*, el sistema implementa dos de los criterios tradicionales para la selección de oraciones en la generación de resúmenes por extracción: el *criterio posicional* y el *criterio de similitud con el título del documento*. A pesar de su simplicidad, ambos criterios continúan siendo muy utilizados incluso en los trabajos más recientes sobre generación de resúmenes (Bossard, Généreux, y Poibeau, 2008; Bawakid y Oussalah, 2008).

- **Criterio posicional:** La posición de las oraciones en el documento ha sido considerada tradicionalmente como un factor importante a la hora de determinar qué oraciones son las más relacionadas con el tema principal del documento. Aunque se trata de un criterio muy dependiente del tipo de documento, en general se observa que las oraciones situadas al inicio y al final del documento condensan la información relevante y que como tal, debería formar parte del resumen (Brandow, Mitze, y Rau, 1995; Bossard, Généreux, y Poibeau, 2008; Bawakid y Oussalah, 2008). Siguiendo este razonamiento, este criterio asigna una mayor peso a las oraciones cercanas al inicio y final del documento. En concreto, en este trabajo, se calcula una puntuación $Posición \in \{0, 1\}$ para cada oración, de acuerdo con la Ecuación 4.9.

$$Posición(O_j) = max\{\frac{1}{n_j}, \frac{1}{N - n_j + 1}\} \tag{4.9}$$

- **Similitud con el título:** De acuerdo con este criterio, el título que el propio autor asigna al documento comprende la información más significativa del mismo, y por ello se utiliza como referencia para cuantificar la importancia del resto de oraciones del documento (Bawakid y Oussalah, 2008). En este trabajo, la similitud de las oraciones con respecto al título se calcula como el ratio entre el número de conceptos que la oración y el título tienen en común, y el número de total de conceptos en ambos, de acuerdo con la Ecuación 4.10.

$$Título(O_j) = \frac{Conceptos_{O_j} \cap Conceptos_{título}}{Conceptos_{O_j} \cup Conceptos_{título}} \tag{4.10}$$

Habiendo implementado estos tres criterios de relevancia, la selección final de las oraciones para el resumen se basa en la suma ponderada de las puntuaciones asignadas a cada oración según cada uno de los criterios (Ecuación 4.11). De nuevo, los valores de λ, θ y χ en esta ecuación deberán ser ajustados experimentalmente, pues dependen en gran medida del dominio y del tipo de documentos que se desea resumir.

$$Puntuación(O_j) = \lambda \times GrSem(O_j) + \theta \times Posición(O_j) + \chi \times Título(O_j) \tag{4.11}$$

4.8. Generación de Resúmenes Multi-documento

Tal y como se adelantara en la introducción de este capítulo, el método diseñado puede ser utilizado para generar resúmenes a partir de múltiples documentos sobre un mismo tema o suceso. Para ello, añadimos dos nuevos módulos, un *módulo de integración* y un *módulo de eliminación de redundancia*, a la arquitectura mostrada en la Figura 4.1. De este modo, la arquitectura modificada para contemplar la generación de resúmenes multi-documento se muestra en la Figura 4.6. El proceso realizado para generar el resumen multi-documento consta de las tres siguientes etapas:

- **Módulo de integración**: Dado un directorio que almacena el conjunto de documentos a partir de los cuales se desea generar el resumen, este módulo realiza las siguientes acciones:

    1. Si los documentos incluyen un título, se extrae su contenido y se concatena, creando un único título que representa al conjunto de

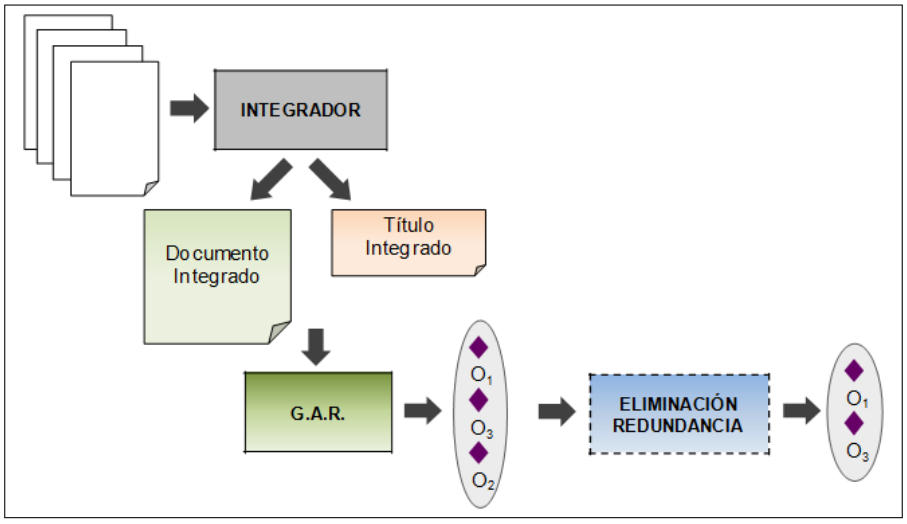


Figura 4.6: Arquitectura del método de generación de resúmenes multi-documento

documentos, y del que se elimina la información repetida.

2. Se extrae el contenido del cuerpo de los distintos documentos.
3. Se concatena la información extraída del cuerpo de los documentos y se crea un nuevo documento "integrado" que recopila toda esta información.

- **Generación del resumen**: El documento y el título producidos por el módulo de integración se utilizan como entrada al generador de resúmenes, como si de un resumen mono-documento se tratase.

- **Módulo de eliminación de redundancia**: Sobre el resumen generado en la etapa anterior, se ejecuta un algoritmo de detección de redundancia. En la Sección 2.4.1 vimos cómo, al proceder la información de distintas fuentes y versar todas ellas sobre un mismo tema, el resumen generado a menudo presenta información repetida, lo que no es, obviamente, una característica deseable en un resumen. Para eliminar estas repeticiones, el sistema implementado no incorpora ningún mecanismo de eliminación de redundancia, si bien su desarrollo se contempla como trabajo futuro, aunque es posible invocar alguno de los muchos algoritmos disponibles para esta tarea, como por ejemplo, el presentado en Ferrández *et al.* (2007). Como resultado, se obtiene un nuevo resumen del que ha sido eliminada la información redundante.

Capítulo 5

Caso de Estudio: Resúmenes Mono-documento de Artículos Científicos de Biomedicina

El primer caso de estudio tiene como objetivo configurar el método genérico propuesto en el capítulo previo para generar resúmenes de artículos científicos de biomedicina. Para ello, es necesario especializar todos aquellos aspectos que, en cada una de las etapas del algoritmo, dependen del dominio y del tipo de documento a resumir.

El presente capítulo se organiza como sigue. En primer lugar, se presenta un breve estudio de las características que hacen a los artículos científicos en biomedicina especialmente interesantes desde el punto de vista de los sistemas de acceso a la información en general, y de los sistemas de generación de resúmenes en particular. En segundo lugar, se describe en detalle el proceso realizado para configurar el método genérico para elaborar resúmenes de textos biomédicos. Con el objetivo de clarificar este proceso, a lo largo de la exposición se hará referencia a un documento concreto del corpus de BioMed Central[1], cuyo resumen será elaborado progresivamente a través de las distintas etapas del algoritmo.

[1]BioMed Central Corpus. http://www.biomedcentral.com/info/about/datamining/. Consultada el 1 de noviembre de 2010

5.1. Peculiaridades del Dominio y del Tipo de Documentos

Los textos sobre biomedicina presentan ciertas características que los diferencian de los textos de otras disciplinas. En primer lugar, la información biomédica se presenta bajo la forma de muy diversos tipos de documentos, desde historiales clínicos hasta artículos científicos, pasando por bases de datos semiestructuradas, imágenes de rayos X e incluso vídeos (Afantenos, Karkaletsis, y Stamatopoulos, 2005). Cada uno de estos tipos de documentos presenta características que lo distingue del resto y que deben ser consideradas en el proceso de generación del resumen. En este trabajo, el interés se centra en los artículos científicos, una categoría de documentos que, si bien están compuestos principalmente por texto, frecuentemente contienen información en otros formatos, como tablas o imágenes. Los artículos biomédicos que presentan resultados experimentales generalmente se encuentran estructurados en los mismos o similares apartados: *Introducción*, *Métodos*, *Resultados* y *Discusión*, la denominada estructura IMRAD (*Introduction, Method, Results* And *Discussion*). Además, frecuentemente presentan otras secciones, como *Abreviaturas*, *Limitaciones del estudio* y *Conflicto de intereses*. En la mayoría de los casos, y dependiendo del uso final deseado de los resúmenes, este conocimiento acerca de la estructura de los documentos puede contribuir a mejorar la calidad de los resúmenes automáticos.

En segundo lugar, las peculiaridades de la terminología y del estilo de escritura empleado por los profesionales en medicina hacen de la detección de conceptos y del análisis de la información una tarea muy compleja y ambiciosa (Nadkarni, 2000). El primer desafío es el problema de los **sinónimos** (utilización de diferentes términos para designar un mismo concepto) y los **homónimos** (uso de términos con múltiples significados). Por ejemplo, los términos *infarto* y *paro cardíaco* comparten el mismo significado; mientras que el término *anestesia* puede hacer referencia tanto a la pérdida de la sensibilidad dolorosa como al procedimiento o fármaco utilizado para inducirla. En relación al fenómeno de la sinonimia, su solución es relativamente sencilla si se utiliza alguna ontología o terminología como UMLS, ya que en ellas se espera que los términos sinónimos se presenten asociados a un mismo concepto. Por el contrario, la solución al problema de la homonimia es más compleja, ya que requiere del uso de algún algoritmo de desambiguación que

asigne a cada término el significado pertinente en función del contexto en el que se utiliza.

Otro problema al que se enfrentan las aplicaciones informáticas al tratar con textos médicos es la frecuente presencia de **elisiones**, **neologismos** y **abreviaciones**. Una elisión es una oración en la que se ha suprimido algún elemento del discurso, sin que ello conlleve consecuencias gramaticales (por ejemplo, en la oración "la cuenta blanca fue de 1800", la expresión *cuenta blanca* hace referencia al recuento de glóbulos blancos). Un neologismo es un nuevo vocablo o término que no se espera se encuentre en diccionarios o terminologías (por ejemplo, los términos *positividad* o *sobrecrecimiento*). Finalmente, una abreviación es una reducción de una palabra, mediante la supresión de determinadas letras o sílabas, que puede crear confusión a la hora de reconocer automáticamente los conceptos inmersos en el texto (por ejemplo, *adenoca* se utiliza con frecuencia en lugar de adenocarcinoma; mientras que *AH* se utiliza habitualmente en lugar de ácido hialurónico). De nuevo, el uso de terminologías, en particular UMLS, es de gran utilidad a la hora de resolver estos fenómenos lingüísticos, ya que contiene numerosas entradas de elisiones, neologismos y abreviaciones que, además, se encuentran en continua actualización.

5.2. Especialización del Método para el Dominio Biomédico

El objetivo de esta sección es detallar, para cada una de las etapas en que se subdivide el algoritmo de generación de resúmenes, el proceso y los recursos necesarios para adaptar el método genérico al dominio biomédico. Con el objetivo de clarificar el funcionamiento del algoritmo, a lo largo de la exposición se hará referencia a un documento concreto del corpus de BioMed Central. El artículo completo presenta un total de 58 oraciones, y puede encontrarse en el Apéndice B.1 de este documento.

5.2.1. Etapa I: Pre-procesamiento

En primer lugar, y como parte del pre-procesamiento del documento (Sección 4.1), se han considerado innecesarias para la realización del resumen las siguientes secciones: *Autores*, *Instituciones*, *Publicación*, *Año*, *ISSN*, *Volumen*, *Número*, *Url*, *Conflicto de intereses*, *Agradecimientos*, *Contribuciones*

y *Referencias*. Por lo tanto, si alguna de dichas secciones se encuentra presente en el documento, simplemente se elimina. También se eliminan las cabeceras o títulos de las diferentes secciones en que se divide el documento, y se extraen las tablas y figuras.

Por otra parte, los artículos científicos en este dominio suelen presentar una sección, *Abreviaciones*, en la que los autores definen los acrónimos, las siglas y las abreviaturas que se utilizan en el resto del documento. Como parte del pre-procesamiento, se extraen de esta sección tanto las formas abreviadas como sus expansiones. Esta información se utiliza para reemplazar en el resto del documento las formas abreviadas por sus correspondientes formas expandidas.

En cuanto a la eliminación de palabras genéricas se refiere, se utiliza la lista de palabras vacías de *MEDLINE*[2].

5.2.2. Etapa II: Traducción de las Oraciones a Conceptos del Dominio

De entre todas las terminologías biomédicas analizadas en la Sección 3.1, se ha seleccionado UMLS para este caso de estudio, por los siguientes motivos:

- En primer lugar, el propio propósito para el que han sido desarrolladas las tres terminologías 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 la catalogación de documentos, asignando titulares y categorías a

[2]PubMed StopWords.
http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords. Consultada el 1 de noviembre de 2010

las publicaciones. Tanto los contenidos como la estructuración obedecen a dicho propósito y se muestran insuficientes 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. Por su parte, SNOMED-CT está más orientado a ofrecer una terminología común para ser utilizada por expertos sanitarios (médicos, investigadores) y enfermos, y no tanto por aplicaciones (aunque cada vez más se contempla esta posibilidad).

- UMLS cuenta con el respaldo de un considerable número de aplicaciones que lo utilizan (*PubMed*, *NLM Gateway*, *ClinicalTrials.gov*, la *Indexing Initiative* de NLM o los *Enterprise Vocabulary Services* del *National Cancer Institute*). Por su parte, SNOMED-CT 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. 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 a UMLS, SNOMED-CT y MeSH adolecen de no disponer de herramientas que faciliten el análisis léxico de los textos.

Por otro lado, UMLS proporciona varios mecanismos de acceso a los datos del Metatesauro, el Léxico Especializado y la Red Semántica:

- A través de aplicaciones java, utilizando un API que permite la conexión de los programas de usuario al servidor *UMLSKSS* (*UMLS Knowledge Source Server*).

- Cargando los archivos relacionales de UMLS en una base de datos local y accediendo a ellos mediante consultas *SQL*.

Ambas alternativas han sido implementadas, si bien se ha constatado que el acceso local constituye la opción más adecuada, por tres razones:

1. Debido al elevado tiempo que consume la primera alternativa. Sirva para clarificar esta afirmación los experimentos realizados sobre distintos documentos 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 al *UMLSKSS* se necesitan del orden de varias horas.

2. Se elimina así la excesiva dependencia del servidor *UMLSKSS*, que en ocasiones no se encuentra disponible por tareas de mantenimiento y mejora.

3. Mantener una copia local de la base de datos permite utilizar el programa MetamorphoSys para restringir los vocabularios utilizados, reduciendo así tanto el espacio en disco como el tiempo de cómputo necesarios.

No obstante, el acceso local tiene el inconveniente de que la base de datos debe ser actualizada periódicamente, cada vez que la NLM publique una nueva versión, lo que sucede aproximadamente dos veces al año.

Para establecer la correspondencia entre el texto del documento y los conceptos de UMLS, se ha utilizado el programa MetaMap (ver Sección 3.3). 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 de la oración en el Metatesauro. Tal y como se aprecia en la Tabla 5.1, 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 a construir, y en una consiguiente mejora de la eficiencia del algoritmo. 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*

en el primer caso es más precisa que la recuperación de los dos conceptos distintos *Angiotensin* y *Enzyme* en el segundo caso.

Conceptos recuperados por MetaMap
1. C0012798:Diuretics
2. C0008294:Chlorthalidone
3. C0205166:Long
4. C0439228:Day
5. C0013227:Pharmaceutical Preparations
6. C0443199:Differential quality
7. C0456387:Class
8. C0003015:Angiotensin-Converting Enzyme Inhibitors
9. C0065374:Lisinopril
10. C0006684:Calcium Channel Blockers
11. C0051696:Amlodipine
12. C0001641:Adrenergic alpha-Antagonists
13. C0114873:Doxazosin
Conceptos recuperados indexando con unigramas
1. C0012798:Diuretics
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
15. C0114873:Doxazosin

Tabla 5.1: Conceptos recuperados por MetaMap e indexando con unigramas, respectivamente

Por otro lado, los documentos biomédicos a menudo sufren de ambigüedad léxica en cuanto al significado de sus términos se refiere. Esta ambigüedad se refleja, a su vez, en las propias terminologías. Así, por ejemplo, la palabra *cold* se corresponde en el Metatesauro de UMLS con 7 posibles significados. Por este motivo, MetaMap a menudo devuelve más de un concepto candidato para un único término o conjunto de términos, a los que además asigna la misma puntuación, tal y como se observa en la Figura 5.1, en la que el término *cold* en la oración *Tissues are often cold* es traducido a tres

posibles conceptos ("C0234192:Cold Sensation", "C0009443:Common Cold" y "C0009264:Cold Temperature"), todos ellos con la misma puntuación.

Phrase: "Tissues"
Meta Mapping (1000):
1000 C0040300:Tissues (Body tissue)

Phrase: "are"

Phrase: "often cold"
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0234192:Cold (Cold Sensation)
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0009443:Cold (Common Cold)
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0009264:Cold (Cold Temperature)

Figura 5.1: Salida de MetaMap para la oración *Tissues are often cold*

Es necesario, por lo tanto, disponer de algún mecanismo que permita elegir, de entre los distintos significados devueltos por MetaMap, el más adecuado al contexto en el que se encuentra. Para ello, el generador de resúmenes permite el uso de diferentes algoritmos de desambiguación léxica:

- La primera posibilidad y la más simple consiste en utilizar la propia opción *-y* de MetaMap, que implementa un algoritmo de desambiguación que favorece a aquellos conceptos semánticamente consistentes con el resto de conceptos en su contexto, a través de la información contenida en sus tipos semánticos (Humphrey et al., 2006).

- La segunda opción es utilizar el algoritmo de desambiguación basado en grafos *PPR (Personalized PageRank)* (Agirre y Soroa, 2009), descrito en la Sección 3.4, en cualquiera de sus dos modos de funcionamiento: estándar o palabra-por-palabra. Para ello, ha sido necesario adaptar el algoritmo para su uso en el dominio biomédico. Recordemos que este algoritmo necesita dos recursos de conocimiento: un *diccionario* y una *base de conocimiento léxico*. Los autores distribuyen una versión que dispone de los recursos necesarios para operar con WordNet, pero que resulta inapropiada para trabajar con documentos de biomedicina y desambiguar los conceptos del Metatesauro de UMLS recuperados por MetaMap. Por ello, se ha utilizado el Metatesauro para construir

un diccionario y una base de conocimiento léxico. En primer lugar, se ha generado un grafo en el que cada nodo es un CUI (identificador de concepto) del Metatesauro, y las aristas representan todas las posibles relaciones establecidas entre los respectivos conceptos (obtenidas de la tabla *MRREL* del Metatesauro). En cuanto al diccionario se refiere, simplemente se utiliza la salida de MetaMap al aplicarlo sobre cada palabra del documento (es decir, la lista de CUI candidatos que MetaMap devuelve para cada término).

Por otro lado, dado que 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, se ha decidido ignorar estos conceptos a la hora de construir los grafos de las oraciones y del documento. Los tipos semánticos de UMLS son de gran utilidad a la hora de identificar los términos asociados a conceptos muy generales. De esta forma, tras un estudio del significado de cada tipo semántico y de los conceptos que agrupa, se ha determinado que los conceptos pertenecientes a tipos tan genéricos como *Quantitative Concept*, *Qualitative Concept*, *Temporal Concept*, *Functional Concept*, *Idea or Concept*, *Intellectual Product*, *Mental Process*, *Spatial Concept* y *Language* serán omitidos. 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 en la construcción del grafo del documento.

- **Quantitative Concept:** Lowered, Two, Four, Several.
- **Qualitative Concept:** Firstly, Initial, Possibly, Definite.
- **Temporal Concept:** Previous, Year, Seconds, Frequent.
- **Functional Concept:** Purpose, Designate, Treat, Lead.
- **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.

Para ilustrar todo el proceso descrito en esta etapa, 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 hypertension. Como resultado de la ejecución de MetaMap, se obtienen los conceptos mostrados en la Tabla 5.2. Los conceptos de tipos semánticos muy generales, y que por lo tanto son ignorados por el algoritmo, aparecen tachados.

Concept	MetaMap Score	Semantic Type
~~**Goals**~~	~~1000~~	~~Intellectual Product~~
Clinical Trials	1000	Research Activity
Cardiovascular system	694	Body System
~~**Mortality vital statistics**~~	~~861~~	~~Quantitative Concept~~
~~**Morbidity - disease rate**~~	~~1000~~	~~Quantitative Concept~~
Cerebrovascular accident	1000	Disease or Syndrome
Coronary heart disease	1000	Disease or Syndrome
Congestive heart failure	1000	Disease or Syndrome
~~**Evidence of**~~	~~660~~	~~Functional Concept~~
~~**Basis**~~	~~660~~	~~Functional Concept~~
Clinicians	1000	Prof. or Occup. Group
~~**Treatment intent**~~	~~1000~~	~~Functional Concept~~
Hypertensive disease	1000	Disease or Syndrome

Tabla 5.2: Conceptos asociados a 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 hypertension*

El número final de conceptos asociados al documento de ejemplo, una vez eliminados los conceptos que se han considerado excesivamente genéricos, es de 345. Es importante resaltar que se trata de un documento relativamente pequeño. Los experimentos realizados sobre artículos más extensos muestran en torno a 1500-2000 conceptos.

5.2.3. Etapa III: Representación de las Oraciones como Grafos de Conceptos

Los conceptos descubiertos por MetaMap para cada oración se recuperan del Metatesauro de UMLS, junto con su jerarquía completa de hiperónimos (relaciones *es un*). Estas relaciones se obtienen de la tabla *MRHIER* del Metatesauro, que lista todas las jerarquías en las que aparece el concepto, presentando el camino completo desde dicho concepto hasta la raíz de la jerarquía. Así, por ejemplo, dicha tabla establece que el concepto "C0035243:Respiratory Tract Infections" es un padre del concepto "C0009443:Common Cold". A modo de ejemplo, la Figura 5.2 muestra el resultado del proceso descrito

para el concepto la palabra “C0007226:Cardiovascular”.

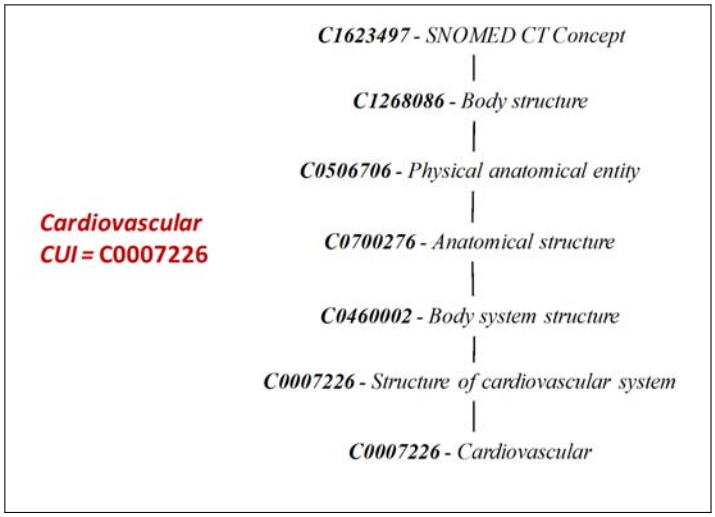


Figura 5.2: Hiperónimos del concepto *cardiovascular*

Las jerarquías de todos los conceptos de una misma oración se mezclan de manera que, como resultado, se obtiene una estructura de árbol que representa a la oración. Los dos niveles superiores de esta jerarquía se eliminan, una vez más por representar conceptos con un significado muy general. La Figura 5.3 muestra el grafo correspondiente a la oración de ejemplo *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.*

5.2.4. Etapa IV: Construcción del Grafo del Documento

Tras unir los grafos de todas las oraciones en un único grafo que representa al conjunto del documento, este grafo se extiende con dos nuevas relaciones semánticas entre nodos: la relación *associated with*, definida entre tipos de la Red Semántica de UMLS; y la relación *related to* o *other related*, definida entre conceptos del Metatesauro. Según la especificación de UMLS, la relación *associated with* une dos tipos semánticos entre los que existe una relación significativa. Esta relación se encuentra definida en la tabla *SRSTR* de la Red Semántica. Por su parte, la relación *related to* define otros tipos de relaciones existentes entre conceptos dentro de su propio vocabulario (i.e. dentro de SNOMED-CT o MeSH, pero no entre ellos), excluyendo la relación de co-ocurrencia. Esta relación se encuentra definida en la tabla

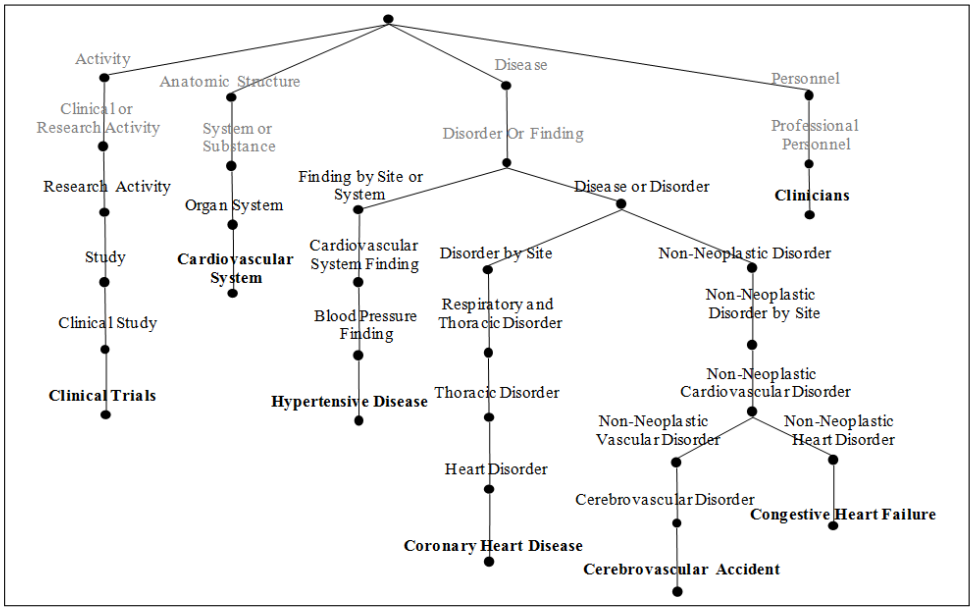


Figura 5.3: Grafo semántico de la oración *The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension*

MRREL del Metatesauro (ver Sección 3.1.3). Posteriormente, las aristas del grafo son etiquetadas utilizando alguno de los coeficientes de similitud explicados en la Sección 4.4. La Figura 5.4 muestra, a modo de ejemplo, el grafo de un documento ficticio compuesto de las siguientes dos oraciones, donde las aristas han sido etiquetadas utilizando el coeficiente de Jaccard:

1. *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.*

2. *While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke to occur more often in the doxazosin group, than in the group taking chlorthalidone.*

5.2.5. Etapa V: Clustering de Conceptos e Identificación de Temas

En esta etapa, el algoritmo no requiere ninguna matización o adaptación con respecto a lo explicado en la Sección 4.5. No obstante, y continuando

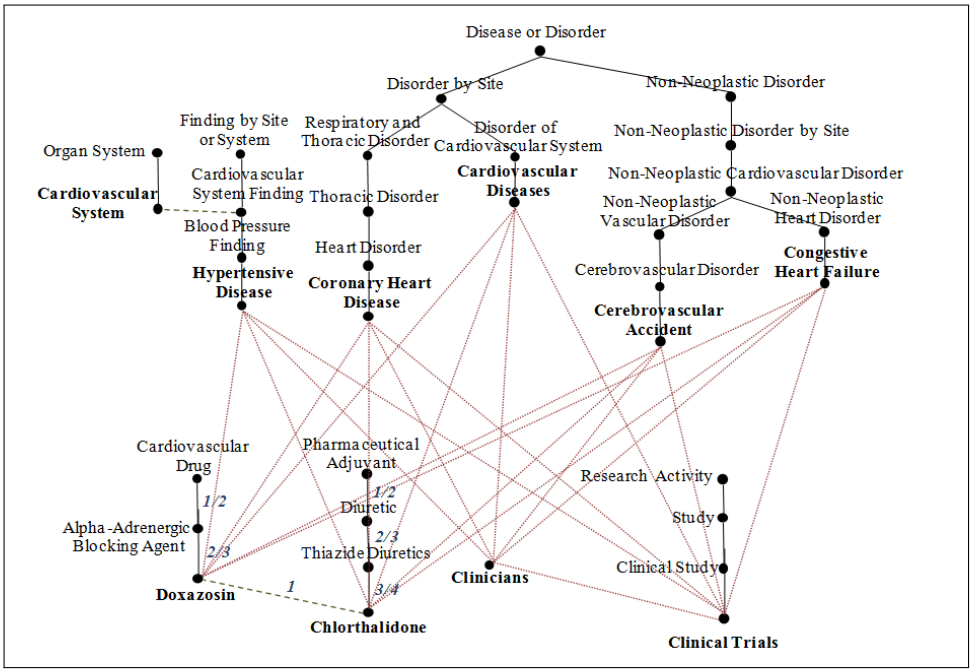


Figura 5.4: Ejemplo de grafo semántico de un documento ficticio

con el ejemplo que nos ocupa, señalaremos algunas observaciones con respecto a la influencia que la elección del porcentaje de *hub vertices* y de las relaciones semánticas tienen sobre las características estructurales del grafo del documento, y por tanto, sobre el resultado del algoritmo de agrupamiento. La Tabla 5.3 muestra el número de *HVS* generados para diferentes configuraciones de relaciones semánticas y porcentajes de *hub vertices*.

Relaciones	**Número de *Hub Vertices***	**Número de HVS**
Hiperonimia + *Associated with*	5 %	3
	10 %	5
	20 %	4
Hiperonimia + *Related to*	5 %	3
	10 %	8
	20 %	13
Hiperonimia + *Associated with* + *Related to*	5 %	2
	10 %	5
	20 %	3

Tabla 5.3: Efecto sobre el número de *HVS* del número de *hub vertices* y del número y tipo de las relaciones semánticas consideradas

En primer lugar, se observa en la Tabla 5.3 que la relación *related to*, cuando se utiliza en exclusividad, da como resultado un número de *HVS*

demasiado elevado. Esto se debe a que esta relación conecta entre sí a pocos conceptos del Metatesauro, por lo que el grafo del documento resulta demasiado inconexo. En cuanto a utilizar la relación *associated with* únicamente o junto a la relación *related to*, 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. No parece existir una relación directa entre el número de *hub vertices* y el número de *HVS* generados por el algoritmo de agrupamiento. La razón parece ser que, si bien al aumentar el número de *hub vertices* aumentan las posibilidades de crear un número mayor de grupos, también lo hacen las posibilidades de establecer relaciones semánticas entre ellos, de modo que los distintos *HVS* se agrupan entre sí reduciendo su número pero aumentando su tamaño.

A continuación, se muestran los *HVS* generados para el documento de ejemplo, cuando el número de *hub vertices* utilizado es del 20 % y el grafo del documento ha sido construido a partir de relaciones *is a* y *associated with* (Tabla 5.4).

HVS 1 (3 conceptos)		
Study	Provide	View
HVS 2 (28 conceptos)		
Analysis of substances	Hepatic	Audiological observations
Blood	Entire upper arm	Age
receptor	Base	Very large
Other therapy NOS	Reserpine	Related personal status
Reduction - action	Agent	In care
Therapeutic procedure	Discontinued	Cardiovascular event
Finding	Admin. occup. activities	Cardiovascular system
Primary operation	Descriptor	heart rate
Entire lung	Guide device	Adverse reactions
Entire heart		
HVS 3 (4 conceptos)		
Systolic hypertension	May	Person
Blood Pressure		
HVS 4 (32 conceptos)		
Duplicate concept	Lower	Clonidine
Articular system	Support, device	Hydralazine
Entire hand	Antihypertensive Agents	Adrenergic beta-Antag.
Body system structure	Falls	Diuretics
Diastolic blood pressure	Angiotensin-Conv. Enz. Inh.	PREVENT
Expression procedure	Prazosin	CONCEPT Drug
Prevention	Immune Tolerance	Calcium Channel Blockers
Chlorthalidone	Tissue damage	Assessment procedure
Hopelessness	Ramipril	Qualifier value
Lisinopril	Amlodipine	Unapproved attribute
Doxazosin	Reporting	

Tabla 5.4: Conceptos que conforman los distintos HVS

5.2.6. Etapa VI: Asignación de Oraciones a Grafos

De acuerdo a lo explicado en la Sección 4.6, en esta etapa cada una de las oraciones asigna una puntuación a cada uno de los clusters, en función de su similitud semántica. En concreto, para el ejemplo que nos ocupa, la Tabla 5.5 muestra las puntuaciones asignadas por cada oración a cada uno de los clusters generados en la etapa anterior.

Oración	**C.1**	**C.2**	**C.3**	**C.4**	**Oración**	**C.1**	**C.2**	**C.3**	**C.4**
1	98.0	172.0	74.0	208.0	30	6.0	7.0	4.0	7.0
2	13.0	21.0	9.0	18.0	31	9.0	14.0	5.0	12.0
3	8.0	13.0	4.0	17.0	32	4.0	6.0	2.0	5.0
4	8.0	19.0	4.0	16.0	33	3.0	3.0	2.0	2.0
5	9.0	13.0	4.0	16.0	34	18.0	18.0	9.0	20.0
6	6.0	4.0	4.0	8.0	35	8.0	15.0	5.0	17.0
7	1.0	1.0	1.0	3.0	36	1.0	6.0	1.0	2.0
8	8.0	15.0	4.0	25.0	37	8.0	12.0	4.0	12.0
9	3.0	9.0	2.0	10.0	38	6.0	6.0	4.0	9.0
10	9.0	10.0	6.0	10.0	39	0.0	2.0	0.0	6.0
11	4.0	9.0	2.0	14.0	40	2.0	2.0	1.0	2.0
12	5.0	7.0	3.0	8.0	41	1.0	4.0	1.0	8.0
13	16.0	29.0	11.0	25.0	42	9.0	14.0	6.0	17.0
14	7.0	5.0	5.0	15.0	43	14.0	14.0	8.0	26.0
15	3.0	10.0	1.0	9.0	44	11.0	12.0	7.0	22.0
16	11.0	11.0	7.0	15.0	45	3.0	6.0	2.0	9.0
17	4.0	5.0	2.0	6.0	46	5.0	16.0	2.0	18.0
18	5.0	11.0	3.0	17.0	47	3.0	5.0	2.0	11.0
19	5.0	6.0	2.0	4.0	48	4.0	11.0	3.0	20.0
20	17.0	24.0	10.0	25.0	49	7.0	9.0	4.0	12.0
21	16.0	31.0	12.0	27.0	50	5.0	10.0	4.0	13.0
22	11.0	23.0	8.0	27.0	51	3.0	9.0	3.0	6.0
23	3.0	8.0	4.0	14.0	52	2.0	6.0	1.0	6.0
24	14.0	17.0	9.0	27.0	53	2.0	2.0	2.0	4.0
25	7.0	27.0	6.0	20.0	54	0.0	0.0	0.0	0.0
26	5.0	7.0	3.0	12.0	55	0.0	0.0	0.0	0.0
27	5.0	6.0	2.0	8.0	56	2.0	5.0	2.0	4.0
28	2.0	5.0	1.0	2.0	57	12.0	14.0	7.0	22.0
29	10.0	22.0	8.0	18.0	58	4.0	7.0	2.0	7.0

Tabla 5.5: Similitud entre oraciones y clusters

5.2.7. Etapa VII: Selección de Oraciones para el Resumen

Finalmente, la Tabla 5.6 recoge las oraciones seleccionadas por cada heurística (ver Sección 4.7), junto con su puntuación, para el documento de ejemplo. Se ha utilizado un ratio de compresión del 15 %, y no se utilizan los criterios de posición y de similitud con el título. Es importante señalar que, puesto que en el tipo de documentos que nos ocupa es posible encontrar tablas y figuras con información relevante para el resumen, las tablas y figuras a las

que se hace referencia desde alguna de las oraciones incluidas en el resumen, se incluyen también en dicho resumen. Sin embargo, la información en tales tablas y figuras no será tenida en cuenta a la hora de evaluar los respectivos resúmenes con las métricas ROUGE.

Heurística 1		**Heurística 2**		**Heurística 3**	
Oración	**Puntuación**	**Oración**	**Puntuación**	**Oración**	**Puntuación**
1	208.0	1	208.0	1	101.33
22	27.0	21	31.0	21	16.47
24	27.0	13	29.0	13	15.81
43	26.0	22	27.0	31	15.53
8	25.0	24	27.0	8	15.06
20	25.0	25	27.0	10	13.2
44	22.0	29	26.0	25	12.67
3	22.0	8	25.0	20	12.5
34	20.0	20	25.0	2	11.53

Tabla 5.6: Oraciones seleccionadas por cada heurística y puntuación asignada a cada una de ellas

Para terminar, a continuación, se muestra el resumen generado por cada una de las heurísticas (Tablas 5.7, 5.8 y 5.9, respectivamente).

	Heurística 1
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.

Tabla 5.7: Resumen generado por la heurística 1

	Heurística 2
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.

Tabla 5.8: Resumen generado por la heurística 2

	Heurística 3
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.
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.
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.

Tabla 5.9: Resumen generado por la heurística 3

Capítulo 6

Caso de Estudio: Resúmenes Mono-documento de Noticias Periodísticas

El segundo caso de estudio tiene como objetivo configurar el método genérico para producir resúmenes de noticias periodísticas. Se trata pues de un dominio y de una tipología de documentos sustancialmente distintos de los artículos científicos en biomedicina. El presente capítulo se desarrolla según el mismo esquema que el capítulo anterior. En primer lugar, se presenta un breve estudio de las características del lenguaje periodístico y de la estructura del subgénero que nos ocupa: la noticia periodística. En segundo lugar, se describe el proceso de especialización del método genérico a través de un ejemplo. En concreto, se utiliza un documento del corpus de noticias de las conferencias DUC 2002.

6.1. Peculiaridades del Dominio y del Tipo de Documentos

El texto periodístico presenta una serie de características lingüísticas y estructurales definitorias. En primer lugar, abarca una amplia variedad de subgéneros, desde la noticia hasta el reportaje, pasando por la entrevista o el artículo de opinión. Estos subgéneros, a su vez, se diferencian unos de otros en la estructura y el lenguaje que utilizan. En este trabajo, el interés se centra en la noticia. La estructura de una noticia presenta generalmente los

siguientes elementos: un *Titular*, con el que se informa al lector, en pocas palabras, sobre la temática de la noticia; una *Entrada* o *Lead*, que puede verse como un resumen muy condensado de la noticia; y el *Cuerpo*, donde se aclara, matiza y completa la información. El cuerpo de la noticia suele venir estructurado en forma de **pirámide invertida**: los datos de mayor interés se incluyen en primer lugar y, a continuación, se desarrollan aspectos secundarios.

En cuanto al lenguaje y el estilo utilizado, uno de los rasgos más distintivos de la noticia periodística es la **concisión**. La noticia suele ser concisa, lo que significa que en general no contendrá información redundante, sino que la mayoría de la información será de interés para el lector. Esto complica la labor de generación del resumen, tanto si se aborda de manera automática como manual. Además, la amplitud temática es inmensa, como también lo es el vocabulario utilizado. Otro problema al que se enfrentan las sistemas de información al tratar con textos periodísticos es la presencia de un lenguaje con cierta *tendencia al cliché*, plagado de **frases hechas**, **metáforas** y **tópicos**. Por el contrario, suelen emplearse oraciones cortas, y de sintaxis sencilla. Por último, idealmente tratan la información de manera objetiva, y la presentan de manera impersonal.

6.2. Especialización del Método para el Dominio Periodístico

El objetivo de esta sección es detallar, para cada una de las etapas en que se subdivide el algoritmo de generación de resúmenes, el proceso y los recursos necesarios para adaptar el método genérico al dominio periodístico. Como ya se ha mencionado, a lo largo de la exposición se hará referencia a un ejemplo de generación de resumen para un documento concreto del corpus de la conferencia DUC 2002. La noticia completa presenta un total de 16 oraciones, y puede encontrarse en el Apéndice B.2 de este documento.

6.2.1. Etapa I: Pre-procesamiento

En primer lugar, se han considerado innecesarias para la realización del resumen las secciones *Número de documento*, *Nombre del fichero*, *Entrada* (si existe), *Fecha de publicación*, *Autores* y *Nombre de la publicación*; y por lo tanto, han sido eliminadas del documento.

Además, se ha observado que en este tipo de documentos es muy frecuente la presencia de acrónimos y abreviaturas (por ejemplo, NY (*New York*), Al (*Alabama*), Co (*Company*)). Puesto que no disponen de una sección en la que se especifiquen las formas expandidas correspondientes, se han utilizado las listas de abreviaturas y acrónimos disponibles en el módulo de nomenclatura *ANNIE Gazetteer* de GATE (Sección 3.2). Así, cada vez que se encuentra en el texto una forma abreviada cuya expansión aparece en dichas listas, simplemente se sustituye la una por la otra.

En cuanto a la eliminación de palabras genéricas se refiere, se utiliza la lista de palabras vacías elaborada por Ted Pedersen[1], especialmente diseñada para su uso en textos procesados o que se quieren procesar con WordNet.

6.2.2. Etapa II: Traducción de las Oraciones a Conceptos del Dominio

En este caso de estudio se utiliza la base de datos léxica WordNet (Sección 3.1.6) para la traducción de los documentos a conceptos. No se trata, pues, de una recurso específico de un dominio concreto, sino de un recurso de propósito general, ya que los documentos que nos conciernen versan sobre temas muy dispares. De entre todos los recursos de propósito general analizados en la Sección 3.1, se ha seleccionado WordNet por los siguientes motivos:

- El primer motivo se desprende del propósito para el que han sido desarrollados los tres recursos. El objetivo de Cyc es proporcionar una base de conocimiento de sentido común que pueda ser utilizada en cualquier sistema de Inteligencia Artificial. El objetivo de WordNet es menos pretencioso: crear una base de conocimiento léxico adecuada para el inglés. Puesto que nuestro sistema de generación de resúmenes no requiere el uso de complejos mecanismos de inferencia, el alcance de WordNet resulta suficiente para nuestro propósito. Por el contrario, el hecho de que el tesauro Roget no diferencie explícitamente entre los tipos de relaciones que conectan sus conceptos, limita las posibilidades de utilizarlo en el generador de resúmenes.

- Mientras que Cyc está optimizado para realizar razonamiento lógico, WordNet está optimizado para realizar tareas de procesamiento de len-

[1] WordNet Stop List.
http://www.d.umn.edu/~tpederse/Group01/WordNet/wordnet-stoplist.html. Consultada el 1 de noviembre de 2010

guaje natural, como categorización léxica y determinación de similitud semántica entre términos.

- WordNet es, a día de hoy, el recurso léxico más utilizado por la comunidad de lingüística computacional. Esto ha propiciado el desarrollo de un amplio abanico de herramientas que facilitan tanto su uso como la realización de otras tareas de procesamiento de lenguaje. Así, dispone de interfaces de acceso en más de 20 lenguajes de programación, sistemas de desambiguación capaces de distinguir el contexto en el que se utilizan los diferentes términos, algoritmos destinados a calcular la similitud entre ellos, etc.

- WordNet es relativamente sencillo de utilizar, mientras que la dificultad de utilizar Cyc para resolver problemas reales de razonamiento textual constituye el principal factor que limita su uso en muchas tareas de procesamiento de lenguaje natural.

Para traducir el texto del documento a conceptos de WordNet, se ha utilizado el programa WordNet::SenseRelate (Sección 3.6). Para ello, se ha integrado en el código del generador de resúmenes la llamada al programa *wsd.pl*, con la siguiente configuración de parámetros:

- **Formato del texto:** plano.
- **Esquema de desambiguación:** normal.
- **Medida de similitud:** cualquiera de las definidas en la herramienta, a especificar por el usuario en el archivo de configuración del generador de resúmenes.
- **Lista de parada:** Lista de palabras vacías *WordNet Stop List*.
- **Tamaño de la ventana de contexto:** Cuatro palabras.

Para justificar el uso de WordNet::SenseRelate, sirva el siguiente apunte: a lo largo de la noticia de ejemplo, se utiliza reiteradamente la palabra *storm*, término que, cuando utilizado como sustantivo, presenta tres significados muy distintos en WordNet: una primera acepción que hace referencia a una perturbación atmosférica violenta acompañada de aparato eléctrico y viento fuerte, lluvia, nieve o granizo; una segunda acepción que se refiere a una manifestación violenta de un estado de ánimo o conmoción; y una tercera acepción como sinónimo de asalto o ataque a una fortaleza militar o bastión. Parece obvio que, si en lugar de seleccionar la primera acepción del

término *storm*, seleccionamos cualquiera de las siguientes, nuestro sistema trabajará con conceptos erróneos e intentará seleccionar las oraciones para el resumen en base a significados que ni siquiera están presentes en el texto. WordNet::SenseRelate permite solucionar este problema utilizando alguno de los algoritmos de desambiguación léxica implementados en la herramienta para seleccionar el significado apropiado de cada palabra utilizando la información presente en su contexto.

El resultado de aplicar WordNet:SenseRelate puede observarse en la Tabla 6.1, donde se muestran los conceptos de WordNet recuperados para la oración *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.* WordNet::SenseRelate reconoce 18 conceptos, definidos cada uno de ellos por su identificador de concepto y su *sense* o acepción en WordNet. El número total de conceptos asociados al documento de ejemplo es de 209.

Término	***Sense***	**Término**	***Sense***
Hurricane	1	**populate**	2
Gilbert	2	**south**	1
sweep	1	**coast**	1
Dom. Rep	1	**prepare**	4
Sunday	1	**high**	2
civil	1	**wind**	1
defense	9	**heavy**	1
alert	1	**rain**	1
heavily	2	**sea**	1

Tabla 6.1: Conceptos asociados a la oración del ejemplo

6.2.3. Etapa III: Representación de las Oraciones como Grafos de Conceptos

De los conceptos descubiertos en la etapa anterior, sólo aquellos que se corresponden con sustantivos se recuperan de WordNet junto con su jerarquía completa de hiperónimos. Los resultados experimentales han mostrado que el algoritmo se comporta mejor si sólo se tienen en cuenta los conceptos de esta categoría gramatical. A modo de ejemplo, la Figura 6.1 muestra la jerarquía de hiperónimos del concepto *hurricane*.

Para acceder a WordNet y recuperar tales conceptos y sus relaciones, se

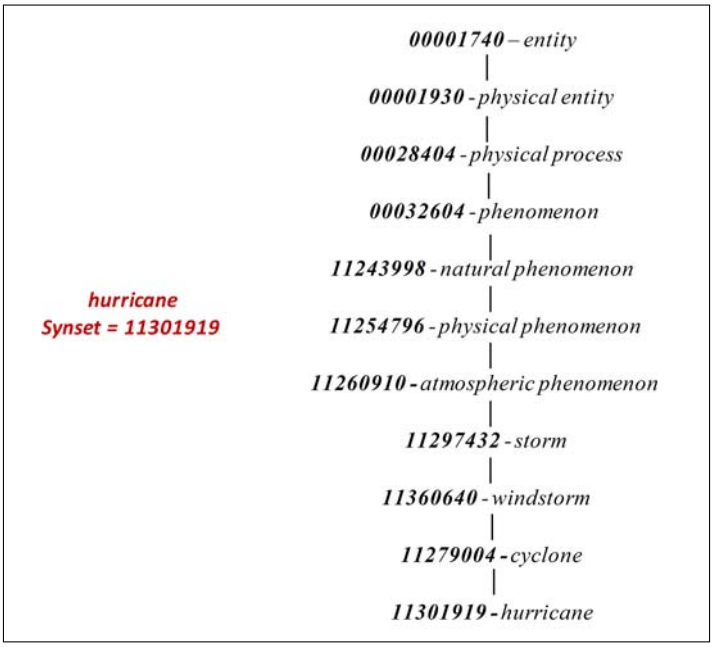


Figura 6.1: Hiperónimos del concepto *hurricane*

puede utilizar cualquiera de las APIs[2] desarrolladas a tal efecto y para distintos lenguajes de programación (.NET/C#, COM, dBase, Java, MySQL, Perl, etc.). En concreto, y puesto que nuestro sistema se ha desarrollado en JAVA, se han considerado y evaluado las siguientes APIs:

- JAWS (Java API for WordNet Searching), desarrollada por Brett Spell.
- JWordNet, de la Universidad George Washington.
- JWI (MIT Java Interface to WordNet), del Massachusetts Institute of Technology.

De todas ellas, se ha elegido JWI por su compatibilidad con la versión de WordNet utilizada y su facilidad de uso.

De nuevo, las jerarquías de todos los conceptos de una misma oración se mezclan, de manera que como resultado se obtiene una estructura de árbol que representa a la oración. La Figura 6.2 muestra el árbol correspondiente a la oración *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.* Una vez más, los niveles superiores

[2]WordNet APIs. http://wordnet.princeton.edu/wordnet/related-projects/#Java. Consultada el 1 de noviembre de 2010

de esta jerarquía, esta vez tres, se eliminan por representar conceptos muy genéricos.

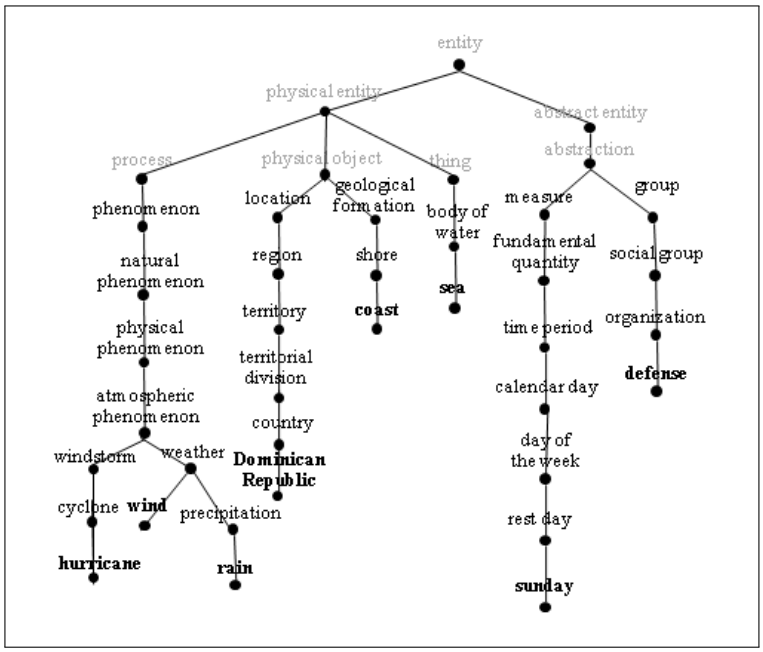


Figura 6.2: Grafo semántico de la oración *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas*

6.2.4. Etapa IV: Construcción del Grafo del Documento

En esta etapa se unen todos los grafos de las oraciones en un único grafo que representa al documento. A su vez, este nuevo grafo se extiende con nuevas relaciones semánticas entre conceptos. En concreto, se han realizado distintos experimentos utilizando la que denominaremos *relación de similitud semántica*. Para ello, se ha utilizado el paquete WordNet::Similarity (ver Sección 3.5) para calcular la similitud entre pares de conceptos representados en el grafo. Puesto que paquete define distintas medidas de similitud, el sistema permite al usuario especificar cuál de ellas desea utilizar. Para expandir el grafo del documento con estas relaciones, se añade una nueva arista entre cada par de nodos hoja si la similitud entre ellos supera un determinado *umbral de similitud*. Posteriormente, las aristas del grafo son etiquetadas según lo explicado en la Sección 4.4. La Figura 6.3 muestra, a modo de ejemplo el grafo de un documento ficticio compuesto únicamente

por la oración *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.* Para construir este grafo, se ha utilizado como medida de similitud el algoritmo *Lesk* y un umbral igual a 0.01. Las aristas han sido etiquetadas utilizando el coeficiente de similitud de Jaccard.

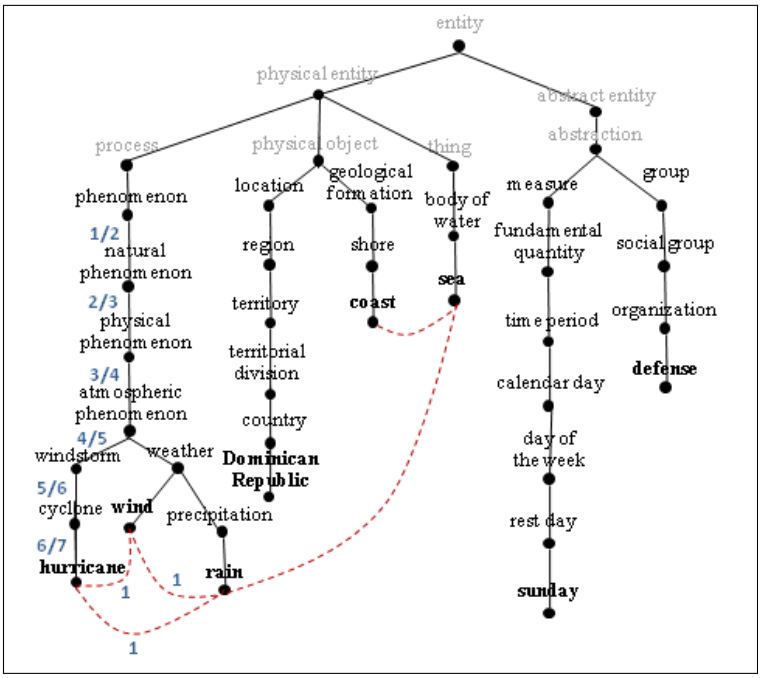


Figura 6.3: Ejemplo de grafo semántico de un documento ficticio

6.2.5. Etapa V: Clustering de Conceptos e Identificación de Temas

Siguiendo con el ejemplo que nos ocupa, la Tabla 6.2 muestra cómo la elección de los parámetros del algoritmo influye en las características estructurales del grafo del documento, y por tanto, en el resultado del algoritmo de agrupamiento y en la delimitación de los temas. Se puede observar que al aumentar el número de *hub vertices*, el número de clusters aumenta, mientras que al aumentar la conectividad del grafo (i.e. al reducir el umbral de similitud) el número de clusters disminuye. Los experimentos realizados incluyen ambos tipos de relaciones (hiperonimia y similitud semántica), pues de considerar únicamente la relación de hiperonimia, el grafo generado re-

sulta tan inconexo que el algoritmo de agrupamiento no consigue converger sin necesidad de limitar el número de iteraciones.

Umbral de Similitud	**Número de *Hub Vertices***	**Número de HVS**
0.001	5 %	2
	10 %	7
	20 %	14
0.01	5 %	2
	10 %	6
	20 %	13
0.5	5 %	2
	10 %	5
	20 %	10

Tabla 6.2: Efecto sobre el número de *HVS* del número de *hub vertices* y del umbral de similitud considerados

La Tabla 6.3 muestra los *HVS* generados para el documento de ejemplo. En este caso, el número de *hub vertices* se ha fijado al 10 % del total de conceptos del documento, y el umbral de similitud se ha fijado a *0.01*. Asimismo, se han utilizado ambas relaciones semánticas (hiperonimia y similitud).

HVS 1 (7 conceptos)		
atmospheric condition	wind	flood
precipitation	weather	cyclone
hurricane		
HVS 2 (3 conceptos)		
people	resident	inhabitant
HVS 3 (3 conceptos)		
casualty	fatality	accident
HVS 4 (5 conceptos)		
island	gulf	region
republic	Caribbean	
HVS 5 (3 conceptos)		
Saturday	Sunday	weekday

Tabla 6.3: Conceptos que conforman los distintos *HVS* para el documento de ejemplo

6.2.6. Etapa VI: Asignación de Oraciones a Grafos

De acuerdo a lo explicado en la Sección 4.6, en esta etapa cada una de las oraciones asigna una puntuación a cada cluster en función la similitud

semántica entre el grafo de la oración y cada uno de estos clusters. En concreto, para el ejemplo que nos ocupa, la Tabla 6.4 muestra las puntuaciones asignadas para cada oración y cluster.

Oración	**C.1**	**C.2**	**C.3**	**C.4**	**C.5**
1	54.0	37.0	34.5	49.5	19.5
2	27.0	3.5	4.5	20.5	0.0
3	18.0	15.5	48.0	5.0	32.5
4	17.5	14.5	12.0	38.5	2.0
5	5.0	38.0	21.0	40.0	2.0
6	48.0	10.0	22.5	45.5	19.5
7	29.5	9.0	14.5	48.0	18.0
8	50.0	8.0	15.0	51.5	3.5
9	21.5	6.5	18.0	29.5	17.0
10	46.5	9.0	20.5	41.5	1.5
11	1.0	0.5	5.5	0.0	0.0
12	41.5	25.0	20.5	38.5	28.0
13	40.0	21.0	18.0	35.5	16.5
14	35.5	28.0	20.5	21.5	2.5
15	48.5	16.0	22.5	18.0	0.0
16	30.5	12.5	18.0	39.0	9.0

Tabla 6.4: Similitud entre oraciones y clusters

6.2.7. Etapa VII: Selección de Oraciones para el Resumen

Finalmente, la Tabla 6.5 recoge las oraciones seleccionadas por cada heurística (ver Sección 4.7), junto con su puntuación, para el documento de ejemplo. Se ha utilizado un ratio de compresión del 30 %. No se utilizan los criterios de posición y similitud con el título.

Heurística 1		**Heurística 2**		**Heurística 3**	
Oración	**Puntuación**	**Oración**	**Puntuación**	**Oración**	**Puntuación**
1	54.0	1	54.0	1	9.33
8	50.0	8	50.0	3	8.11
15	48.5	15	48.5	12	7.99
6	48.0	7	48.0	6	6.31
10	46.5	3	48.0	13	6.13

Tabla 6.5: Oraciones seleccionadas por cada heurística y puntuación asignada a cada una de ellas

Para terminar, se muestra el resumen generado por cada una de las heurísticas (Tablas 6.6, 6.7 y 6.8, respectivamente).

	Heurística 1
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
6	Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
8	The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a broad area of cloudiness and heavy weather rotating around the center of the storm.
10	Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico s south coast.
15	Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.

Tabla 6.6: Resumen generado por la heurística 1

	Heurística 2
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
3	There is no need for alarm, Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.
7	The National Hurricane Center in Miami reported its position at 2 a.m Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico and 200 miles southeast of Santo Domingo.
8	The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a broad area of cloudiness and heavy weather rotating around the center of the storm.
15	Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.

Tabla 6.7: Resumen generado por la heurística 2

	Heurística 3
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
3	There is no need for alarm, Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.
6	Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
12	San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsi ded during the night.
13	On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

Tabla 6.8: Resumen generado por la heurística 3

Capítulo 7

Caso de Estudio: Resúmenes Multi-documento de Páginas Web de Información Turística

El tercer y último caso de estudio tiene como objetivo configurar el sistema de generación de resúmenes para elaborar resúmenes multi-documento de páginas web con información sobre destinos de interés turístico. El capítulo se organiza como sigue. En primer lugar, se presenta un breve estudio de las características del lenguaje y de la estructura del tipo de documentos que nos ocupa. En segundo lugar, se describe el proceso realizado para configurar el método genérico presentado en el Capítulo 4 para realizar resúmenes a partir de múltiples documentos con información sobre monumentos u otros lugares de visita turística.

7.1. Peculiaridades del Dominio y del Tipo de Documentos

El tipo de documentos con el que nos enfrentamos presenta ciertas características que los hacen especialmente interesantes como caso de estudio para un sistema automático de generación de resúmenes.

En primer lugar, el léxico turístico utiliza una terminología amplia procedente de diversos campos (geografía, economía, historia del arte, etc.). En

él encontramos palabras relativas a hostelería y restauración, transacciones comerciales, transportes, burocracia internacional, tiempo libre y espectáculos, historia y arte, etc. Se trata, pues, como en el caso de estudio anterior, de un lenguaje poco especializado: si bien es cierto que incluye un núcleo léxico más específico, que comprende términos técnicos relativos a las organizaciones turísticas, servicios, etc., en general, utiliza un vocabulario amplio y cotidiano.

En segundo lugar, el tipo concreto de documentos que nos ocupa (páginas web sobre monumentos o destinos de interés turístico) se caracteriza, fundamentalmente, por la amplia variedad de información presentada. En general, cuando se describe un monumento o una ciudad, dicha descripción incluye información sobre el tipo de lugar al que nos referimos (e.g. una iglesia o una fortaleza), su localización, la fecha de su construcción o fundación, información histórica o artística sobre el mismo, detalles sobre otros monumentos o lugares de interés en los alrededores, información sobre oficinas de turismo, horarios y precios de visita, planificación de excursiones, etc. Sin embargo, y sobre todo cuando se trata de páginas web, nos encontramos además con otro tipo de información, a menudo no relacionada con el lugar que se describe, que plantea problemas adicionales a la generación automática de resúmenes. Así, por ejemplo, muchas de estas páginas web incluyen un foro en el que los visitantes expresan sus opiniones o experiencias, y que a menudo contienen información de escaso interés para el lector que acude a ellas con el objetivo de recabar información para preparar su viaje. El sistema de resúmenes deberá, por tanto, discernir qué información de estos foros es importante y cuál no. Además, estas páginas web a menudo incluyen publicidad, tanto de la empresa que aloja la información como de otras empresas turísticas que se anuncian a través de la misma. Es el caso, por ejemplo, de las páginas de *tripadvisor.com*, en las que se anuncian hoteles y restaurantes situados en los alrededores del lugar descrito. Toda esta información, cuando se desea realizar un resumen con una elevada tasa de compresión, deberá sacrificarse en beneficio de otra más relacionada con el objeto del resumen.

En tercer lugar, y al tratarse de resúmenes multi-documento, la información procede de distintas páginas web que describen el mismo lugar, con lo que es previsible que parte del contenido se repita de unas páginas a otras.

7.2. Especialización del Método para el Dominio Turístico

Como en los casos de estudio anteriores, en esta sección se pretende detallar el proceso realizado para adaptar el algoritmo genérico para realizar resúmenes a partir de múltiples páginas web de información turística. La principal diferencia de este caso de estudio con respecto a los anteriores es que nos enfrentamos a una tarea de generación de resúmenes multi-documento. Esto significa, como ya se ha mencionado, que cada resumen ha de ser producido a partir de la información contenida en un número variable (pero siempre superior a uno) de documentos. Tal y como se estudiara en la Sección 2.4.1, la generación de resúmenes multi-documento presenta dos retos fundamentales. El primero de ellos, la agrupación de documentos similares, no será tratado en este caso de estudio. Suponemos, pues, que disponemos de los documentos ya agrupados, de tal modo que todos aquellos en un mismo grupo describen un mismo monumento o destino turístico. Por el contrario, sí será objeto de estudio el segundo de los retos, la eliminación de la redundancia.

El procesamiento a realizar para generar el resumen es equivalente al desarrollado en el caso de estudio de generación de resúmenes de noticias periodísticas. Se utiliza, por tanto, como parte del pre-procesamiento, la lista de WordNet para detectar y eliminar palabras vacías, y la lista de abreviaturas del *ANNIE Gazetteer* de GATE para expandir acrónimos y abreviaturas. Por tratarse de documentos con un vocabulario amplio y general, se utiliza la base de datos léxica WordNet (Sección 3.1.6) para representar los conceptos del dominio y el sistema de desambiguación WordNet::SenseRelate (Sección 3.6) para resolver posibles ambigüedades léxicas. Asimismo, se utiliza la *relación de similitud semántica* explicada en la Sección 6.2.4 para asociar los distintos conceptos en el grafo del documento.

La eliminación del contenido que se repite a través de los distintos documentos a resumir se realiza como un paso posterior a la generación del resumen, conforme a lo descrito en la Sección 4.8. Para ello, recordemos que, en primer lugar, se concatena el contenido de todos los documentos. A continuación, se ejecuta el algoritmo de generación de resúmenes, como si de un resumen mono-documento se tratase, con la configuración indicada. Como resultado, se obtiene un resumen intermedio, en el que es previsible la presencia de información repetida. Para soslayar esta limitación, en es-

te caso de estudio se utiliza la herramienta para detección de implicación textual (*textual entailment*) desarrollada por Ferrández *et al.* (2007) para identificar el contenido redundante. Si como resultado de la aplicación de dicha herramienta sobre cada par de oraciones del resumen se determina que una de ellas presenta una relación de implicación con respecto a la otra (es decir, la información contenida en una de las oraciones, desde una perspectiva semántica, se encuentra contenida en la otra), dicha oración se elimina del resumen. De este modo, se evita incluir información repetida.

Sin ánimo de exhaustividad, y puesto que el proceso es idéntico al detallado en el caso de estudio anterior, nos limitamos aquí a reproducir en las Tablas 7.1, 7.2 y 7.3 los resúmenes generados por las tres heurísticas de selección de oraciones para un conjunto de 10 páginas web de información sobre la Acrópolis de Atenas, páginas que se pueden consultar en el Apéndice B.3 de este documento, y que han sido obtenidas del corpus descrito en Aker y Gaizauskas (2009). Para elaborar estos resúmenes, se han considerado los siguientes valores de los parámetros que intervienen en el algoritmo:

- **Ratio de compresión**: 200 ± 10 palabras.
- **Porcentaje de *hub vertices***: 10 %.
- **Relaciones semánticas**: hiperonimia y similitud semántica (*jcn*).
- **Umbral de similitud**: 0.25.
- **Criterios adicionales de selección de oraciones**: ninguno.

Heurística 1
Acropolis (Gr akros, akron, edge, extremity + polis, city, pl. acropoleis) literally means city on the edge (or extremity)
In Greek, Acropolis means Highest City.
For purposes of defense, early settlers naturally chose elevated ground, frequently a hill with precipitous sides.
Although originating in the mainland of Greece, use of the acropolis model quickly spread to Greek colonies such as the Dorian Lato on Crete during the Archaic Period.
Because of its classical Greco-Roman style, the ruins of Mission San Juan Capistrano's Great Stone Church in California, United States has been called the American Acropolis.
The word Acropolis, although Greek in origin and associated primarily with the Greek cities Athens, Argos, Thebes, and Corinth (with its Acrocorinth), may be applied generically to all such citadels, including Rome, Jerusalem, Celtic Bratislava, many in Asia Minor, or even Castle Rock in Edinburgh.
The term acropolis is also used to describe the central complex of overlapping structures, such as plazas and pyramids, in many Mayan cities, including Tikal and Copán.
In Central Italy, many small rural communes still cluster at the base of a fortified habitation known as La Rocca of the commune.

Tabla 7.1: Resumen generado por la heurística 1

Heurística 2
Acropolis (Gr akros, akron, edge, extremity + polis, city, pl acropoleis) literally means city on the edge (or extremity). In Greek, Acropolis means Highest City. For purposes of defense, early settlers naturally chose elevated ground, frequently a hill with precipitous sides. I think it would be good for school children to learn to think and enjoy words. I love words and this is a great game. In many parts of the world, these early citadels became the nuclei of large cities, which grew up on the surrounding lower ground, such as modern Rome. The word Acropolis, although Greek in origin and associated primarily with the Greek cities Athens, Argos, Thebes, and Corinth (with its Acrocorinth), may be applied generically to all such citadels, including Rome, Jerusalem, Celtic Bratislava, many in Asia Minor, or even Castle Rock in Edinburgh. By: arizonalady on 06 february 09 Easy Challenging Relaxing Fast Paced Clicky Thinky This game is: Addictive , Good Replay , Original , Involved , Good Value , Kid Friendly I just love all word games. The most famous example is the Acropolis of Athens, which, by reason of its historical associations and the several famous buildings erected upon it (most notably the Parthenon), is known without qualification as the Acropolis.

Tabla 7.2: Resumen generado por la heurística 2

Heurística 3
Acropolis (Gr. akros, akron, edge, extremity + polis, city, pl. acropoleis) literally means city on the edge (or extremity). The Acropolis was designated as a UNESCO World Heritage site in 1987, for its, illustrating the civilizations, myths, and religions that flourished in Greece over a period of more than 1,000 years, the Acropolis, the site of four of the greatest masterpieces of classical Greek art - the Parthenon, the Propylaea, the Erechtheum, and the Temple of Athena Nike-can be seen as symbolizing the idea of world heritage. The Acropolis of Athens, a hill c.260 ft (80m) high, with a flat oval top c.500 ft (150m) wide and 1,150 ft (350m) long, was a ceremonial site beginning in the Neolithic Period and was walled before the 6th cent. B.C. by the Pelasgians. Devoted to religious rather than defensive purposes, the area was adorned during the time of Cimon and Pericles with some of the world's greatest architectural and sculptural monuments. This temple is the first building visitors see as they make their way up the Acropolis. The first stone temple to Athena, the patron goddess and protector of the city, was built on the Acropolis at the beginning of the 6th century B.C.

Tabla 7.3: Resumen generado por la heurística 3

Capítulo 8

Evaluación

El propósito de este capítulo es evaluar el comportamiento del método descrito en el Capítulo 4 a la hora de seleccionar oraciones de un documento en función de su importancia relativa para elaborar un resumen; así como realizar una comparación entre dicho comportamiento y el que presentan otros sistemas comerciales y prototipos de investigación. Asimismo, se persigue evaluar la capacidad del sistema para realizar su cometido en distintos dominios, con el objetivo de determinar si las características del conocimiento presente en un texto (especificidad, ambigüedad, formalismo, etc.), así como las características del propio documento (longitud, estructura, etc.) influyen en el rendimiento del sistema. Para ello, se ha realizado una evaluación en dos fases. La primera de ellas va dirigida a determinar los valores óptimos para cada uno de los parámetros que intervienen en el algoritmo; la segunda es una evaluación a gran escala siguiendo las métricas y directrices observadas en las conferencias *DUC* de los años 2004 y 2005 (Litkowski, 2004; Dang, 2005). Así mismo, se estudia el efecto de la ambigüedad léxica en la calidad de los resúmenes generados, mediante la integración en el sistema de distintos algoritmos de desambiguación y su evaluación. Todo lo anterior, a su vez, se realizará para cada uno de los dominios estudiados en los casos de estudio.

El capítulo se organiza en cinco secciones. La Sección 8.1 introduce la metodología de evaluación, incluyendo las métricas empleadas y las colecciones de documentos utilizadas en la evaluación de los distintos casos de estudio. La Sección 8.2 presenta los resultados experimentales obtenidos en la generación de resúmenes de artículos científicos en biomedicina, tanto por el método propuesto como por otros sistemas evaluados en las mismas

condiciones experimentales. La Sección 8.3 presenta estos mismos resultados para el caso de estudio de generación de resúmenes de noticias periodísticas. La Sección 8.4 presenta la evaluación del caso de estudio de generación de resúmenes de páginas web turísticas. Para concluir, la Sección 8.5 analiza y discute estos resultados.

8.1. Metodología de Evaluación

8.1.1. Métricas de Evaluación

Para la evaluación del sistema, se ha utilizado el paquete de evaluación intrínseca de resúmenes automáticos ROUGE (ver Sección 2.3.3.5). Los motivos por los que se han seleccionado estas métricas son, esencialmente, dos: en primer lugar, se trata de métricas automáticas, y por lo tanto, no requieren la intervención de jueces humanos que evalúen la calidad de los resúmenes; en segundo lugar, porque ROUGE se ha convertido en el estándar *de facto* desde que fuera elegido como el conjunto de métricas oficial para la evaluación de los sistemas participantes en las conferencias *DUC* (y posteriormente, en las conferencias *TAC*) desde 2004. En concreto, en este trabajo se han utilizado las métricas ROUGE-1, ROUGE-2, ROUGE-4, ROUGE-L, ROUGE-W-1.2 y ROUGE-S4. Conviene recordar que ROUGE realiza la evaluación comparando cada resumen generado por el sistema automático con uno o varios resúmenes *ideales* (*modelos* en la terminología de ROUGE), generados por humanos. Por lo tanto, su cálculo requiere disponer de una colección de documentos a resumir y de, al menos, un resumen modelo por cada documento de la colección.

Por otro lado, y para el caso de estudio de páginas web de información turística, se completa esta evaluación con la medida, por parte de jueces humanos, de diferentes características deseables encaminadas a valorar la legibilidad del resumen. En particular, se evalúan las características propuestas en las conferencias DUC y TAC, descritas en la Sección 2.3.3.7: *calidad gramátical*, *redundancia*, *claridad referencial*, *foco* y *estructura y coherencia*. Para ello, se ha solicitado a tres personas que puntúen 50 de los resúmenes automáticos seleccionados aleatoriamente, asignando a cada uno una puntuación de 1 a 5 conforme al grado de cumplimiento de cada una de las características anteriores.

8.1.2. Colecciones de Evaluación

Así pues, para evaluar el sistema en cada uno de los dominios de aplicación, será necesario disponer de sendas colecciones de evaluación.

En cuanto al primer caso de estudio se refiere, no se tiene constancia de la existencia de ningún corpus especialmente concebido para la evaluación de resúmenes de artículos biomédicos. Por este motivo, los trabajos realizados al respecto generalmente elaboran sus propias colecciones de evaluación *ad hoc*. Reeve *et al.* (2007), por ejemplo, utilizan una colección de 24 documentos extraídos de una base de datos de artículos científicos sobre ensayos clínicos en oncología. Sin embargo, no se especifican los artículos concretos utilizados. En este trabajo, se ha optado por utilizar como colección de evaluación un conjunto de 150 artículos científicos sobre biomedicina, seleccionados de forma aleatoria del corpus de BioMed Central. Dicho corpus contiene más de 60.000 artículos sobre investigación biomédica, disponibles además en formato XML, lo que permite identificar fácilmente las secciones que componen cada documento (título, *abstract*, referencias bibliográficas,etc.); además de otros objetos que pudieran estar incrustados en el documento, como tablas o imágenes. La Tabla 8.1 muestra la estructura simplificada de un documento del corpus. Tal y como los propios autores de ROUGE defienden (Lin, 2004a), el tamaño de la muestra seleccionada es suficiente para asegurar que los resultados obtenidos en la evaluación son estadísticamente significativos.

En cuanto a la elaboración de los resúmenes modelo con los que comparar los resúmenes automáticos, se ha pedido a estudiantes de quinto curso de medicina que generen por extracción un resumen de cada uno de los documentos de la colección de evaluación, de una longitud comprendida entre el 20 y el 30 por ciento de la longitud del documento original. La longitud del resumen ha sido elegida conforme a la afirmación comúnmente aceptada de que el tamaño de un resumen debería estar comprendido entre el 15 y el 35 por ciento del tamaño del documento de partida (Hovy, 2005). Por otra parte, se ha preferido dar a los expertos cierta libertad a la hora de decidir qué es o no es importante, y es por ello que se ha permitido que la longitud de los resúmenes oscile dentro de un intervalo. Dada la complejidad de la tarea encomendada, junto a la dificultad añadida de que los documentos están escritos en inglés, sólo se han obtenido 10 resúmenes válidos, y se ha decidido utilizar estos resúmenes en el proceso de parametrización del algoritmo. Para la evaluación final, se ha optado por utilizar los *abstracts* o resúmenes

```
<artículo>
  <título> <p>Increased capsaicin receptor TRPV1 in skin nerve fibres and related
      vanilloid receptors TRPV3 and TRPV4 in keratinocytes in human breast
      pain</p>
  </título>
  <autores>...</autores>
  <instituciones>...</instituciones>
  <fuente> BMC Women's Health </fuente>
  <issn> 1472-6874 </issn>
  <año> 2005 </año>
  <volume> 5 </volume>
  <issue> 1 </issue>
  <url> http://www.biomedcentral.com/1472-6874/5/2 </url>
  ...
  <abstract><p>Breast pain and tenderness affects 70 % of women at some...</p>
  </abstract>
  <cuerpo>
    <sec>
      <st><p>Background</p></st>
      <p>Breast pain is a common problem, which can affect up to 70 % ... </p>
    </sec>
    ...
    <sec>
      <st><p>List of abbreviations</p></st>
      <p>TRPV = transient receptor potential vanilloid;
    </sec>
    <sec>
      <st><p>Competing interests</p></st>
      <p>The author(s) declare that they have no competing interests.</p>
    </sec>
    <sec>
      <st> <p> Authors'contributions </p> </st>
      <p> PG and EW recruited patients, collected biopsies and ... </p>
    </sec>
  </cuerpo>
  <contribuciones> ...</contribuciones>
  <abreviaciones> ...</abreviaciones>
  <agradecimientos> ...</agradecimientos>
  <referencias>...</referencias>
</artículo>
```

Tabla 8.1: Estructura simplificada de un documento del corpus de BioMed Central

que acompañan a cada uno de los 150 artículos de la colección de evaluación, y que han sido elaborados por los propios autores de los artículos.

En relación al segundo caso de estudio, no ha sido necesario elaborar una colección de evaluación a medida, pues sí se encuentra disponible una colección especialmente construida para la evaluación de resúmenes automáticos sobre noticias periodísticas, que además ha sido y es ampliamente utilizada

por la comunidad científica con dicho propósito. Se trata de la colección de evaluación utilizada en la conferencia DUC del año 2002[1]. Dicha colección se compone de 567 artículos de noticias periodísticas en inglés, acompañados de dos o más resúmenes (dependiendo de la noticia en cuestión), realizados por expertos. Estás 567 noticias han sido extraídas de distintos periódicos y agencias de noticias (por ejemplo, *Financial Times* o *Associated Press*) y, dentro de estos, de diferentes secciones, por lo que los temas tratados en el conjunto de la colección son muy diversos (desastres naturales, enfrentamientos armados, evolución bursátil, entre otros). Los resúmenes que acompañan a cada una de las noticias han sido elaborados mediante abstracción, y tienen una longitud de 100 palabras, lo que supone aproximadamente entre el 15 % y el 35 % de la longitud de las noticias. La Tabla 8.2 muestra la estructura simplificada de un documento del corpus.

```
<doc>
  <docno>AP891018-0301 </docno>
  <fichero>AP-NR-10-18-89 1832EDT</fichero>
  <titular>Many Homeowners Not Insured;
        Analysts Say Disaster May Benefit Insurers</titular>
  <autor> By DEAN GOLEMBESKI </autor>
  <fuente> Associated Press Writer </fuente>
  <fecha> 10/19/89 </fecha>
  <cuerpo>
    <p>Most San Francisco-area homeowners may have to pay for damage from
    Tuesday's earthquake out of their own pockets, while insurance companies may
     reap long-term benefits from higher rates, industry spokesmen said...</p>
    <p> Only 15 percent to 20 percent of California homeowners have earthquake,
     insurance which typically requires a 10 percent deductible and costs between
    200to400 a year for a $100,000 home, according to ... </p>
  </cuerpo>
</doc>
```

Tabla 8.2: Estructura simplificada de un documento del corpus de evaluación de la conferencia DUC 2002

Al igual que en el dominio anterior, sólo 10 documentos han sido utilizados para la parametrización del algoritmo. De nuevo, los resúmenes modelo utilizados han sido generados mediante extracción. Para ello, se ha solicitado a una persona que seleccione las oraciones más importantes de cada una de las noticias hasta que la longitud del texto seleccionado supere las 100 palabras.

Finalmente, para la evaluación del tercer caso de estudio se ha utilizado

[1]DUC corpora. http://www-nlpir.nist.gov/projects/duc/data.html. Consultada el 1 de noviembre de 2010

la colección descrita en Aker y Gaizauskas (2009). Dicha colección consta de 308 imágenes a las que manualmente se les ha asignado el nombre del lugar o monumento que representan (por ejemplo, *Eiffel Tower* o *Akershus Fortress*), diez documentos o páginas web describiendo el objeto presentado en la imagen, y que han sido obtenidos automáticamente utilizando el motor de búsqueda *Yahoo!* y el nombre del objeto como consulta, y un número variable de hasta cuatro resúmenes modelo creados manualmente para cada una de las imágenes de la colección. Estos resúmenes modelo presentan una longitud entre 190 y 210 palabras. Tanto los documentos como los resúmenes se presentan como texto plano. Un ejemplo del conjunto de documentos del corpus que describen la Acrópolis de Atenas puede encontrarse en el Apéndice B.3 de este documento.

8.1.3. Parametrización del Algoritmo

Antes de realizar la evaluación definitiva, ha sido necesario realizar una evaluación que podríamos calificar como preliminar, con el objetivo de determinar los valores óptimos de los diferentes parámetros que intervienen en el algoritmo de generación de resúmenes. En particular, se pretende dar respuesta a las siguientes preguntas:

1. ¿Qué porcentaje de vértices deben tomarse como *hub vertices* en el algoritmo de agrupamiento? (ver Sección 4.5).

2. ¿Qué combinación de relaciones semánticas produce mejores resultados? (ver Sección 4.4).

3. ¿Qué coeficiente de similitud (Jaccard *vs.* Dice-Sorensen) se comporta mejor a la hora de determinar los pesos de las aristas del grafo del documento? (ver Sección 4.4).

4. En caso de utilizar la relación de similitud semántica, ¿qué umbral de similitud se debería establecer? (ver Sección 6.2.4).

5. El uso de criterios estadísticos tradicionales, como la posición de las oraciones y su similitud con el título, combinados con el algoritmo propuesto, ¿se traduce en una mejora de la calidad de los resúmenes generados? (ver Sección 4.7).

6. Finalmente, ¿cuál de las tres heurísticas de selección de oraciones produce mejores resúmenes? (ver Sección 4.7).

8.1.4. Comparación con otros Sistemas

Con el objetivo de comparar el comportamiento del algoritmo presentado frente al de otros sistemas de generación de resúmenes, se han evaluado, utilizando las mismas colecciones de evaluación y respetando las mismas condiciones experimentales, otros sistemas, incluyendo prototipos de investigación y aplicaciones comerciales, cuyos principios y funcionamiento se explican a continuación.

LexRank (Sección 2.2.3) es un algoritmo basado en grafos que permite determinar la relevancia de las oraciones en el conjunto del documento, lo que no es más que otro modo de definir la generación de resúmenes mediante extracción. Se encuentra integrado dentro de la plataforma de generación de resúmenes MEAD[2]. Para su uso en la evaluación que nos ocupa, se han utilizado los valores por defecto de los distintos parámetros que intervienen en la generación del resumen.

SUMMA (Saggion, 2008) es una herramienta para la generación de resúmenes mono-documento y multi-documento, desarrollada por Horacio Saggion en la Universidad de Sheffield. SUMMA proporciona la implementación de un conjunto de métodos estadísticos y basados en similitud que pueden ser utilizados, de manera independiente o combinados, para estimar la relevancia de las oraciones de un documento. En concreto, en esta evaluación se han utilizado los siguientes: la posición de las oraciones en el texto y dentro del propio párrafo, su similitud con las secciones de título y *abstract*, la similitud con la primera oración, y la frecuencia de los términos en el documento.

Microsoft Autosummarize[3] es una herramienta para la generación de resúmenes incluida en Microsoft Word que implementa un algoritmo basado en la frecuencia de los términos del documento, pero cuyos detalles concretos de implementación no son públicos.

Para el caso de estudio de noticias periodísticas se muestran, además, los resultados obtenidos por otros sistemas que participaron en la tarea competitiva de la conferencia DUC 2002, cuya colección de evaluación se utiliza en este trabajo (DUC 19, DUC 21, DUC 27, DUC 28 y DUC 29). Sin ánimo

[2]MEAD. http://www.summarization.com/mead/. Consultada el 1 de noviembre de 2010

[3]Microsoft Corporation. Microsoft Office online: automatically summarize a document. http://office.microsoft.com/en-us/word/HA102552061033.aspx. Consultada el 1 de noviembre de 2010

de exhaustividad, DUC 19 utiliza plantillas de representación de diferentes temas para extraer la información relevante, DUC 21, 27 y 28 combinan distintas técnicas de aprendizaje máquina para determinar el mejor conjunto de atributos para la extracción de las oraciones (frecuencia de términos, posición de las oraciones, etc.) y DUC 29 utiliza cadenas léxicas. Además, se muestran los resultados obtenidos y publicados por otros sistemas evaluados sobre la misma colección de documentos: LeLSA+AR, un sistema basado en frecuencias de términos mejorado con información sobre referencias anafóricas (Steinberger et al., 2007) y Freq+TextEnt, un método de extracción de oraciones que combina criterios simples como la frecuencia de términos con técnicas de implicación textual (Lloret et al., 2008).

Para el caso de estudio de páginas web turísticas, se incluyen los resultados obtenidos por otros dos sistemas, uno basado en modelos de lenguaje (Language Models) y otro en frecuencias de palabras y grupos nominales (COMPENDIUM), ambos publicados en (Plaza, Lloret, y Aker, 2010) y evaluados sobre la misma colección. También se muestran los resultados obtenidos por la herramienta de generación de resúmenes MEAD[4], utilizando como atributos para la extracción la posición de las oraciones, su longitud, y su "centralidad" en relación al cluster de documentos al que pertenece.

Además, se han implementado dos líneas base. Una línea base sirve como indicador de referencia para la evaluación, ya que da una idea del rendimiento esperado de una implementación *naïve* de un generador de resúmenes. La primera línea base consiste en seleccionar las n primeras oraciones del documento, y comúnmente se conoce como *línea base posicional* (*Lead*). La segunda consiste en seleccionar aleatoriamente n oraciones del documento, y se conoce habitualmente con el nombre de *línea base aleatoria* (*Random*).

8.2. Evaluación del Caso de Estudio de Generación de Resúmenes Mono-documento de Artículos Científicos en Biomedicina

En primer lugar, evaluamos el rendimiento del método de generación de resúmenes con la configuración presentada en el Capítulo 5 para el tratamiento de artículos científicos en biomedicina. Como ya se ha mencionado, la evaluación incluye, por un lado, la determinación de los valores apropia-

[4]MEAD. http://www.summarization.com/mead/

dos para cada uno de los parámetros de configuración del algoritmo; y por otro lado, una evaluación a gran escala cuyo objetivo principal es comparar la calidad de los resúmenes generados por nuestro método con la de aquellos producidos por otros sistemas.

8.2.1. Parametrización

Con el propósito de responder a las cuestiones planteadas en la Sección 8.1.3 en relación a la parametrización del algoritmo, se ha llevado a cabo una experimentación preliminar sobre un subconjunto de 10 documentos del corpus de evaluación. Es importante recordar que, para estos experimentos, los resúmenes modelo no son los *abstracts* que acompañan a los documentos, sino extractos elaborados por expertos en biomedicina.

8.2.1.1. Definición del Número Óptimo de *Hub Vertices* y del Mejor Conjunto de Relaciones Semánticas

El objetivo del primer grupo de experimentos es determinar la mejor combinación de relaciones semánticas para la construcción del grafo del documento (ver Sección 4.4), junto al porcentaje óptimo de vértices utilizados como *hub vertices* en el algoritmo de agrupamiento (ver Sección 4.5). Nótese que ambos parámetros deben ser evaluados de forma conjunta, puesto que las relaciones consideradas influyen sobre la conectividad del grafo del documento y, por lo tanto, también sobre el número óptimo de *hub vertices*. En la realización de estos experimentos, se ha utilizado el coeficiente de Jaccard para el cálculo de los pesos de las aristas del grafo (ver Sección 4.5) y ninguno de los criterios de selección de oraciones estudiados en la Sección 4.7 (i.e. criterio posicional y de similitud con el título).

Las Tablas 8.3, 8.4 y 8.5 muestran las puntuaciones medias de ROUGE obtenidas por los resúmenes generados con distintas combinaciones de valores de los parámetros estudiados, para cada una de las tres heurísticas implementadas para la selección de oraciones. En estas tablas, los mejores resultados para cada conjunto de relaciones se muestran en cursiva, mientras que los mejores resultados globales por heurística se muestran en negrita. En caso de conflicto entre las distintas métricas ROUGE a la hora de elegir entre dos configuraciones, se considera mejor aquella con mayor valor de ROUGE-2, por ser esta métrica la que mayor correlación ha demostrado con respecto a las valoraciones de los jueces en las conferencias DUC.

En relación a la primera de las heurísticas, en la Tabla 8.3 se puede observar que los resultados obtenidos son similares cuando se utilizan todas las relaciones semánticas junto a un porcentaje de *hub vertices* del 5 % y cuando se usan únicamente las relaciones de hiperonimia y *related to* junto al 2 % de los vértices del grafo del documento como vértices *hub*.

Heurística 1							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	0.7770	*0.6229*	*0.5530*	0.7673	0.2339	*0.5812*
	5 %	0.7739	0.5802	0.5133	0.7638	0.2328	0.5625
	10 %	*0.7794*	0.5721	0.4997	*0.7693*	*0.2354*	0.5528
	20 %	0.7730	0.4791	0.4037	0.7612	0.2307	0.4615
Hiperonimia + *Associated with*	2 %	*0.7810*	*0.6188*	*0.5468*	*0.7715*	*0.2355*	*0.6011*
	5 %	0.7129	0.6129	0.5425	0.7004	0.2110	0.5950
	10 %	0.7422	0.6168	0.5444	0.7299	0.2215	0.5977
	20 %	0.6680	0.6014	0.5309	0.6557	0.1910	0.5858
Hiperonimia + *Related to*	2 %	*0.7829*	**0.6275**	**0.5599**	*0.7740*	*0.2366*	**0.6088**
	5 %	0.7476	0.5730	0.5024	0.7359	0.2220	0.5578
	10 %	0.7474	0.5804	0.5095	0.7374	0.2246	0.5635
	20 %	0.6961	0.5181	0.4474	0.6811	0.1977	0.5012
Hiperonimia + *Associated with* + *Related to*	2 %	0.7752	0.6148	0.5438	0.7657	0.2333	0.5960
	5 %	**0.7853**	*0.6250*	*0.5549*	**0.7751**	**0.2371**	*0.6068*
	10 %	0.7558	0.5960	0.5268	0.7442	0.2220	0.5776
	20 %	0.7666	0.6042	0.5331	0.7545	0.2317	0.5870

Tabla 8.3: Resultados de la evaluación conjunta de las relaciones semánticas y del porcentaje de *hub vertices* para la heurística 1

La Tabla 8.4 sintetiza los valores de ROUGE para la segunda heurística. Se puede comprobar que el mejor resultado se obtiene cuando se considera el conjunto completo de relaciones semánticas (i.e. *associated with*, *related to* e hiperonimia), y se utiliza un porcentaje del 10 % de los vértices del grafo del documento como vértices *hub*. Así pues, si bien el conjunto de relaciones se puede considerar equivalente, o cuanto menos muy cercano, al obtenido para la primera heurística, el número de *hub vertices* se duplica. La razón puede encontrarse en las hipótesis que subyacen en cada una de las heurísticas y en el tipo de resúmenes que pretenden generar cada una de ellas. Recordemos que el objetivo de la segunda heurística es producir un resumen que cubra, por igual, todos los temas tratados en el documento, independientemente de su importancia. Por ello, no es suficiente considerar como centroides de los grafos aquellos conceptos que representan el tema principal del documento, sino también otros que se refieren a información complementaria.

Heurística 2							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	0.7673	*0.6179*	**0.5497**	0.7575	0.2303	*0.5821*
	5 %	*0.7770*	0.5909	0.5201	*0.7673*	*0.2339*	0.5728
	10 %	0.7686	0.5916	0.5153	0.7577	0.2282	0.5711
	20 %	0.7636	0.4984	0.4194	0.7519	0.2262	0.4794
Hiperonimia + *Associated with*	2 %	*0.7758*	0.6012	0.5300	*0.7667*	**0.2348**	0.5977
	5 %	0.7589	*0.6168*	*0.5444*	0.7483	0.2247	0.5831
	10 %	0.7564	0.6035	0.5308	0.7451	0.2257	*0.6137*
	20 %	0.6851	0.5952	0.5260	0.6712	0.1970	0.5784
Hiperonimia + *Related to*	2 %	*0.7734*	0.6143	0.5462	*0.7639*	*0.2332*	*0.5960*
	5 %	0.7585	*0.6151*	*0.5391*	0.7477	0.2260	0.5806
	10 %	0.6955	0.5193	0.4427	0.6820	0.1993	0.4989
	20 %	0.7296	0.5615	0.4852	0.7162	0.2082	0.5391
Hiperonimia + *Associated with* + *Related to*	2 %	0.7752	0.6148	0.5438	0.7657	*0.2333*	0.5960
	5 %	0.7751	0.6099	0.5404	0.7651	*0.2333*	0.5941
	10 %	**0.7795**	**0.6189**	*0.5491*	**0.7677**	0.2329	**0.6135**
	20 %	0.7777	0.6166	0.5449	0.7675	0.2326	0.5997

Tabla 8.4: Resultados de la evaluación conjunta de las relaciones semánticas y del porcentaje de *hub vertices* para la heurística 2

Por último, en relación a la tercera de las heurísticas, se puede observar en la Tabla 8.5 que el mejor resultado se obtiene, al igual que ocurriera para la segunda heurística, cuando se considera el conjunto completo de relaciones semánticas, aunque ahora el número óptimo de *hub vertices* es igual al 5 % de los vértices del grafo del documento. Por lo tanto, el número de *hub vertices* es un valor intermedio entre los obtenidos para las dos primeras heurísticas. Recordemos que el objetivo de esta heurística es seleccionar, prioritariamente, la información relacionada con el tema principal del documento, aunque permitiendo también la inclusión de cierta información secundaria. Se trata, pues, de un compromiso entre las dos heurísticas anteriores que el algoritmo de agrupamiento parece reconocer.

A la vista de estos resultados, parece obvio que tanto el conjunto de relaciones como el número de *hub vertices* dependen de la heurística, y por tanto, del tipo de resumen que se desee elaborar. En general, el número óptimo de vértices *hub* oscila entre el 2 % y el 10 % del número total de vértices del grafo del documento. En cuanto al mejor conjunto de relaciones semánticas se refiere, en general, el comportamiento es mejor cuanto más conexo es el grafo del documento; es decir, cuantas más relaciones se utilizan para construirlo. Por otro lado, quizás el resultado más interesante sea la

Heurística 3							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	0.7752	*0.6208*	*0.5521*	0.7657	0.2333	*0.6016*
	5 %	0.7745	0.6093	0.5368	0.7647	0.2332	0.5924
	10 %	*0.7817*	0.6048	0.5287	*0.7721*	*0.2363*	0.5853
	20 %	0.7752	0.5931	0.5218	0.7630	0.2309	0.5759
Hiperonimia + *Associated with*	2 %	*0.7796*	0.6148	0.5438	*0.7704*	*0.2355*	0.5960
	5 %	0.7761	*0.6251*	0.5533	0.7636	0.2329	0.6068
	10 %	0.7728	0.5425	*0.5555*	0.7593	0.2295	*0.6137*
	20 %	0.7623	0.6070	0.5344	0.7505	0.2283	0.5891
Hiperonimia + *Related to*	2 %	0.7810	0.6223	0.5538	0.7719	0.2357	0.6041
	5 %	*0.7827*	*0.6234*	*0.5547*	*0.7721*	*0.2360*	*0.6062*
	10 %	0.7594	0.5888	0.5149	0.7482	0.2248	0.5698
	20 %	0.7625	0.5991	0.5263	0.7506	0.2295	0.5819
Hiperonimia + *Associated with* + *Related to*	2 %	0.7749	0.6146	0.5438	0.7655	0.2334	0.5962
	5 %	**0.7886**	**0.6324**	**0.5635**	**0.7793**	**0.2388**	**0.6151**
	10 %	0.7830	0.6277	0.5572	0.7727	0.2338	0.6086
	20 %	0.7791	0.6151	0.5443	0.7674	0.2346	0.5981

Tabla 8.5: Resultados de la evaluación conjunta de las relaciones semánticas y del porcentaje de *hub vertices* para la heurística 3

observación de cierta tendencia a aumentar el número óptimo de vértices concentradores conforme aumenta el número de relaciones consideradas.

8.2.1.2. Definición de la Mejor Combinación de Criterios de Selección de Oraciones

El segundo grupo de experimentos persigue investigar si el uso de los criterios tradicionales de posición de la oración y similitud con el título permiten mejorar el resultado conseguido por el generador de resúmenes basado en grafos semánticos, utilizando la Ecuación 4.11 definida en la Sección 4.7 para seleccionar las oraciones del resumen. Para realizar estos experimentos, el porcentaje de vértices utilizados como *hubs* en el algoritmo de agrupamiento se ha establecido al 5 % para las heurísticas 1 y 3, y al 10 % para la heurística 2. Además, se han utilizado todas las relaciones (i.e. hiperonimia, asociación entre tipos semánticos y relación entre conceptos) en la construcción del grafo del documento y el coeficiente de Jaccard para el cálculo de los pesos de las aristas del grafo.

Las Tablas 8.6, 8.7 y 8.8 muestran las puntuaciones obtenidas por los resúmenes generados con los distintos sistemas que resultan de combinar nuestro método con los dos criterios estadísticos mencionados. En cuanto a

los pesos utilizados para ponderar los criterios en la función de selección (i.e. los parámetros λ, θ y χ definidos en la Ecuación 4.11), sus valores óptimos (determinados empíricamente) se muestran en estas mismas tablas.

Los resultados presentados en la Tabla 8.6 muestran que la heurística 1 se comporta mejor cuando no se combina con ninguno de los criterios considerados. De entre estos dos criterios, la similitud con el título produce unos resultados ligeramente mejores que la posición de la oración.

Heurística 1									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	**0.7853**	**0.6250**	0.5549	**0.7751**	**0.2371**	**0.6068**
Grafos Sem. + Posición	0.9	0.1	0.0	0.7826	0.6202	0.5471	0.7738	0.2371	0.5998
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.7817	0.6233	**0.5581**	0.7729	0.2370	0.6056
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	0.7823	0.6198	0.5490	0.7736	0.2371	0.6029

Tabla 8.6: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 1

A diferencia de lo que ocurriera con la heurística 1, la Tabla 8.7 muestra que la heurística 2 se comporta mejor cuando se combina con ambos criterios (i.e. la posición de la oración en el documento y su similitud con el título). De nuevo, si se analizan individualmente los dos criterios, se observa que la similitud con el título produce mejores resultados que la posición de la oración.

Heurística 2									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	0.7795	0.6189	0.5491	0.7677	0.2329	**0.6135**
Grafos Sem. + Posición	0.9	0.1	0.0	0.7804	0.6113	0.5395	0.7711	0.2364	0.5967
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.7786	0.6201	0.5505	0.7690	0.2354	0.6048
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	**0.7823**	**0.6225**	**0.5530**	**0.7737**	**0.2375**	0.5956

Tabla 8.7: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 2

Finalmente, podemos observar en la Tabla 8.8 que, al igual que en el caso de la heurística 1, la heurística 3 se comporta mejor cuando no se combina con otros criterios. Además, si se consideran por separado los dos criterios estudiados, se comprueba que el criterio de similitud con el título produce resultados ligeramente mejores que el criterio posicional.

Heurística 3									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	**0.7886**	**0.6324**	**0.5635**	**0.7793**	**0.2388**	**0.6151**
Grafos Sem. + Posición	0.9	0.1	0.0	0.7860	0.6149	0.5433	0.7773	0.2385	0.5984
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.7801	0.6173	0.5543	0.7711	0.2362	0.6038
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	0.7808	0.6171	0.5463	0.7720	0.2366	0.6005

Tabla 8.8: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 3

A modo de resumen, la Tabla 8.9 recopila los mejores resultados obtenidos por cada una de las heurísticas, así como la combinación de criterios de selección de oraciones que dan como fruto estos resultados.

Heurística	**Criterios**	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 1	Grafos Sem.	0.7853	0.6250	0.5549	0.7751	0.2371	0.6068
Heurística 2	Grafos Sem. + Pos. + Tit.	0.7823	0.6225	0.5530	0.7737	0.2375	0.5956
Heurística 3	Grafos Sem.	**0.7886**	**0.6324**	**0.5635**	**0.7793**	**0.2388**	**0.6151**

Tabla 8.9: Combinación óptima de criterios de selección de oraciones para cada heurística

Las configuraciones obtenidas por las distintas heurísticas no son, en absoluto, inesperadas. Hay que tener en cuenta que la información presente en los resúmenes modelo que se utilizan en la evaluación, obedeciendo a las instrucciones dadas a los expertos encargados de elaborarlos, está principalmente relacionada con el tema principal del documento. Por tanto, la heurística 2, por su propia definición, se encuentra en clara desventaja frente a las otras heurísticas. De hecho, ésta es la heurística que peores resultados obtiene, incluso cuando se apoya o beneficia de la información proporcionada por los otros criterios (posición de la oración y similitud con el título).

No quiere decir esto que sea peor que las demás, simplemente obedece a un objetivo distinto. Además, se observa que el criterio posicional, en general, no ayuda a mejorar los resultados, lo que sugiere que en un artículo científico el orden en que se presenta la información no indica necesariamente la importancia de la misma.

8.2.1.3. Definición del Mejor Coeficiente de Similitud

A continuación, queremos comprobar cuál de los dos índices de similitud implementados en la arquitectura de generación de resúmenes para el cálculo de los pesos de las aristas del grafo del documento (i.e. los coeficientes de similitud de Jaccard y de Dice-Sorensen) produce mejores resultados en relación a la calidad informativa de los resúmenes finales generados.

Para ello, repetimos los experimentos mostrados en la Tabla 8.9, esta vez utilizando el coeficiente de similitud de Dice-Sorensen. La Tabla 8.10 muestra los resultados obtenidos en estos nuevos experimentos. Se observa cómo, para todas las heurísticas, la calidad de los resúmenes generados disminuye. No obstante, las diferencias no son significativas, dado que ambos coeficientes producen pesos muy similares.

Heurística	**Criterios**	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 1	Grafos Sem.	0.7728	0.6143	0.5324	0.7680	0.2292	0.5981
Heurística 2	Grafos Sem. + Pos. + Tit.	0.7704	0.6146	0.5546	0.7658	0.2264	0.5928
Heurística 3	Grafos Sem.	**0.7823**	**0.6285**	**0.5586**	**0.7748**	**0.2352**	**0.6123**

Tabla 8.10: Combinación óptima de criterios de selección de oraciones para cada heurística, utilizando el coeficiente de similitud de Dice-Sorensen para el cálculo de los pesos de las aristas del grafo del documento

8.2.1.4. Recopilación: Parametrización Óptima del Algoritmo

La Tabla 8.11 recopila los valores óptimos de los distintos parámetros estudiados para cada una de las tres heurísticas de selección de oraciones. Se puede comprobar que estos valores dependen, en cierto grado, de la heurística utilizada para la selección final de las oraciones, lo que a su vez no resulta sorprendente puesto que tales heurísticas persiguen distintos objetivos y aspiran a capturar oraciones con diferentes propiedades, y por tanto, generar distintos tipos de resúmenes. No obstante, se pueden extraer algunas conclusiones generales. En primer lugar, se observa que la relación

related to por sí sola resulta insuficiente a la hora de extender el grafo del documento. La razón parece ser que se trata de una relación que no se da con una elevada frecuencia entre los conceptos inmersos en los documentos analizados. El resultado es un grafo excesivamente inconexo que invalida la hipótesis de partida (i.e. no se trata pues, de una red libre de escala). En segundo lugar, el uso del criterio posicional no aporta información de utilidad a la hora de determinar la relevancia de las oraciones. El motivo está relacionado con la naturaleza de los documentos utilizados para la evaluación: se ha comprobado que muchos de los documentos del corpus comienzan la redacción introduciendo el problema desde un punto de vista muy general, antes de presentar el problema concreto y las propuestas de solución. Esta información genérica del problema no es, a priori, significativa para el resumen.

Heurística	**Conjunto de Relaciones**	**Nº de *Hub Vertices***	**Criterios de Selección**	**Coeficiente de Similitud**
Heurística 1	Hiperonimia *Related to* *Associated with*	2 %-5 %	Grafos Semánticos	Jaccard
Heurística 2	Hiperonimia + *Related to* *Associated with*	10 %	Grafos Semánticos + Posición + Similitud Título	Jaccard
Heurística 3	Hiperonimia + *Related to* *Associated with*	5 %	Grafos Semánticos	Jaccard

Tabla 8.11: Recopilación: Mejor parametrización por heurística

8.2.2. Estudio del Efecto de la Ambigüedad Léxica

Con el objetivo de determinar en qué medida afecta la ambigüedad léxica presente en el documento de entrada a la calidad de los resúmenes generados, se han realizado distintos experimentos utilizando distintos algoritmos de desambiguación. Para ello, se han generado resúmenes automáticos de los 150 documentos del corpus de BioMed Central utilizando cuatro estrategias distintas de desambiguación:

1. Seleccionar el primero de los posibles candidatos recuperados por MetaMap *(1er candidato)*, lo que en realidad equivale a no realizar desambiguación alguna.

2. Utilizar el propio algoritmo de desambiguación de MetaMap, que se invoca mediante la opción *-y (MetaMap -y)*.
3. Utilizar la versión estándar del algoritmo Personalized PageRank *(PPR)*.
4. Utilizar la versión palabra-por-palabra del algoritmo Personalized PageRank *(PPR-w2w)*.

Sistema	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 1 - *1er Candidato*	0.7514	0.3304	0.1255	0.6700	0.1921	0.3128
Heurística 2 - *1er Candidato*	0.7305	0.3093	0.1115	0.6528	0.1811	0.2856
Heurística 3 - *1er Candidato*	0.7504	0.3283	0.1237	0.6689	0.1915	0.3117
Heurística 1 - *MetaMap -y*	**0.7724**	**0.3453**	**0.1327**	**0.6702**	**0.1936**	**0.3189**
Heurística 2 - *MetaMap -y*	**0.7772**	0.3421	0.1240	**0.6784**	**0.1969**	0.3205
Heurística 3 - *MetaMap -y*	**0.7845**	**0.3538**	**0.1308**	**0.6813**	**0.1983**	**0.3267**
Heurística 1 - *PPR*	0.7692	0.3383	0.1236	0.6682	0.1933	0.3150
Heurística 2 - *PPR*	0.7718	0.3380	0.1220	0.6684	0.1935	0.3145
Heurística 3 - *PPR*	0.7737	0.3419	0.1228	0.6709	0.1937	0.3178
Heurística 1 - *PPR-w2w*	0.7704	0.3379	0.1238	0.6680	0.1926	0.3108
Heurística 2 - *PPR-w2w*	0.7751	**0.3438**	**0.1255**	0.6716	0.1965	**0.3210**
Heurística 3 - *PPR-w2w*	0.7804	0.3530	0.1294	0.6754	0.1966	0.3262

Tabla 8.12: Evaluación del sistema de generación de resúmenes para distintas estrategias de desambiguación léxica

A la vista de los resultados de la Tabla 8.12, resulta evidente que el uso de desambiguación mejora las puntuaciones obtenidas por todas las heurísticas de generación de resúmenes, aunque esta diferencia es más evidente en el caso de la segunda heurística. La versión estándar del algoritmo *PPR* mejora significativamente (prueba de los signos de Wilcoxon, $p < 0{,}01$) en las métricas ROUGE-1, ROUGE-2 y ROUGE-4 para las heurísticas 1 y 3 con respecto a la estrategia del *1er candidato*, y en todas las métricas para la heurística 2. Por su parte, los algoritmos *PPR-w2w* y *MetaMap -y* lo hacen para todas las métricas y todas las heurísticas. En cuanto al algoritmo de desambiguación que mejor se comporta, la opción *-y* de MetaMap produce, para todas las métricas ROUGE, el mejor resultado en las heurísticas 1 y 3, mientras que el algoritmo *PPR-w2w* es el que mejor se comporta en el caso de la segunda heurística.

8.2.3. Comparación con otros Sistemas

El objetivo de esta sección es comparar los resultados del método propuesto en la generación de resúmenes de artículos científicos en biomedicina con

aquellos obtenidos por otros sistemas. Para ello, se han generado resúmenes automáticos utilizando LexRank, SUMMA y AutoSummarize. Asimismo, se han generado resúmenes posicionales (Lead) y resúmenes aleatorios (Random). La Tabla 8.13 recopila el resultado obtenido por todos estos sistemas para las diferentes métricas de evaluación, así como los obtenidos por las tres versiones de nuestro sistema utilizando la parametrización previamente determinada (Sección 8.2.1.4) y el algoritmo de desambiguación de MetaMap. En esta tabla, los resultados se muestran ordenados según el valor de ROUGE-2, y el mejor resultado obtenido para cada métrica se muestra en negrita.

Sistema	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 3	**0.7845**	**0.3538**	0.1308	**0.6813**	**0.1983**	**0.3267**
Heurística 2	0.7772	0.3421	0.1240	0.6784	0.1969	0.3205
Heurística 1	0.7724	0.3453	**0.1327**	0.6702	0.1936	0.3189
LexRank	0.7317	0.3248	0.1229	0.6508	0.1873	0.3097
SUMMA	0.7123	0.3187	0.1259	0.6320	0.1812	0.2989
AutoSummarize	0.5994	0.2446	0.0886	0.4874	0.1380	0.2318
Lead	0.6483	0.2566	0.0905	0.5751	0.1621	0.2646
Random	0.4998	0.1777	0.0540	0.4362	0.1207	0.2315

Tabla 8.13: Comparación de los resultados con los obtenidos por otros sistemas

Se observa en la Tabla 8.13 que las tres versiones del algoritmo propuesto producen mejores resultados que el resto de sistemas estudiados. Con el objetivo de evaluar si existen diferencias estadísticamente significativas entre los distintos sistemas, se ha realizado el test de los signos de Wilcoxon. Con un nivel de confianza del 99 %, todas las heurísticas son significativamente mejores que el resto de sistemas estudiados para todas las métricas ROUGE consideradas. Asimismo, la heurística 3 presenta mejores resultados que las otras dos heurísticas, aunque en este caso, las diferencias no son estadísticamente significativas.

8.2.4. Discusión de los Resultados

Los resultados presentados en la Tabla 8.13 parecen indicar que el uso de conceptos específicos del dominio, junto con un algoritmo de desambiguación, mejora la calidad de los resúmenes automáticos. Sin embargo, y a la vista de estos resultados, cabe preguntarse el por qué de las diferencias entre las heurísticas y con respecto al resto de sistemas. En concreto, llama

la atención el hecho de que la heurística 1 se comporte peor que las otras. Esto ocurre, además, únicamente en la evaluación final, pero no ocurrían en la parametrización. La razón principal parece estar en el tipo de documentos utilizados, para los que generalmente no es suficiente con mostrar en el resumen la información relacionada con el tema principal del documento, sino que es necesario incluir otra información secundaria o dependiente que, en ocasiones, pudiera ser de interés para el lector (Reeve, Han, y Brooks, 2007). Así se refleja, de hecho, en los *abstracts* de los artículos utilizados como modelos en la evaluación. Esta es, pues, también la explicación de los buenos resultados obtenidos por la tercera heurística. En cuanto a la segunda heurística se refiere, y a pesar de que su principal objetivo es capturar todo tipo de información (no sólo información principal), lo cierto es que el examen de los resúmenes generados por esta heurística demuestra que esto no siempre es así, y que a menudo presenta resultados similares a los de las heurísticas 1 y 3. Desde luego, el hecho de haber combinado esta heurística con los criterios posicional y de similitud con el título en los experimentos explica, al menos en parte, esta desviación de su objetivo inicial.

Por otro lado, el análisis de las puntuaciones ROUGE obtenidas por cada uno de los documentos del corpus en los distintos experimentos muestra que existen importantes divergencias entre las mismas. Esto no se observa en las tablas anteriores, puesto que presentan los resultados medios de los experimentos, pero se puede apreciar en la Tabla 8.14, donde se muestra la desviación típica de los valores de las diferentes métricas ROUGE en la colección de evaluación, para los resúmenes obtenidos con la heurística 3.

Métrica	**Desviación típica**
ROUGE-1	0.0813
ROUGE-2	0.1228
ROUGE-4	0.1187
ROUGE-L	0.0946
ROUGE-W-1.2	0.0497
ROUGE-S4	0.1074

Tabla 8.14: Desviación típica en los resultados de las distintas métricas ROUGE para los resúmenes generados por la heurística 3

Con el objetivo de vislumbrar las razones de tales diferencias, se han estudiado en detalle los dos casos extremos, el "mejor documento" y el "peor documento"; es decir, el documento con las puntuaciones más bajas de ROU-

GE, y el documento con las puntuaciones más altas. El mejor documento se corresponde con uno de los de mayor longitud en el corpus, mientras que el peor documento es uno de los más cortos (6 páginas *vs.* 3 páginas). De acuerdo con la hipótesis de partida por la que asumimos que el grafo del documento se comporta como una red libre de escala, este resultado no es de extrañar, puesto que dicha hipótesis resulta más válida cuanto mayor sea la red considerada (en este caso, el grafo del documento). Otra diferencia interesante entre ambos documentos se encuentra en el tema subyacente. Mientras que el mejor documento pertenece a la publicación *BMC Biochemistry* y trata sobre las reacciones de ciertos tipos de proteínas sobre las membranas sinápticas cerebrales, el peor documento pertenece a la publicación *BMC Bioinformatics* y trata sobre el uso de coincidencias de patrones para realizar búsquedas en bases de datos. Se ha podido verificar que la terminología utilizada para la identificación de conceptos (i.e. UMLS) cubre mejor el vocabulario presente en el primer documento que el del segundo, en términos tanto de los conceptos identificados como del número de relaciones reconocidas. También se ha podido observar que el mejor caso contiene una gran cantidad de bi-gramas, tri-gramas en incluso cuatri-gramas que se corresponden con un único concepto en UMLS, y que aparecen tanto en el cuerpo del documento como en el *abstract*, lo que obviamente influye positivamente en los resultados de las métricas de evaluación. Finalmente, en el peor documento, el uso de sinónimos es muy frecuente, algo que prácticamente no ocurre en el mejor documento. Se utilizan, por ejemplo, términos distintos para designar un mismo concepto en el *abstract* que en el cuerpo del documento (*pattern matching vs. string searching*). No resulta difícil darse cuenta que, puesto que las métricas ROUGE calculan el número de términos coincidentes entre el resumen automático y el modelo, aquellos resúmenes que contengan sinónimos de los términos presentes en el *abstract* resultarán injustamente penalizados.

Otro problema identificado tiene que ver con la frecuencia de acrónimos y abreviaturas en los documentos del corpus. A pesar de que la publicación ofrece la posibilidad (e incluso recomienda) el uso de una sección independiente en la que definir las formas abreviadas utilizadas en la redacción junto a sus formas expandidas, lo cierto es que rara vez se utiliza dicha sección, y en general, la mayoría de los acrónimos y abreviaturas se definen *ad hoc* en el propio documento, a medida que aparecen en él. Además, a menudo no

se corresponden con formas estándar, por lo que no suelen existir en UMLS. Por ejemplo, en uno de los documentos del corpus en el que se presenta una herramienta para el análisis de repeticiones de secuencias simples de ADN, sólo la primera ocurrencia de *simple sequence repeat* se presenta como tal, en el resto del documento se utiliza su acrónimo "SSR". Lo mismo ocurre en un documento que investiga la producción de survivina durante el desarrollo embrionario de las glándulas salivares sub-mandibulares, donde la expresión "SMG" se utiliza siempre en lugar de *embryonic submandibular*. La inmediata consecuencia de esto es que, cuando se traduce el documento a conceptos de UMLS, MetaMap no encuentra los conceptos asociados a dichos acrónimos. Se ha observado además, que dichos términos o grupos de términos representados de forma abreviada frecuentemente se corresponden con conceptos relevantes del documento. Por tanto, cuando aparecen en una oración, al no traducirse a ningún concepto de UMLS, la oración disminuye su probabilidad de ser seleccionada para el resumen.

Para soslayar esta limitación, se ha utilizado BioText[5] (Schwartz y Hearst, 2003), una herramienta para reconocer y resolver las apariciones de acrónimos en textos biomédicos. Está desarrollada en java, resulta muy sencilla de instalar y utilizar, y presenta una precisión en torno al 95 % y una cobertura del 82 %. La integración del software en el sistema de generación de resúmenes se realiza como primer paso dentro del proceso de elaboración del resumen. Antes incluso de la fase de pre-procesamiento del documento, se invoca la clase *ExtractAbbrev* con el documento que se desea resumir como argumento. Como resultado, se obtiene un fichero en el que se listan las abreviaturas o acrónimos encontrados en el documento, junto con la definición o forma expandida sugerida. A continuación, simplemente se analiza el fichero resultado y se sustituye cada acrónimo del documento por su forma expandida. La Tabla 8.15 muestra el resultado de aplicar el algoritmo de generación de resúmenes sobre el corpus de evaluación, una vez resueltos los acrónimos y las abreviaturas. Si comparamos estos resultados con los presentados en la Tabla 8.13, comprobamos que, efectivamente, la presencia de acrónimos y abreviaturas no reconocidos por MetaMap influye negativamente en los resultados de la generación de resúmenes, si bien la mejora obtenida tras su expansión ha resultado ser menor de lo esperado.

[5]BioText. http://biotext.berkeley.edu/software.html. Consultada el 1 de noviembre de 2010

Heurística	R-1	R-2	R-4	R-L	R-W	R-S4
Heurística 3	**0.7874**	**0.3560**	**0.1339**	**0.6837**	**0.2017**	**0.3300**
Heurística 2	0.7800	0.3440	0.1250	0.6804	0.1996	0.3228
Heurística 1	0.7754	0.3476	0.1337	0.6738	0.1953	0.3232

Tabla 8.15: Resultados del sistema una vez resueltos los acrónimos y las abreviaturas

8.3. Evaluación del Caso de Estudio de Generación de Resúmenes Mono-documento de Noticias Periodísticas

Evaluamos ahora el rendimiento del método de generación de resúmenes con la configuración presentada en el Capítulo 6 para el tratamiento de noticias periodísticas. De nuevo, esta evaluación comprende tanto la determinación de los valores de los parámetros del algoritmo como la comparación del método propuesto con respecto a otros sistemas de resúmenes comerciales y de investigación.

8.3.1. Parametrización

Con el propósito de responder a las cuestiones planteadas en la Sección 8.1.3 y determinar los valores óptimos de los parámetros que intervienen en el algoritmo de generación de resúmenes, se ha llevado a cabo una evaluación preliminar sobre un conjunto de 10 documentos. A continuación se describen los experimentos realizados y los resultados obtenidos.

8.3.1.1. Definición del Número Óptimo de *Hub Vertices* y del Umbral de Similitud

El objetivo del primer grupo de experimentos es establecer el porcentaje de vértices utilizados como *hub vertices* en el algoritmo de agrupamiento (ver Sección 4.5), junto con el mejor umbral considerado a la hora de añadir las relaciones de similitud al grafo del documento (ver Sección 4.4). Se utilizan, pues, en esta experimentación, ambas relaciones semánticas (i.e. hiperonimia y similitud) a la hora de construir el grafo. Se utiliza *jcn* (Sección 3.5) como métrica para calcular la similitud semántica entre conceptos de WordNet. Nótese que es necesario evaluar ambos parámetros simultáneamente,

pues del umbral de similitud utilizado depende la conectividad del grafo del documento, y esta, a su vez, influye sobre el número óptimo de *hub vertices*. En la realización de estos experimentos, se ha utilizado el coeficiente de Jaccard para el cálculo de los pesos de las aristas del grafo (ver Sección 4.5) y ninguno de los criterios de selección de oraciones estudiados en la Sección 4.7 (i.e. criterio posicional y de similitud con el título).

Las Tablas 8.16, 8.17 y 8.18 muestran las puntuaciones medias de ROUGE obtenidas por los resúmenes generados con distintas combinaciones de valores de los parámetros estudiados, para cada una de las tres heurísticas de selección de oraciones.

Heurística 1							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
0.01	2 %	**0.5562**	**0.2565**	**0.1193**	**0.5038**	**0.1898**	**0.2261**
	5 %	0.5307	0.2288	0.1126	0.4825	0.1816	0.2054
	10 %	0.5307	0.2288	0.1126	0.4825	0.1816	0.2054
	20 %	0.4811	0.1914	0.0743	0.4379	0.1661	0.1691
0.05	2 %	*0.5438*	*0.2513*	*0.1184*	*0.4914*	*0.1868*	*0.2232*
	5 %	0.5305	0.2285	0.1166	0.4841	0.1805	0.2064
	10 %	0.4915	0.1921	0.0901	0.4404	0.1633	0.1715
	20 %	0.4426	0.1628	0.0724	0.4095	0.1543	0.1499
0.1	2 %	0.5205	0.2244	0.0999	0.4785	0.1775	0.1983
	5 %	*0.5347*	*0.2353*	*0.1081*	*0.4858*	*0.1852*	*0.2081*
	10 %	0.4592	0.1682	0.0693	0.4058	0.1517	0.1547
	20 %	0.5104	0.1950	0.0866	0.4568	0.1718	0.1783
0.2	2 %	*0.5419*	*0.2384*	*0.1096*	*0.4924*	*0.1872*	*0.2116*
	5 %	0.4723	0.1638	0.0663	0.4218	0.1581	0.1571
	10 %	0.4723	0.1638	0.0663	0.4218	0.1581	0.1571
	20 %	0.4723	0.1638	0.0663	0.4218	0.1581	0.1571

Tabla 8.16: Evaluación del umbral de similitud y del porcentaje de *hub vertices* para la heurística 1

En relación a la primera de las heurísticas, en la Tabla 8.16 se puede observar cómo el mejor resultado se consigue utilizando un porcentaje del 2 % de los vértices del grafo del documento como nodos *hub*, junto con un umbral de similitud de 0.01 en la construcción del grafo del documento. Se observa, además, que casi con completa independencia de cuál sea el umbral de similitud utilizado, al aumentar el número de *hub vertices* el resultado empeora. La razón es que, al tratarse de documentos que únicamente tratan un único tema o suceso, éste queda perfectamente definido por unos pocos conceptos, por lo que el aumentar el número de conceptos utilizados como

centroides de los clusters no hace sino otorgar importancia a conceptos muy alejados o en absoluto relacionados con el tema principal de la noticia.

La Tabla 8.17 muestra los resultados correspondientes a la segunda heurística. Se puede comprobar que el mejor resultado se consigue cuando se utilizan como *hub vertices* el 20 % de los nodos del grafo del documento, junto con un umbral de similitud de 0.05. Así pues, ahora los resultados parecen favorecer un número elevado de vértices *hub*. La razón, de nuevo, se encuentra en la naturaleza de esta heurística, que "obliga" a seleccionar oraciones de todos los temas tratados en el documento, con independencia de que representen o no información central en el mismo.

Heurística 2							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
0.01	2 %	0.5362	0.2365	0.1093	0.4838	0.1798	0.2061
	5 %	0.5231	0.2206	0.1053	0.4740	0.1788	0.1982
	10 %	0.5231	0.2206	0.1053	0.4740	0.1788	0.1982
	20 %	*0.5455*	*0.2415*	*0.1148*	*0.4899*	*0.1841*	*0.2171*
0.05	2 %	0.5403	0.2422	0.1163	0.4909	0.1885	0.2132
	5 %	0.5279	0.2225	0.1014	0.4853	0.1812	0.1968
	10 %	0.5391	0.2060	0.0813	0.4894	0.1784	0.1848
	20 %	*0.5463*	**0.2455**	**0.1180**	*0.4919*	*0.1895*	**0.2184**
0.1	2 %	0.5485	0.2141	0.0972	0.5056	*0.1881*	0.1956
	5 %	0.5125	0.2097	0.0968	0.4671	0.1745	0.1865
	10 %	0.5421	0.2378	*0.1098*	0.4934	0.1874	0.2084
	20 %	*0.5488*	*0.2405*	0.1096	*0.4983*	0.1865	*0.2131*
0.2	2 %	0.5419	0.2336	0.1006	0.4868	0.1822	0.2060
	5 %	0.5371	0.2179	0.0904	0.4771	0.1760	0.1872
	10 %	**0.5524**	*0.2437*	*0.1202*	0.5058	**0.1950**	*0.2161*
	20 %	0.5483	0.2204	0.0887	**0.5254**	0.1911	0.1948

Tabla 8.17: Evaluación del umbral de similitud y del porcentaje de *hub vertices* para la heurística 2

Por último, la Tabla 8.18 muestra los resultados de la tercera heurística. En esta ocasión, el mejor resultado se obtiene utilizando un umbral de 0.01 y un número de vértices *hub* igual al 5 % de los vértices del grafo de documento. Se trata, pues, de un valor intermedio entre los utilizados por las heurísticas anteriores, pues nos encontramos ante una heurística que, si bien premia la selección de oraciones relacionadas con el tema principal del documento, también permite incluir otros contenidos secundarios.

Así pues, al igual que ocurriera en el dominio biomédico, el porcentaje de vértices utilizados como *hubs* depende de la heurística. En general, el

Heurística 3							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
0.01	2 %	0.5405	0.2360	0.1126	0.4896	0.1850	0.2079
	5 %	**0.5608**	**0.2565**	**0.1193**	**0.5094**	**0.1910**	**0.2260**
	10 %	0.5405	0.2360	0.1126	0.4896	0.1850	0.2079
	20 %	0.5464	0.2509	0.1189	0.5023	0.1906	0.2200
0.05	2 %	0.5400	0.2324	0.1098	0.4926	0.1842	0.2058
	5 %	*0.5428*	*0.2475*	*0.1184*	*0.4905*	*0.1861*	*0.2219*
	10 %	0.5346	0.2320	0.1082	0.4884	0.1826	0.2024
	20 %	0.5285	0.2387	0.1123	0.4841	0.1822	0.2067
0.1	2 %	0.5438	0.2414	0.1075	0.4974	0.1875	0.2133
	5 %	*0.5544*	*0.2472*	*0.1096*	*0.5039*	*0.1885*	*0.2164*
	10 %	0.5216	0.2384	0.1126	0.4765	0.1808	0.2076
	20 %	0.5271	0.2305	0.1059	0.4847	0.1822	0.2007
0.2	2 %	0.5419	0.2384	0.1096	0.4924	0.1872	0.2116
	5 %	*0.5500*	*0.2539*	*0.1202*	*0.4976*	*0.1891*	*0.2251*
	10 %	*0.5500*	*0.2539*	*0.1202*	*0.4976*	*0.1891*	*0.2251*
	20 %	*0.5500*	*0.2539*	*0.1202*	*0.4976*	*0.1891*	*0.2251*

Tabla 8.18: Evaluación del umbral de similitud y del porcentaje de *hub vertices* para la heurística 3

valor óptimo varía entre el 2 % y el 5 %, a excepción de la heurística 2, que requiere un número mucho mayor de *hub vertices* para alcanzar sus mejores resultados. Lo mismo ocurre con el umbral de similitud, que depende del número de vértices *hub* y que, por tanto, también varía en función de la heurística considerada.

8.3.1.2. Definición del Mejor Conjunto de Relaciones Semánticas

El objetivo del segundo grupo de experimentos es determinar la mejor combinación de relaciones semánticas utilizadas en la construcción del grafo del documento (ver Sección 4.4). Recordemos que este parámetro debe ser evaluado en combinación con el número óptimo de *hub vertices*, puesto que las relaciones consideradas influyen sobre la conectividad del grafo del documento y, por lo tanto, también sobre el número óptimo de *hub vertices*. Puesto que el efecto del porcentaje de *hub vertices* sobre el resultado cuando se utiliza la relación de similitud semántica ya ha sido analizado en la sección anterior, aquí nos limitamos a repetir esos mismos experimentos para el caso en el que sólo se considera la relación de hiperonimia, y a recordar el mejor resultado de cada heurística para el conjunto completo de relaciones.

Las Tablas 8.19, 8.20 y 8.21 muestran las puntuaciones obtenidas por

los resúmenes generados. Podemos comprobar en estas tablas que, con independencia de la heurística, el mejor conjunto de relaciones semánticas está formado por la relación de hiperonimia y la relación de similitud semántica. La otra alternativa, utilizar relación de hiperonimia en exclusividad, resulta insuficiente, ya que produce un grafo del documento excesivamente inconexo. El resultado es que un elevado número de conceptos que semánticamente guardan una estrecha relación, no se encuentran directamente relacionados en este grafo, sino a través del largo camino que construyen sus jerarquías.

Heurística 1							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	*0.5387*	0.2209	0.1157	0.4801	*0.1840*	*0.2102*
	5 %	0.4748	0.1883	0.0913	0.4322	0.1635	0.1746
	10 %	0.5235	*0.2410*	*0.1193*	*0.4810*	0.1836	0.2084
	20 %	0.5176	0.2045	0.0889	0.4733	0.1760	0.1804
Hiperonimia + Sim. Sem.		**0.5562**	**0.2565**	**0.1193**	**0.5038**	**0.1898**	**0.2261**

Tabla 8.19: Evaluación del mejor conjunto de relaciones para la heurística 1

Heurística 2							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	0.5186	0.2237	*0.1157*	0.4741	0.1793	0.1984
	5 %	0.5034	0.2143	0.0977	0.4608	0.1746	0.1902
	10 %	0.5590	0.2291	0.0936	0.5074	0.1884	0.1970
	20 %	**0.6063**	*0.2405*	0.0985	**0.5545**	**0.1949**	*0.2095*
Hiperonimia + Sim. Sem.		0.5463	**0.2455**	**0.1180**	0.4919	0.1895	**0.2184**

Tabla 8.20: Evaluación del mejor conjunto de relaciones para la heurística 2

Heurística 3							
Parámetros		**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Hiperonimia	2 %	*0.5357*	*0.2479*	*0.1157*	*0.4971*	0.1824	*0.2181*
	5 %	0.5152	0.2181	0.0967	0.4642	0.1769	0.1937
	10 %	0.5326	0.2312	0.1065	0.4857	0.1832	0.2026
	20 %	0.5411	0.2395	0.1133	0.4950	*0.1860*	0.2082
Hiperonimia + Sim. Sem.		**0.5608**	**0.2565**	**0.1193**	**0.5094**	**0.1910**	**0.2260**

Tabla 8.21: Evaluación del mejor conjunto de relaciones para la heurística 3

8.3.1.3. Definición de la Mejor Combinación de Criterios de Selección de Oraciones

El tercer grupo de experimentos pretende comprobar si el uso de los criterios estadísticos de posición de la oración y similitud con el título permiten mejo-

rar el resultado conseguido por el generador de resúmenes basado en grafos semánticos, utilizando para ello la Ecuación 4.11 definida en la Sección 4.7 para seleccionar las oraciones del resumen. Para realizar estos experimentos, el resto de parámetros del algoritmo que no son objeto de evaluación (i.e. el número de *hub vertices*, el conjunto de relaciones semánticas y el umbral de similitud) se han establecido a los valores que mejores resultados han ofrecido para cada una de las heurísticas según la evaluación realizada hasta el momento. Se ha utilizado el coeficiente de similitud de Jaccard para etiquetar las aristas del grafo del documento.

Las Tablas 8.22, 8.23 y 8.24 muestran las puntuaciones de ROUGE obtenidas por los resúmenes generados con los distintos sistemas que resultan de combinar nuestro método con los dos criterios mencionados.

Heurística 1									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	0.5562	0.2565	0.1193	0.5038	0.1898	0.2261
Grafos Sem. + Posición	0.9	0.1	0.0	**0.5593**	**0.2585**	**0.1217**	**0.5108**	**0.1921**	**0.2332**
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.5359	0.2358	0.1087	0.4836	0.1816	0.2089
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	0.5526	0.2548	0.1193	0.5031	0.1906	0.2241

Tabla 8.22: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 1

Heurística 2									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	0.5463	0.2455	0.1180	0.4919	0.1895	0.2184
Grafos Sem. + Posición	0.9	0.1	0.0	**0.5592**	**0.2596**	**0.1193**	**0.5088**	**0.1921**	**0.2286**
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.5248	0.2320	0.1087	0.4743	0.1779	0.2054
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	0.5383	0.2481	0.1193	0.4907	0.1850	0.2214

Tabla 8.23: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 2

A la vista de los resultados, se puede comprobar que todas las heurísticas

Heurística 3									
Criterios	λ	θ	χ	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Grafos Sem.	1.0	0.0	0.0	0.5608	0.2565	0.1193	0.5094	0.1910	0.2260
Grafos Sem. + Posición	0.9	0.1	0.1	**0.5612**	**0.2596**	**0.1206**	**0.5098**	**0.1921**	**0.2286**
Grafos Sem. + Sim. Tit.	0.9	0.0	0.1	0.5359	0.2358	0.1087	0.4836	0.1816	0.2089
Grafos Sem. + Posición +Sim. Tit.	0.8	0.1	0.1	0.5526	0.2548	0.1193	0.5031	0.1906	0.2271

Tabla 8.24: Resultados de la evaluación de las distintas combinaciones de criterios de selección de oraciones para la heurística 3

se comportan mejor cuando se combinan con el criterio posicional, mientras que la similitud con el título produce unos resultados considerablemente peores que el resto de las combinaciones de criterios estudiadas.

La Tabla 8.25 recopila los mejores resultados obtenidos por cada una de las heurísticas, así como la combinación de criterios de selección de oraciones que dan como fruto estos resultados.

Heurística	**Criterios**	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 1	Grafos Sem. + Posición	0.5593	0.2585	0.1217	**0.5108**	0.1921	0.2332
Heurística 2	Grafos Sem. + Posición	0.5592	0.2596	0.1193	0.5088	0.1921	0.2286
Heurística 3	Grafos Sem. + Posición	**0.5612**	**0.2596**	**0.1206**	0.5098	**0.1921**	**0.2286**

Tabla 8.25: Combinación óptima de criterios de selección de oraciones para cada heurística

8.3.1.4. Definición del Mejor Coeficiente de Similitud

A continuación evaluamos cuál de los dos índices de similitud implementados en el sistema de generación de resúmenes para el cálculo de los pesos de las aristas del grafo del documento (i.e. los coeficientes de similitud de Jaccard y de Dice-Sorensen) resulta más apropiado. Para ello, repetimos los experimentos mostrados en la Tabla 8.25, esta vez utilizando el coeficiente de similitud de Dice-Sorensen. La Tabla 8.26 muestra los resultados obtenidos en estos nuevos experimentos. De nuevo, y al igual que ocurriera en el caso de estudio biomédico, los resultados para todas las heurísticas son

inferiores a los obtenidos utilizando el coeficiente de Jaccard, aunque las diferencias no son significativas, dado que ambos coeficientes producen pesos muy similares.

Heurística	**Criterios**	**R-1**	**R-2**	**R-4**	**R-L**	**R-W**	**R-S4**
Heurística 1	Grafos Sem. + Posición	0.5523	0.2534	0.1186	0.5064	0.1896	0.2277
Heurística 2	Grafos Sem. + Posición	0.5517	0.2522	0.1176	0.5042	0.1888	0.2265
Heurística 3	Grafos Sem. + Posición	**0.5586**	**0.2543**	**0.1192**	**0.5072**	**0.1904**	**0.2265**

Tabla 8.26: Combinación óptima de criterios de selección de oraciones para cada heurística, utilizando el coeficiente de similitud de Dice-Sorensen para el cálculo de los pesos de las aristas del grafo del documento

8.3.1.5. Recopilación: Parametrización Óptima del Algoritmo

Una vez estudiados todos los parámetros que intervienen en el algoritmo de generación de resúmenes, resulta apropiado recordar el resultado de dicha parametrización. La Tabla 8.27 recopila los valores óptimos de cada uno de los parámetros para cada una de las tres heurísticas de selección de oraciones.

Heurística	**Conjunto de Relaciones**	**Umbral de Similitud**	**Nº de *Hub Vertices***	**Criterios de Selección**	**Coeficiente de Similitud**
Heurística 1	Hiperonimia + Sim.Semántica	0.01	2 %	Grafos Sem. + Posición	Jaccard
Heurística 2	Hiperonimia + Sim.Semántica	0.05	20 %	Grafos Sem. + Posición	Jaccard
Heurística 3	Hiperonimia + Sim.Semántica	0.01	5 %	Grafos Sem. + Posición	Jaccard

Tabla 8.27: Recopilación: Mejor parametrización por heurística

Por lo tanto, a la vista de la Tabla 8.27, y tal y como sucediera en el caso de estudio anterior, se deduce que la parametrización del algoritmo varía en función de la heurística. No obstante, en este caso nos encontramos con tres parámetros cuyos valores óptimos no dependen de la heurística: el conjunto de relaciones semánticas utilizado para construir el grafo del documento, la combinación adecuada de criterios de selección de oraciones y el coeficiente para etiquetar las aristas del grafo del documento. Por un lado, el mejor

conjunto de relaciones semánticas es siempre aquel que incluye a la relación de similitud. Esto es lógico puesto que, como ya se ha comentado, la relación de hiperonimia por sí sola produce un grafo del documento muy inconexo, y del que por tanto no puede suponerse que cumpla las propiedades de una red libre de escala. Por otro lado, y en relación a la mejor combinación de criterios de selección, el uso de información sobre la posición de la oración en el documento mejora la calidad de los resúmenes generados por el algoritmo basado en grafos, con independencia de la heurística implementada. Esto queda explicado por la naturaleza de los documentos que nos ocupan, noticias periodísticas, que, como ya sabemos, generalmente se redactan atendiendo a una estructura "piramidal" en la que la información relevante ocupa las primeras líneas del cuerpo de la noticia.

8.3.2. Estudio del Efecto de la Ambigüedad Léxica

Al igual que para el dominio anterior, nos interesa conocer en qué medida afecta la ambigüedad léxica presente en el documento a la calidad de los resúmenes generados. Para ello, se han realizado diversos experimentos utilizando distintos algoritmos de desambiguación, sobre el conjunto completo de 567 documentos de la conferencia DUC 2002. En concreto, se han evaluado los resúmenes generados utilizando las siguientes estrategias de desambiguación:

1. Seleccionar aleatoriamente un significado de entre todos los posibles en WordNet (*Aleatorio*), lo que equivale a no realizar desambiguación.
2. Seleccionar el primero de los posibles significados en WordNet *(1er sense)*, lo que equivale a seleccionar el significado más frecuente.
3. Utilizar el algoritmo de desambiguación de Lesk (Lesk, 1986) *(Lesk)*.

De nuevo, podemos observar en la Tabla 8.28 que el uso de desambiguación mejora las puntuaciones obtenidas por todas las heurísticas. Según la prueba de los signos de Wilcoxon, $p < 0{,}05$, ambas estrategias de desambiguación (*Lesk* y *1er Sense*) producen resultados significativamente mejores que la estrategia aleatoria para todas las métricas ROUGE. Sin embargo, y aunque la desambiguación utilizando el algoritmo *Lesk* es la que mejor se comporta, las diferencias con respecto a utilizar la estrategia *1er Sense* es menor de lo esperado. La razón parece ser, en primer lugar, que utilizar el significado más frecuente es un buen criterio de desambiguación; pero

Sistema	R-1	R-2	R-4	R-L	R-W	R-S4
Heurística 1 - *Aleatorio*	0.4214	0.1932	0.0902	0.4027	0.1503	0.1691
Heurística 2 - *Aleatorio*	0.4253	0.1972	0.0934	0.4085	0.1555	0.1713
Heurística 3 - *Aleatorio*	0.4322	0.2001	0.0976	0.4109	0.1576	0.1780
Heurística 1 - *1er Sense*	0.4584	0.2057	0.0996	0.4217	0.1626	0.1794
Heurística 2 - *1er Sense*	0.4594	0.2074	0.1027	0.4224	0.1631	0.1810
Heurística 3 - *1er Sense*	0.4619	0.2104	**0.1043**	0.4252	0.1643	0.1838
Heurística 1 - *Lesk*	**0.4641**	**0.2191**	**0.1015**	**0.4268**	**0.1647**	**0.1919**
Heurística 2 - *Lesk*	**0.4651**	**0.2193**	**0.1374**	**0.4276**	**0.1650**	**0.1927**
Heurística 3 - *Lesk*	**0.4648**	**0.2196**	0.1023	**0.4277**	**0.1652**	**0.1928**

Tabla 8.28: Evaluación del sistema de generación de resúmenes para distintas estrategias de desambiguación léxica

además, que el algoritmo *Lesk* introduce cierto ruido, ya que favorece la elección del primer significado posible. De hecho, se ha comprobado que la desambiguación realizada por las dos estrategias coincide para el 61 % de los conceptos del documento.

8.3.3. Comparación con otros Sistemas

El objetivo de esta sección es comparar los resultados obtenidos por el método propuesto en la generación de resúmenes de noticias periodísticas con aquellos obtenidos por otros sistemas de resúmenes automáticos. Para ello, se han generado resúmenes de los 567 documentos que componen el corpus de evaluación de la conferencia DUC 2002 (ver Sección 8.1.2), utilizando los sistemas LexRank, SUMMA y AutoSummarize. De nuevo, y al igual que en el caso de estudio anterior, se han generado resúmenes posicionales (Lead) y resúmenes aleatorios (Random). Además, se muestran los resultados obtenidos y publicados por otros sistemas evaluados sobre la misma colección de documentos y utilizando las mismas métricas de evaluación: *LeLSA+AR*, *Freq+TextEnt*, y los cinco sistemas participantes en DUC 2002 que mejores resultados obtienen en términos de la métrica ROUGE (DUC 19, DUC 21, DUC 27, DUC 28 y DUC 29). Todos estos sistemas han sido descritos en la Sección 8.1.4.

Para calcular las métricas ROUGE, todos los resúmenes a evaluar se truncan de modo que su longitud sea exactamente 100 palabras, tal y como se hace en las competiciones de las conferencias DUC y TAC. La Tabla 8.13 muestra el resultado obtenido por todos estos sistemas para las diferentes

métricas de evaluación, así como los obtenidos por las tres versiones de nuestro sistema utilizando la parametrización determinada en la Sección 8.3.1.5 y el algoritmo de desambiguación Lesk. En esta tabla, los resultados se muestran ordenados según el valor de ROUGE-2, y el mejor resultado obtenido para cada métrica se muestra en negrita.

Sistema	**R-1**	**R-2**	**R-L**	**R-S4**
Heurística 3	0.4648	**0.2196**	**0.4277**	**0.1928**
Heurística 2	**0.4651**	0.2193	0.4276	0.1927
Heurística 1	0.4641	0.2191	0.4268	0.1919
LexRank	0.4558	0.2115	0.4173	0.1846
Freq+TextEnt	0.4518	0.1942	0.4104	-
LeLSA+AR	0.4228	0.2074	0.3928	0.1661
DUC 28	0.4278	0.2177	0.3865	0.1732
SUMMA	0.4217	0.1952	0.3876	0.1516
AutoSummarize	0.4216	0.1887	0.3671	0.1429
Lead	0.4113	0.2108	0.3754	0.1660
DUC 19	0.4082	0.2088	0.3735	0.1638
DUC 27	0.4052	0.2022	0.3691	0.1600
DUC 21	0.4149	0.2104	0.3754	0.1655
DUC 29	0.3993	0.2006	0.3617	0.1576
Random	0.2996	0.1110	0.2795	0.0900

Tabla 8.29: Comparación de los resultados con los obtenidos por otros sistemas

Para comprobar si las diferencias entre estos sistemas son estadísticamente significativas, se ha realizado la prueba de los signos de Wilcoxon con un nivel de confianza del 95 %. El resultado indica que las tres heurísticas del método propuesto se comportan significativamente mejor que LexRank, todos los sistemas DUC y ambas líneas base, en al menos dos de las cuatro métricas ROUGE analizadas. Por el contrario, no existen diferencias significativas entre las tres heurísticas. Con relación a Freq+TextEnt y LeLSA+AR, al no disponerse de los resultados desagregados sino únicamente de las puntuaciones promedio, no es posible aplicar dicho test. No obstante, las tres versiones de nuestro generador de resúmenes superan a ambos sistemas en todas las métricas ROUGE (para el sistema Freq+TextEnt, no se dispone del valor de ROUGE-S4).

8.3.4. Discusión de los Resultados

Los resultados de la evaluación muestran el buen comportamiento del método propuesto en comparación con otros sistemas que implementan un amplio

abanico de técnicas de generación de resúmenes. Sin embargo, llama especialmente la atención el hecho de que las tres heurísticas, incluso habiendo sido diseñadas para perseguir distintos objetivos, presentan resultados muy similares. Para entender el por qué de esta situación, se han examinado los resultados intermedios del algoritmo, en concreto, los diferentes clusters generados y la asignación de oraciones a estos clusters según cada una de dichas heurísticas. Se ha comprobado que, con frecuencia, y para algunos documentos, el algoritmo produce un cluster de gran tamaño junto con un numero variable de clusters de tamaño mucho menor. Como consecuencia, aunque la heurística 2 haya sido diseñada para seleccionar oraciones de todos los clusters, al determinar el número de oraciones correspondientes a cada cluster en función del tamaño de éste, el resultado es que la mayoría de las oraciones se seleccionan del cluster de mayor tamaño. En estos casos, los resúmenes producidos por las tres heurísticas son muy similares. Sin embargo, los mejores resultados se consiguen con la heurística 3. De nuevo, se ha comprobado que esta heurística selecciona la mayoría de las oraciones del cluster de mayor tamaño, pero además selecciona algunas oraciones de otros clusters cuando las puntuaciones asignadas a estos son altas. De este modo, además de incluir información relacionada con el tema central de la noticia, esta heurística también incluye otra información secundaria que podría ser de interés para el lector. Por el contrario, la heurística 1 no consigue capturar este tipo de información, mientras que la heurística 2 incluye demasiada información secundaria en detrimento de otra información principal.

Por otro lado, y a diferencia de lo que sucediera en el dominio anterior, en este dominio los acrónimos y las abreviaturas no suponen un grave problema, ya que, en general, se trata de formas estándar, la mayoría de ellas incluidas en WordNet, como por ejemplo, “Mr”, “Mrs” o “Al” (Alabama). No obstante, recordemos que el generador de resúmenes utiliza las listas de abreviaturas y acrónimos del módulo de nomenclatura *ANNIE Gazetteer* de GATE, de tal forma que, cada vez que se encuentra en el texto una forma abreviada cuya expansión aparece en dichas listas, se sustituye la una por la otra.

De nuevo, el análisis de las puntuaciones ROUGE obtenidas por cada uno de los documentos del corpus muestra que existen importantes divergencias entre las mismas. Esto se puede apreciar en la Tabla 8.30, donde se muestra la desviación típica de los valores de las diferentes métricas ROUGE en la

colección de evaluación, para los resúmenes obtenidos con la heurística 3. Se puede comprobar que tales desviaciones son muy similares a las que se obtuvieron para el caso de estudio anterior (ver Tabla 8.30). Sin embargo, en esta ocasión la única causa evidente parece ser la diferencia de tamaño entre los documentos del corpus.

Métrica	**Desviación típica**
ROUGE-1	0.1001
ROUGE-2	0.1179
ROUGE-4	0.0974
ROUGE-L	0.0998
ROUGE-W-1.2	0.0432
ROUGE-S4	0.1052

Tabla 8.30: Desviación típica en los resultados de las distintas métricas ROUGE para los resúmenes generados por la heurística 3

8.4. Evaluación del Caso de Estudio de Generación de Resúmenes Multi-documento de Páginas Web de Información Turística

A continuación evaluamos el rendimiento del método propuesto en la tarea de generar resúmenes a partir de múltiples páginas web de información turística sobre un mismo objeto o monumento (Capítulo 7).

A diferencia de los casos de estudio anteriores, en el que ahora nos ocupa no se ha realizado un proceso de determinación de los valores óptimos de los distintos parámetros del algoritmo. Por el contrario, se han utilizado los valores determinados para el caso de estudio de noticias periodísticas (ver Tabla 8.27), por emplear ambos la misma base de conocimiento y el mismo algoritmo de desambiguación. El objetivo es comprobar si el método puede ser aplicado satisfactoriamente a nuevos dominios y tipos de documentos sin necesidad de realizar esta parametrización inicial.

De nuevo, la evaluación se realiza mediante el cálculo automático de distintas métricas ROUGE, y su comparación con las obtenidas por otros sistemas de resúmenes. Además, se completa esta evaluación con una valoración manual, realizada por tres personas distintas, de las distintas características de legibilidad presentadas en la Sección 2.3.3.7.

8.4.1. Comparación con otros Sistemas

Para comparar los resultados obtenidos por nuestro algoritmo con los presentados por otros sistemas, se han generado resúmenes de 200 ± 10 palabras de los 308 grupos de documentos que componen el corpus de evaluación (Aker y Gaizauskas, 2009), utilizando las tres heurísticas de selección de oraciones, SUMMA, MEAD, COMPENDIUM y Language Models. Todos estos sistemas han sido explicados en la Sección 8.1.4.

La Tabla 8.31 muestra las puntuaciones obtenidas para las métricas ROUGE-2 y ROUGE-S4. Nótese que la razón de mostrar únicamente las puntuaciones de dos de las métricas ROUGE es que no se dispone de información sobre las puntuaciones obtenidas por algunos de los sistemas en las otras métricas. En esta tabla, los resultados se muestran ordenados según el valor de ROUGE-2, y el mejor resultado obtenido para cada métrica se muestra en negrita.

Sistema	**R-2**	**R-S4**
Heurística 3	**0.090**	**0.143**
Heurística 1	0.089	0.139
MEAD	0.089	0.138
COMPENDIUM	0.086	0.134
Language Models	0.071	0.119
Heurística 2	0.069	0.117
SUMMA	0.064	0.109

Tabla 8.31: Comparación de los resultados con los obtenidos por otros sistemas

La Tabla 8.31 muestra que las heurísticas 1 y 3 se comportan mejor que el resto de sistemas analizados para todas las métricas ROUGE. De acuerdo con el test de los signos de Wilcoxon ($p < 0.01$), ambas heurísticas y MEAD producen resultados significativamente mejores que SUMMA, COMPENDIUM, Language Models y la segunda de las heurísticas. Sin embargo, no existen diferencias significativas entre ambas heurísticas y con respecto a MEAD. Por el contrario, la heurística 2 se comporta considerablemente peor que en el resto de casos de estudio. La razón parece ser que los documentos a resumir contienen una amplia variedad de información relacionada con temas muy diversos, por lo que el resultado del algoritmo de clustering es un relativamente elevado número de grupos. Por ello, la heurística 2, al verse obligada a seleccionar las oraciones de todos y cada uno de los grupos, incluye demasiada información secundaria o poco relacionada con el tema

principal del conjunto de documentos a resumir.

8.4.2. Evaluación de la Legibilidad de los Resúmenes

Además de la evaluación automática del contenido informativo, los resúmenes generados en este caso de estudio han sido sometidos a una evaluación manual de su legibilidad. Para ello, se han seleccionado aleatoriamente 50 resúmenes de entre los generados por la heurística 3, y solicitado a tres personas que puntúen los resúmenes de acuerdo a los cinco criterios de legibilidad utilizados en las conferencias DUC y TAC (Sección 2.3.3.7), asignando a cada criterio un valor de 1 a 5 ($1 \approx$ Muy baja calidad, $5 \approx$ Muy alta calidad).

La Tabla 8.32 muestra los resultados obtenidos por nuestro método en esta evaluación, junto con los alcanzados por los sistemas que, para cada uno de los criterios considerados, mejor puntuación obtuvieron en la edición 2006 de la conferencia DUC (Dang, 2006).

Criterio	**Puntuación**	**DUC 2006**
Calidad Gramatical	4.11	4.62
Redundancia	3.8	4.0
Claridad referencial	3.72	4.66
Foco	4.1	4.28
Estructura y coherencia	3.15	3.28

Tabla 8.32: Evaluación de la legibilidad de los resúmenes generados. Los valores que se muestran son la media aritmética de las puntuaciones asignadas por tres evaluadores a cada uno de los criterios de legibilidad

A la vista de la tabla anterior, y aunque los resultados obtenidos son peores que los alcanzados por los mejores sistemas de la conferencia DUC 2006, conviene aclarar que en la tarea competitiva de esta conferencia los sistemas disponían de una consulta o pregunta acerca del contenido que debía presentar el resumen; es decir, se trataba de una tarea de generación de resúmenes adaptados a una consulta. Tal y como demuestran estudios recientes, el disponer de este tipo de información influye muy positivamente en la calidad de los resúmenes generados.

Independientemente de lo anterior, los peores resultados se obtienen para los criterios "Claridad referencial" y "Estructura y coherencia". Sin duda alguna, ambos criterios pueden y deberán ser mejorados mediante el uso de técnicas de resolución de anáforas y co-referencias (Steinberger et al., 2007),

y de simplificación y condensación de oraciones (Barzilay y McKeown, 2005; Filippova y Strube, 2008).

8.4.3. Discusión de los Resultados

Los resultados de la evaluación demuestran, en primer lugar, que el algoritmo propuesto puede ser fácilmente adaptado para abordar la tarea de generar resúmenes a partir de múltiples documentos, sin más que disponer de un método para detectar la redundancia derivada del uso de diversas fuentes con información repetida. En segundo lugar, demuestran que, sin necesidad de realizar un nuevo proceso de parametrización, éste puede aplicarse con éxito a nuevos dominios y tipos de documentos.

Al igual que en los casos de estudio anteriores, se han detectado importantes divergencias en las puntuaciones de ROUGE obtenidas por los distintos resúmenes. En particular, el método presenta pobres resultados en la generación de resúmenes de documentos que describen lugares como el *Sacre Coeur*, *Santorini* o *Ipanema*. El motivo parece encontrarse en el hecho de que estos lugares no se encuentran representados en la base de conocimiento (i.e. WordNet), y que por lo tanto, y al tratarse precisamente de los conceptos más importantes de los documentos a resumir, el grafo generado prescinde de información esencial para la identificación de los temas principales de estos documentos. Creemos, además, que esta es la causa de que las diferencias con respecto a otros sistemas en este caso de estudio no sean tan pronunciadas como en los casos de estudio anteriores. Para solucionar este problema, sería recomendable estudiar la posibilidad de incluir una nueva base de conocimiento con cobertura suficiente de estos conceptos geográficos, como por ejemplo, Wikipedia, ya sea en exclusividad o en combinación con WordNet.

En cuanto a la evaluación de la legibilidad se refiere, los resultados de nuestro sistema son cercanos a los obtenidos por los mejores sistemas presentados a la tarea competitiva de la conferencia DUC 2006, a pesar de que estos sistemas disponían de información adicional en forma de consulta de usuario para acometer la tarea. No obstante, los puntos débiles del sistema son, sin duda, la claridad referencial y la estructura y coherencia del resumen generado, que deberán ser mejorados mediante el uso de técnicas de resolución de anáforas y co-referencias (Steinberger et al., 2007), y de simplificación de oraciones (Barzilay y McKeown, 2005; Filippova y Strube, 2008).

Otro criterio con respecto al cual cabe mejorar es el relativo a la detección y eliminación de redundancia, tarea especialmente difícil en el caso de los documentos con los que nos enfrentamos, en el que un elevado porcentaje de la información se encuentra repetida en los distintos documentos a resumir.

8.5. Discusión

A lo largo de este capítulo, se han presentado los resultados de la evaluación del método de generación de resúmenes para tres casos de estudios que trabajan sobre documentos notablemente diferentes entre sí: artículos científicos en biomedicina, noticias periodísticas y páginas web de información turística. Los dos primeros casos de estudio se corresponden con tareas de generación de resúmenes mono-documento, mientras que en el tercero se aborda la tarea de generar resúmenes a partir de múltiples documentos. Se ha comprobado que, con independencia del dominio y la tarea, el sistema obtiene mejores resultados que otros enfoques actuales evaluados respetando las mismas condiciones experimentales.

En primer lugar, se ha realizado una evaluación preliminar encaminada a determinar los valores óptimos de los distintos parámetros que intervienen en el algoritmo. Aunque se ha comprobado que los valores concretos dependen en cierta medida del dominio y la estructura de los documentos considerados, muchas de las conclusiones acerca de dichos valores son equivalentes e independientes del dominio:

- En relación al mejor conjunto de relaciones semánticas a utilizar en la construcción del grafo del documento, los resultados en ambos dominios parecen indicar que cuanto más conexo es el grafo (es decir, a mayor número de relaciones utilizadas), mejores son los resultados obtenidos. Esta conclusión no nos sorprende puesto que, según nuestra hipótesis inicial de que el grafo del documento se comporta como una red libre de escala, la elevada conectividad de los nodos del grafo del documento no es sino una premisa necesaria para que pueda considerarse como una instancia de este tipo de red. Esto explica los pobres resultados obtenidos por el algoritmo cuando sólo se utiliza la relación de hiperonimia para construir el grafo del documento.

- En cuanto al porcentaje de vértices del grafo del documento que se debes utilizar como *hub vertices* en el algoritmo de agrupamiento, en

todos los dominios dicho porcentaje depende de la heurística de selección de oraciones, aunque se observa que el valor es relativamente pequeño para la heurística 1 (2-5 %), algo superior para la heurística 3 (5 %) y relativamente grande para la heurística 2 (10 % y 20 %, respectivamente para cada caso de estudio). Como ya se ha mencionado, la razón subyace en la distinta concepción del resumen por parte de cada una de las heurísticas. En concreto, la heurística 2 concibe el resumen como un compendio de información que cubre someramente todos y cada uno de los temas tratados en el documento, con independencia de la importancia o relevancia de dichos temas. Por ello, no es suficiente considerar como centroides de los grafos aquellos conceptos que representan el tema principal del documento, sino también aquellos que se refieren a información secundaria o complementaria.

- En relación al uso de criterios adicionales de selección de oraciones (i.e. criterio posicional y de similitud con el título), los resultados dependen en gran medida de las características del documento a resumir. Así, al resumir artículos científicos, el uso de ninguno de estos criterios mejora los resultados obtenidos. Sin embargo, cuando se elaboran resúmenes de noticias periodísticas, el uso del criterio posicional se traduce en una mejora de la calidad de estos resúmenes. La razón se encuentra en la estructura de *pirámide invertida* de este tipo de documentos, en los que las primeras oraciones desarrollan los datos de mayor interés y, a continuación, se desarrollan aspectos secundarios. En cuanto al criterio de similitud con el título se refiere, en ambos dominios se observa que este criterio no contribuye a mejorar los resúmenes generados. Este resultado contrasta con el obtenido por otros autores (Edmundson, 1969; Teufel y Moens, 1997), que demuestran que la información contenida en el título del documento puede ayudar a mejorar los resúmenes obtenidos por los enfoques tradicionales basados en las frecuencias de los términos del documento. Por el contrario, este criterio no parece aportar información adicional a la ya capturada con el enfoque de grafos semánticos aquí presentado.

- En cuanto a la mejor heurística se refiere, en todos los casos de estudio la que mejor se comporta es la heurística 3, lo que parece indicar que, en general, un buen resumen incluye cierto grado de información

secundaria y no sólo información directamente relacionada con el tema principal del documento. Esta afirmación, no obstante, depende en cierto grado del ratio de compresión deseado y, en general, a medida que éste disminuye, también lo hace la cantidad de información secundaria contenida en el resumen. Es por ello que en el tercer caso de estudio, donde el ratio de compresión es muy elevado, las heurísticas 1 y 3 presentan resultados muy similares.

En segundo lugar, se ha realizado una evaluación a gran escala de cada uno de los tres casos de estudio analizados. En todos ellos, el sistema ha mostrado comportarse mejor que otros trabajos actuales evaluados respetando las mismas condiciones experimentales. Se ha visto, además, que resulta sencillo adaptar el método para generar resúmenes a partir de múltiples documentos, sin más que utilizar un mecanismo para detectar y eliminar la redundancia como paso previo a la generación del resumen.

En cuanto al efecto de la ambigüedad léxica en la tarea que nos ocupa, se ha comprobado cómo al utilizar un algoritmo de desambiguación para identificar los significados correctos de los términos ambiguos, la calidad de los resúmenes mejora. Este resultado era previsible, puesto que el algoritmo diseñado hace un uso intensivo de conocimiento, y, en caso de malinterpretar el significado de los términos del documento, la selección de oraciones relevantes se ve necesariamente perjudicada.

Finalmente, la evaluación manual de la legibilidad y la calidad gramatical del resumen realizada para el tercer caso de estudio ha puesto de manifiesto la necesidad de utilizar técnicas de detección y resolución de referencias anafóricas como parte de todo sistema de generación de resúmenes mediante extracción; así como la conveniencia de utilizar técnicas de condensación y simplificación de oraciones, e incluso de abstracción y reescritura de texto, para asegurar que el resumen final presenta una estructura coherente y conexa.

Capítulo 9

Conclusiones y Trabajo Futuro

9.1. Conclusiones

En esta memoria de tesis doctoral se ha presentado un método genérico para la realización de resúmenes de texto. La principal aportación radica en el uso de conocimiento específico sobre la estructura y el dominio al que pertenecen los documentos a resumir para mejorar la calidad de los resúmenes finales, a la vez que se proporciona un entorno fácilmente configurable para trabajar sobre nuevos tipos de documentos. El sistema desarrollado permite, además, elaborar tanto resúmenes mono-documento como resúmenes multi-documento. La finalidad de estos resúmenes es la de convertirse en una ayuda eficaz para los usuarios de sistemas de acceso a la información.

El método propuesto para la generación automática de resúmenes se basa en la representación del documento como un grafo extendido de conceptos y relaciones extraídos de una base de conocimiento (cuya elección dependerá del dominio al que pertenezcan los documentos a resumir), y en el cálculo de la relevancia de las oraciones a extraer, en relación al prestigio de los conceptos en el grafo del documento. De este modo, se construye una representación más rica en conocimiento que la utilizada por los enfoques tradicionales basados en términos. El método propuesto implementa distintas heurísticas para la selección de las oraciones para el resumen, cada una de ellas ideada para generar un tipo distinto de resumen en función de

la información que se desee contemplar, a la vez que permite utilizar otras métricas o criterios tradicionales para la selección de las oraciones para el resumen, como la posición de las oraciones en el documento o su similitud con el título del mismo. Dependiendo del dominio de aplicación, algunos de estos criterios pueden ser de gran utilidad.

En el Capítulo 2 se ha realizado una extensa revisión del estado del arte en generación automática de resúmenes, tanto mono-documento como multi-documento, 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 la detección de temas, la desambiguación léxica, la resolución de referencias anafóricas y pronominales, la detección y eliminación de redundancia, la resolución de acrónimos y abreviaturas, o la simplificación y compresión de texto; y que por lo tanto, puede y debe nutrirse de los métodos y técnicas utilizados en cada una de ellas.

De entre todas las sub-tareas que intervienen en la generación automática de resúmenes, el sistema desarrollado en este trabajo aborda la mayoría de ellas: es capaz de detectar los distintos temas tratados en el documento o documentos a resumir, incorpora un módulo para resolver la ambigüedad del texto y asociar cada posible término ambiguo con su significado correcto en función del contexto en el que se utiliza, permite detectar abreviaturas y acrónimos y reemplazarlos por sus correspondientes definiciones o formas expandidas y, en el caso de abordar el problema de generación de resúmenes multi-documento, puede ser fácilmente adaptado para detectar aquella información que se repite a lo largo de los distintos documentos, con el objetivo de no incluirla repetida en el resumen. Para todo lo anterior, bien utiliza desarrollos a medida, bien se nutre de otras herramientas o aplicaciones de libre distribución que resuelven satisfactoriamente las distintas sub-tareas identificadas (ver Capítulo 3). Además, al tratarse de un sistema genérico fácilmente configurable, permite adecuar la elección de dichas herramientas al dominio o tipo de los documentos a resumir.

En los Capítulos 5, 6 y 7 se ha detallado el proceso realizado para con-

figurar la arquitectura implementada con el objetivo de generar resúmenes de tres tipos de documentos muy dispares: artículos científicos sobre biomedicina, noticias periodísticas y páginas web de información turísticas. Los dos primeros casos de estudio se corresponden con tareas de generación de resúmenes mono-documento, mientras que en el tercer caso de estudio se aborda el problema de generar resúmenes a partir de múltiples documentos sobre un mismo tema o suceso. Se ha comprobado que, en cualquier caso, el proceso de configuración resulta relativamente sencillo: basta con modificar la base conocimiento a utilizar y el mecanismo para traducir el documento a conceptos de la base de conocimiento.

En el Capítulo 8 se han presentado los resultados de evaluar los resúmenes generados, utilizando las distintas configuraciones propuestas en los casos de estudio, para documentos de tres colecciones de evaluación de resúmenes automáticos. Para el primer caso de estudio (generación de resúmenes de artículos científicos en biomedicina), el corpus de evaluación ha sido expresamente construido como parte de este trabajo. Para los restantes casos de estudio, se han utilizado colecciones disponibles y utilizadas por otros investigadores en la misma tarea. En cualquier caso, los resultados de estas evaluaciones han sido comparados con los obtenidos por otros sistemas evaluados respetando las mismas condiciones experimentales.

Los resultados de las pruebas realizadas para validar el método propuesto muestran, en primer lugar, que el uso de recursos de conocimiento para capturar el significado y las relaciones entre los términos del documento y construir una representación semántica del mismo, permite elaborar resúmenes de mayor calidad que los enfoques tradicionales basados en términos y en meras representaciones sintácticas. Estos mismos experimentos indican que el uso de técnicas de desambiguación como parte del procesamiento del documento permite conseguir resultados significativamente mejores en la evaluación de los resúmenes. Paradójicamente, la autora no conoce ningún trabajo en el que se estudie la influencia de la ambigüedad en la generación automática de resúmenes. La misma conclusión se extrae sobre la influencia de la presencia de acrónimos y abreviaturas en el documento a resumir: aunque en este caso las ventajas de detectar y expandir tales formas abreviadas no es tan evidente, sí es cierto que traducen en una mejora de la calidad de los resúmenes generados

Otra conclusión interesante es el hecho constatado de que, dependiendo

del dominio y tipo de documentos a resumir, las heurísticas y criterios a utilizar en la selección de las oraciones para el resumen varían: así, hemos comprobado que a la hora de resumir artículos científicos, la información sobre la posición que la oración ocupa en el documento resulta inútil, mientras que esta misma información se muestra muy eficaz a la hora de resumir noticias periodísticas. Por tanto, todo sistema que aspire a generar resúmenes de calidad deberá tener en cuenta la tipología y estructura de los documentos a resumir. Por otro lado, los resultados obtenidos por las distintas heurísticas parecen indicar que, en general, un buen resumen incluye cierto grado de información secundaria y no sólo información directamente relacionada con el tema principal del documento.

Finalmente, la evaluación manual de la legibilidad realizada sobre los resúmenes generados en el tercer caso de estudio (páginas web de información turística) ha permitido vislumbrar cuáles son los problemas más evidentes del método propuesto y, en consecuencia, hacia dónde se ha de dirigir el trabajo futuro más inmediato. Tales problemas no son sino 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 la falta de coherencia y claridad referencial de los resúmenes generados, lo que pone de manifiesto la necesidad de utilizar técnicas de detección y resolución de referencias anafóricas y la conveniencia de utilizar técnicas de condensación y simplificación de oraciones para asegurar que el resumen final presenta una estructura coherente y conexa. Ambos aspectos se analizan en detalle en la siguiente sección.

9.2. Trabajo Futuro

Los resultados de los experimentos presentados en esta memoria de tesis son alentadores, aunque han subrayado la necesidad de continuar con la investigación en varias direcciones. Podemos clasificar el trabajo a realizar en un futuro en dos categorías: acciones necesarias para corregir o solventar los problemas de la versión actual del generador de resúmenes y acciones encaminadas a ampliar el sistema de modo que permita realizar nuevos y más sofisticados tipos de resúmenes.

Dentro de la primera categoría, esto es, de las acciones correctoras, se encontraría el trabajo a realizar para mejorar la estructura y coherencia de

los resúmenes. Esto incluye, como ya se ha comentado, estudiar la aplicabilidad de distintos métodos para la detección y resolución de referencias anafóricas, y de técnicas de simplificación y condensación de oraciones.

En cuanto a resolución de referencias se refiere, en ocasiones se observa cómo en una oración un concepto relevante es referido 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. La solución más simple y evidente a este problema es utilizar un sistema de resolución anafórica como paso previo a la generación del resumen, de manera que se identifiquen todas los elementos referidos y sus respectivas referencias, y se sustituyan en el documento los primeros por las segundas (Steinberger et al., 2007). Pero aún así, quedaría sin resolverse un problema mayor: si al final la oración que contiene la referencia resulta seleccionada, pero no así la oración que contiene el elemento referido, el resultado será un resumen incoherente e inconexo. En este sentido, algunas investigaciones nos muestran posibles soluciones. Se podría optar por no incluir las frases que contengan estas referencias (Brandow, Mitze, y Rau, 1995; Steinberger et al., 2007), incluir la frase anterior a la seleccionada (Nanba 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 tener que seleccionar las oraciones anteriores, debemos dejar de añadir otras frases que pudieran ser más importantes para respetar el ratio de compresión deseado. Cabría pensar que una solución relativamente sencilla sería sustituir, también en el resumen, todos los elementos referidos por sus referencias, pero ello daría lugar a un resumen con excesiva repetición de determinadas referencias. El problema es, por lo tanto, más complejo de lo que a priori cabría esperar, y necesitará ser estudiado a fondo antes de apostar por una propuesta de solución definitiva.

En cuanto a simplificación y condensación de oraciones se refiere, el problema consiste en, dado un conjunto de oraciones, construir una única oración que fusione la información de todas ellas en un menor número de palabras, sin información repetida, y obviando aquella información que pudiera ser secundaria o prescindible. Así, por ejemplo, si se presentan las tres siguientes oraciones en el documento a resumir:

- *Ana, que ahora tiene 40 años, se licenció en Matemáticas en la Universidad Complutense de Madrid.*

- *Posteriormente, Ana se doctoró en Economía en esta misma universidad.*
- *Ana trabaja desde 1998 en la Universidad Complutense.*

El objetivo será condensar toda esta información en una única oración como, por ejemplo, *Ana se licenció en Matemáticas y se doctoró en Economía en la Universidad Complutense de Madrid, donde trabaja desde 1998.*

Como vemos, se trata de un problema complejo que implica, como mínimo, determinar qué información es importante a un nivel más específico que la oración (por ejemplo, a nivel de sintagma nominal o de proposición), analizar las distintas oraciones para identificar su estructura sintáctica, sus constituyentes, y sus elementos primarios de información (por ejemplo, *Ana tiene 40 años. Ana se licenció en Matemáticas.* etc.), construir una representación intermedia de esta información, y traducirla al texto final del resumen utilizando técnicas de generación de lenguaje.

Dentro de la segunda categoría de trabajos futuros, aquellos encaminados a ampliar la funcionalidad del sistema, se está estudiando una posible modificación del algoritmo para generar resúmenes adaptados a una consulta del usuario. Mientras que los resúmenes que podríamos denominar "generalistas" van dirigidos a un amplio espectro de lectores, los resúmenes guiados por una consulta van dirigidos a satisfacer las necesidades concretas de información de un lector o grupo de lectores. El método desarrollado en esta tesis doctoral puede ser fácilmente modificado para aprovechar la información proporcionada por el usuario en beneficio de la generación del resumen. Bastaría con modificar la función destinada a calcular el peso de las aristas del grafo del documento, de manera que, si una arista tiene como origen o destino un vértice que representa un concepto presente en la consulta del usuario, el peso de dicha arista se incrementa. De este modo, el peso se distribuye a través de la red del documento, de tal forma que los conceptos de la consulta y aquellos con los que se encuentran estrechamente relacionados aumentan su *salience* o prestigio en la red. Como resultado, las oraciones que contienen conceptos relacionados con la consulta del usuario incrementan su probabilidad de ser seleccionadas para el resumen.

Finalmente, otra línea de trabajo futuro será el desarrollo e integración en el sistema de generación de resúmenes de un módulo de detección de redundancia, de manera que se pueda abordar la generación de resúmenes multi-documento sin necesidad de acudir a herramientas externas.

Bibliografía

Afantenos, S.D., V. Karkaletsis, y P. Stamatopoulos. 2005. Summarization from Medical Documents: A Survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

Agirre, E. y P. Edmonds, editores. 2006. *Word Sense Disambiguation: Algorithms and Aplications*. Springer.

Agirre, E. y A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 33–41.

Aker, A. y R. Gaizauskas. 2009. Summary Generation for Toponym-Referenced Images using Object Type Language Models. En *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, páginas 6–11.

Aker, A. y R. Gaizauskas. 2010. Generating Image Descriptions using Dependency Relational Patterns. En *Proceedings of the Association of Computational Linguistic*, páginas 1250–1258.

Amini, M-R. y P. Gallinari. 2002. The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. En *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 105–112.

Antiqueira, L., T. A. S. Pardo, M. das G. V. Nunes, y O. N. Oliveira Jr. 2007. Some Issues on Complex Networks for Author Characterization. *Revista Iberoamericana de Inteligencia Artificial*, 11(36):51–58.

Aronson, A. R. y F-M. Lang. 2010. An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association*, 17:229–236.

Banko, M. y L. Vanderwende. 2004. Using N-grams to Understand the Nature of Summaries. En *Proceedings of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies*, páginas 1–4.

Barabási, A.L. y R. Albert. 1999. Emergence of Scaling in Random Networks. *Science*, 268:509–512.

Barzilay, R. y M. Elhadad. 1997. Using Lexical Chains for Text Summarization. En *Proceedings of the Association for Computational Linguistics, Workshop on Intelligent Scalable Text Summarization*, páginas 10–17.

Barzilay, R. y M. Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Barzilay, R. y K. R. McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–327.

Bawakid, A. y M. Oussalah. 2008. A Semantic Summarization System: University of Birmingham at TAC 2008. En *Proceedings of the First Text Analysis Conference*.

Baxendale, P.B. 1958. Man-Made Index for Technical Literature: An Experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Binwahlan, M. S., N. Salim, y L. Suanmali. 2009. Swarm Based Features Selection for Text Summarization. *International Journal of Computer Science and Network Security*, 9(1):175–179.

Blair-Goldensohn, S., D. Evans, V. Hatzivassiloglou, K. McKeown, A. Nenkova, R. Passonneau, B. Schiffman, A. Schlaikjer, A. Siddharthan, y S. Siegelman. 2004. Columbia University at DUC 2004. En *Proceedings of the 4th Document Understanding Conference at North American Chapter of the Association for Computational Linguistics, Human Language Technologies*.

Bodenreider, O., J. A. Mitchell, y A. T. McCray. 2003. Biomedical Ontologies: Session Introduction. En *Proceedings of the Pacific Symposium on Biocomputing*, páginas 562–564.

Boguraev, B. y J. Pustejovsky. 1996. *Corpus Processing for Lexical Acquisition*. The MIT Press.

Bossard, A., M. Généreux, y T. Poibeau. 2008. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. En *Proceedings of the 1st Text Analysis Conference*.

Bouayad-Agha, N., G. Casamayor, G. Ferraro, S. Mille, V. Vidal, y L. Wanner. 2009. Improving the Comprehension of Legal Documentation: The Case of Patent Claims. En *Proceedings of the International Conference on Artificial Intelligence and Law*, páginas 78–87.

Brandow, R., K. Mitze, y L. F. Rau. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, 5(31):675–685.

Brin, S. y L. Page. 1998. The Anatomy of a Largescale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30:1–7.

Cabré, M. T. 1995. La Terminología Hoy: Concepciones, Tendencias y Aplicaciones. *Ciência da Informação*, 24(3).

Carbonell, J., Y. Geng, y J. Goldstein. 1997. Automated Query-relevant Summarization and Diversity-based Reranking. En *Proceedings of the International Joint Conferences on Artificial Intelligence, Workshop on Artificial Intelligence in Digital Libraries*, páginas 12–19.

Carbonell, J. y J. Goldstein. 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*, páginas 335–336.

Carenini, G., R. T. Ng., y X. Zhou. 2008. Summarizing Emails with Conversational Cohesion and Subjectivity. En *Proceedings of the Association for Computational Linguistics - Human Language Technologies*, páginas 353–361.

Cassidy, P. 2000. An Investigation of the Semantic Relations in the Roget's Thesaurus: Preliminary Results. En *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, páginas 181–204.

Chklovski, T. y P. Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, páginas 33–40.

Chuang, W. T. y J. Yang. 2000. Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. En *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 454–457.

Curtis, J., J. Cabral, y D. Baxter. 2006. On the Application of the Cyc Ontology to Word Sense Disambiguation. En *Proceedings of the Florida Artificial Intelligence Research Society Conference*, páginas 652–657.

Curtis, J., M. Gavin, y D. Baxter. 2005. On the Effective Use of Cyc in a Question Answering System. En *Proceedings of the International Joint Conferences on Artificial Intelligence, Knowledge and Reasoning for Answering Questions Workshop*, páginas 61–70.

Dang, H.T. 2005. Overview of DUC 2005. En *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Document Understanding Conference 2005 Workshop*.

Dang, H.T. 2006. Overview of DUC 2006. En *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Document Understanding Conference 2006 Workshop*.

Díaz, A. y P. Gervás. 2005. Personalisation in News Delivery Systems: Item Summarization and Multi-tier Item Selection Using Relevance Feedback. *Web Intelligence and Agent Systems*, 3(2):135–154.

DeJong, G.F. 1982. An Overview of the FRUMP System. En *Strategies for Natural Language Processing*. Lawrence Erlbaum, páginas 149–176.

Divita, G., T. Tse, y L. Roth. 2004. Failure Analysis of MetaMap Transfer (MMTx). *Health Technol Inform*, 107:763–767.

Donaway, R. L., K. W. Drummey, y L. A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. En *Proceedings of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic Summarization*, páginas 69–78.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 2(16):264–285.

Erdös, P. y A. Rényi. 1959. On random graphs. I. *Publicationes Mathematicae 6*, páginas 290–297.

Erkan, G. y D. Radev. 2004a. LexPageRank: Prestige in Multi-Document Text Summarization . En *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, páginas 365–371.

Erkan, G. y D. R. Radev. 2004b. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Ferrández, O., D. Micol, R. Muñoz, y M. Palomar. 2007. A Perspective-Based Approach for Solving Textual Entailment Recognition. En *Proceedings of the Association for Computational Linguistics, PASCAL Workshop on Textual Entailment and Paraphrasing*, páginas 66–71.

Ferrer-Cancho, R. y R. V. Solé. 2001. The Small World of Human Language. En *Proceedings of the Royal Society of London*, volumen 268, páginas 2261–2266.

Filippova, K. y M. Strube. 2008. Sentence Fusion via Dependency Graph Compression. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, páginas 177–185.

Filippova, K., M. Surdeanu, M. Ciaramita, y H. Zaragoza. 2009. Company-oriented Extractive Summarization of Financial News. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 246–254.

Fum, D., G. Gmda, y C. Tasso. 1985. Evaluating Importance: A Step Towards Text Summarization. En *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, páginas 840–844.

Goldstein, J., V. Mittal, J. Carbonell, y M. Kantrowitz. 2000. Multi-document Summarization by Sentence Extraction. En *Proceedings of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic summarization*, páginas 40–48.

González, E. y M. Fuentes. 2009. A New Lexical Chain Algorithm Used for Automatic Summarization. En *Proceeding of the Conference on Artificial Intelligence Research and Development*, páginas 329–338.

Gruber, T. R. 1993. A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, 5(2):199–220.

Hahn, U. y I. Mani. 2000. The Challenges of Automatic Summarization. *Computer*, 33(11):29–36.

Hahn, U. y U. Reimer, 1999. *Advances in Automatic Text Summarization*, capítulo Knowledge-based Text Summarization: Salience and Generalization Operators for Knowledge Base Abstraction, páginas 215–232. The MIT Press.

Halliday, M. 1985. *An Introduction to Functional Grammar*. Edward Arnold.

Halliday, M. y R. Hasan. 1996. *Cohesion in English*. Longsmans.

Hatzivassiloglou, V., J. Klavans, M. Holcombe, R. Barzilay, M.Y. Kan, y K.R. McKeown. 2001. SimFinder: A Flexible Clustering Tool for Summarization. En *Proceedings of the North American Chapter of the Association for Computational Linguistics, Automatic Summarization Workshop*, páginas 41–49.

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

Hendrickx, I., W. Daelemans, E. Marsi, y E. Krahmer. 2009. Reducing Redundancy in Multi-document Summarization Using Lexical Semantic

Similarity. En *Proceedings of the Association for Computational Linguistics - International Joint Conference on Natural Language Processing, Workshop on Language Generation and Summarisation*, páginas 63–66.

Herrera, J., A. Peñas, y F. Verdejo. 2005. Técnicas Aplicadas al Reconocimiento de Implicación Textual. En *Proceedings of the 11th Conferencia de la Asociación Española para la Inteligencia Artificial*, volumen 2.

Hirst, G. y D. St Onge, 1998. *Lexical Chains as Representation of Context for the Detection and Correction Malapropisms*. The MIT Press.

Hobbs, J. 1985. On the Coherence and Structure of Discourse. *CSLI Technical Report*, páginas 85–37.

Hovy, E. 2005. Automated Text Summarisation. En *Handbook of Computational Linguistics*. Oxford University Press.

Hovy, E. y C-Y. Lin. 1999. *Automated Text Summarization in SUMMARIST*. MIT Press.

Humphrey, S. M., W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, y T. C. Rindflesch. 2006. Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113.

Jaccard, P. 1901. Étude Comparative de la Distribution Florale Dans une Portion des Alpes et des Jura. *Bulletin de la Societé Vaudoise des Sciences Naturelles*, 37:547–579.

Jarmasz, M. y S. Szpakowicz. 2003. Not as Easy as It Seems: Automating the Construction of Lexical Chains Using Roget's Thesaurus. En *Proceedings of the 16th Canadian Conference on Artificial Intelligence*, páginas 544–549.

Jiang, J. J. y D. W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. En *Proceedings of the International Conference Research on Computational Linguistics*, páginas 19–33.

Jing, H. 2002. Using Hidden Markov Modeling to Decompose Human-written Summaries. *Computational Linguistics*, 28(4):527–543.

Jing, H., K. R. McKeown, R. Barzilay, y M. Elhadad. 1998. Summarization Evaluation Methods: Experiments and Analysis. En *Proceedings of the AAAI Symposium on Intelligent Text Summarization*, páginas 60–68.

Kazantseva, A. y S. Szpakowicz. 2010. Summarizing Short Stories. *Computational Linguistics*, 36(1):71–109.

Klingberg, T. 2009. *The Overflowing Brain: Information Overload and the Limits of Working Memory*. Oxford University Press Inc.

Knight, K. y D. Marcu. 2000. Statistics-based Summarization - Step One: Sentence Compression. En *Proceedings of the National Conference on Artificial Intelligence*, páginas 703–710.

Kupiec, J., J. O. Pedersen, y F. Chen. 1995. A Trainable Document Summarizer. En *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 68–73.

Leacock, C. y M. Chodorow, 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, capítulo 11, páginas 265–283. The MIT Press.

Lee, L. 1999. Measures of Distributional Similarity. En *Proceedings of the 37th Conference on Association for Computational Linguistics*, páginas 25–32.

Legendre, P. y L. Legendre. 1998. *Numerical Ecology*. The Netherlands, Amsterdam.

Lenat, D. B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Comunications of the ACM Transactions on Speech and Language Processing*, 38(11):33–38.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from a Ice Cream Cone. En *Proceedings of Special Interest Group on Design of Communication*, páginas 24–26.

Levenshtein, V. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Li, W., M. Wu, Q. Lu, W. Xu, y C. Yuan. 2006. Extractive Summarization Using Inter- and Intra- event Relevance. En *Proceedings of the 21st International Conference on Computational Linguistics*, páginas 369–376.

Lin, C-Y. 1999. Training a Selection Function for Extraction. En *Proceedings of the the Eighteenth International Conference on Information and Knowledge Management*, páginas 1–8.

Lin, C-Y. 2004a. Looking for a Few Good Metric: Automatic Summarization Evaluation - How Many Samples are Enough? En *Proceedings of the NII Test Collection for Information Retrieval Systems, Workshop 4.*

Lin, C-Y. 2004b. Rouge: A Package for Automatic Evaluation of Summaries. En *Proceedings of the Association for Computational Linguistics, Workshop: Text Summarization Branches Out*, páginas 74–81.

Lin, C-Y. y E. Hovy. 1997. Identifying Topic by Position. En *Proceedings of the 5th Conference on Applied Natural Language Processing*, páginas 283–290.

Lin, C-Y. y E. Hovy. 2001. NEATS: A Multidocument Summarizer. En *Proceedings of the Workshop on Text Summarization in Conjunction with the ACM SIGIR Conference*, páginas 1–4.

Lin, C-Y. y E. Hovy. 2002a. From Single to Multi-document Summarization: A Prototype System and its Evaluation. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, páginas 457–464.

Lin, C-Y. y E. Hovy. 2002b. Manual and Automatic Evaluation of Summaries. En *Proceedings of the Document Understanding Conference, Workshop on Automatic Summarization*, páginas 45–51.

Lin, C-Y y E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. En *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, páginas 71–78.

Lin, D. 1998. An Information-theoretic Definition of Similarity. En *Proceedings of the 15th International Conference on Machine Learning*, páginas 296–304.

Litkowski, K. C. 2004. Summarization Experiments in DUC 2004. En *Proceedings of the Document Understanding Conference*.

Liu, M., W. Li, M. Wu, y Q. Lu. 2007. Extractive Summarization Based on Event Term Clustering. En *Proceedings of the Association for Computational Linguistics, Demo and Poster Sessions*, páginas 185–188.

Lloret, E., O. Ferrández, R. Muñoz, y M. Palomar. 2008. A Text Summarization Approach under the Influence of Textual Entailment. En *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science in Conjunction with the 10th International Conference on Enterprise Information Systems*, páginas 22–31.

Longacre, R. 1979. The Discourse Structure of the Flood Narrative. *Journal of the American Academy of Religion*, 47(1):89–133.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Maña, M., M. de Buenaga, y J. M. Gómez. 1999. Using and Evaluating User Directed Summaries to Improve Information Access. En *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, páginas 198–214.

Mani, I. 2001. *Automatic Summarization*. Jonh Benjamins Publishing Company.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, y B. Sundheim. 2001. SUMMAC: A Text Summarization Evaluation. *Natural Language Engineering*, 8(1):43–68.

Mann, W. y S. Thompson. 1988. Rethorical Structure Theory: Towards a Functional Theory of Text Organisation. *Text*, 8(3):243–281.

Marcu, D., 1999. *Advances in Automatic Text Summarization*, capítulo Discourse Trees Are Good Indicators of Importance in Text, páginas 123–136. The MIT Press.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Marcu, D. y L. Gerver. 2001. An Inquiry into the Nature of Multidocument Abstracts, Extracts, and their Evaluation. En *Proceedings of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic Summarization*, páginas 1–8.

Mckeown, K., J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, y E. Eskin. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. En *Proceedings of American Association for Artificial Intelligence*, páginas 453–460.

McKeown, K., J. Robin, y K. Kukich. 1995. Generating Concise Natural Language Summaries. *Information Processing and Management*, 31(5):703–733.

Metzler, D. y T. Kanungo. 2008. Machine Learned Sentence Selection Strategies for Query-Biased Summarization. En *Proceedings of the Special Interest Group on Information Retrieval Conference, Learning to Rank for Information Retrieval Workshop*.

Mihalcea, R. y P. Tarau. 2004. TextRank: Bringing Order into Texts. En *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, páginas 404–411.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K. Miller. 1998. Five Papers on WordNet. En *WordNet: An Electronic Lexical Database*. The MIT Press.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K. J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Morris, A., G. Kasper, y D. Adams. 1992. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research*, 3(1):17–35.

Morris, J. y G. Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

Nadkarni, P.M. 2000. Information Retrieval in Medicine: Overview and Applications. *Journal of Postgraduate Medicine*, 46(2):122–166.

Nanba, H. y M. Okumura. 2000. Producing More Readable Extracts by Revising Them. En *Proceedings of the 18th International Conference on Computational Linguistics*, páginas 1071–1075.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Nenkova, A. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. En *Proceedings of the 20th National Conference on Artificial Intelligence*, volumen 3, páginas 1436–1441.

Nenkova, A., R. Passonneau, y K. McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Neto, J. L., A. A. Freitas, y C. A. A. Kaestner. 2002. Automatic Text Summarization Using a Machine Learning Approach. En *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*, páginas 205–215.

Otterbacher, J., G. Erkan, y D. R. Radev. 2005. Using Random Walks for Question-focused Sentence Retrieval. En *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, páginas 915–922.

Ouyang, Y., W. Li, y Q. Lu. 2009. An Integrated Multi-document Summarization Approach Based on Word Hierarchical Representation. En *Proceedings of the Association for Computational Linguistics, International Joint Conference on Natural Language Processing*, páginas 113–116.

Paice, C. D. 1980. The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-indicating Phrases. En *Proceedings of SIGIR*, páginas 172–191.

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

Paice, C. D. y P. A. Jones. 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*, páginas 69–78.

Passonneau, R.J., A. Nenkova, K. McKeown, y S. Sigelman. 2005. Applying the Pyramid Method in DUC 2005. En *Proceedings of the 5th Document Understanding Conference*.

Patwardhan, S. 2003. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota.

Patwardhan, S., S. Banerjee, y T. Pedersen. 2005. SenseRelate::TargetWord: A Generalized Framework for Word Sense Disambiguation. En *Proceedings of the Association for Computational Linguistics*, páginas 73–76.

Pedersen, T., S. Patwardhan, y J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. En *Proceedings of the 19th National Conference on Artificial Intelligence*, páginas 1024–1025.

Perea, J.M., M.T. Martín, A. Montejo, y M.C. Díaz. 2008. Categorización de Textos Biomédicos Usando UMLS. *Procesamiento del Lenguaje Natural*, 40:121–127.

Plaza, L., E. Lloret, y A. Aker. 2010. Improving Automatic Image Captioning Using Text Summarization Techniques. En *Proceedings of the 13th International Conference on Text, Speech and Dialogue*, páginas 165–172.

Radev, D. 2000. A Common Theory of Information Fusion from Multiple Text Sources. Step One: Cross Document Structure. En *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue*, páginas 74–83.

Radev, D., S. BlairGoldensohn, y Z. Zhang. 2001. Experiments in Single and MultiDocument Summarization Using MEAD. En *Proceedings of the Document Understanding Conference*.

Radev, D. R., H. Jing, y M. Budzikowska. 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based

Evaluation, and User Studies. En *Proceeding of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic Summarization*, páginas 21–30.

Radev, D. R. y K. R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):470–500.

Reeve, L. H., H. Han, y A. D. Brooks. 2007. The Use of Domain-specific Concepts in Biomedical Text Summarization. *Information Processing and Management*, 43:1765–1776.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. En *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, páginas 448–453.

Riloff, E. 1996. Automatically Generating Extraction Patterns from Untagged Text. En *Proceedings of the 13th National Conference on Artificial Intelligence*, volumen 2, páginas 1044–1049.

Rindfleisch, T. C., L. Tanabe, y J. N. Weinstein. 2000. EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature. En *Proceedings of the Pacific Symposium on Biocomputing*, páginas 517–528.

Rush, J. E., A. Zamora, y R. Salvador. 1971. Automatic Abstracting and Indexing II. Production of Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. *Journal of the American Society for Information Science*, 22(4):260–274.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Revue Traitement Automatique des Langues*, 49(2):103–125.

Saggion, H., E. Lloret, y M. Palomar. 2010. Using Text Summaries for Predicting Rating Scales. En *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Held in conjunction to ECAI 2010*, páginas 44–51.

Salton, G. y M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Salton, G., A. Singhal, M. Mitra, y C. Buckley. 1997. Automatic Text Structuring and Summarization. *Information Processing and Management*, 33(2):193–207.

Sanderson, M. 1998. Accurate User Directed Summarization from Existing Tools. En *Proceedings of the 7th International Conference on Information and Knowledge Management*, páginas 45–51.

Schwartz, A. y M. Hearst. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. En *Proceedings of the Pacific Symposium on Biocomputing*, páginas 451–462.

Skorochod'ko, E. F. 1972. Adaptive Method of Automatic Abstracting and Indexing. *Information Processing*, 71.

Sneiderman, C. A., T. C. Rindflesch, y C. A. Bean. 1998. Identification of Anatomical Terminology in Medical Text. En *Proceedings of the American Medics Informatics Association Symposium*, páginas 428–432.

Sparck-Jones, K. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–20.

Sparck-Jones, K. 1999. *Automatic Summarising: Factors and Directions*. The MIT Press.

Sparck-Jones, K. y J. R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag New York, Inc.

Steinberger, J., M. Poesio, M. A. Kabadjov, y K. Jezek. 2007. Two Uses of Anaphora Resolution in Summarization. *Information Processing and Management*, 43(6):1663–1180.

Teufel, S. y M. Moens. 1997. Sentence Extraction as a Classification Task. En *Proceedings of the Association for Computational Linguistics, Workshop on Intelligent Scallable Text Summarization*, páginas 58–65.

Tombros, A. y M. Sanderson. 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*, páginas 2–10.

Tseng, Y-H. 2009. Summarization Assistant for News Brief Services on Cellular Phones. *Computational Linguistics and Chinese Language Processing*, 1(14):85–104.

Van Dijk, T. 1988. *News as Discourse*. Erlbaum Associates.

Wan, S. y K. McKeown. 2004. Generating Overview Summaries of Ongoing Email Thread Discussions. En *Proceedings of the 20th International Conference on Computational Linguistics*.

Wan, X., J. Yang, y J. Xiao. 2007. Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. En *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, páginas 552–559.

Watts, D.J. y S.H. Strogatz. 1998. Collective Dynamics of Small World Networks. *Nature*, 393:440–442.

Wei, F., W. Li, y Y. He. 2009. Xo-Feedback Ranking for Query-Focused Summarization. En *Proceedings of the Association for Computational Linguistics, International Joint Conference on Natural Language Processing*, páginas 117–120.

Weigand, H. 1997. A Multilingual Ontology-based Lexicon for News Filtering - The TREVI Project. En *Proceedings of the International Joint Conferences on Artificial Intelligence, Workshop Ontologies and Multilingual NLP*.

Wilks, Y. y M. Stevenson. 1996. The Grammar of Sense: Is Word-sense Tagging Much More than Part-of-speech Tagging? En *Proceedings of the Computing Research Repository , cmp-lg/9607028*.

Witbrock, M., D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, C. Matthews, y A. Vizedom. 2003. An Interactive Dialogue System for Knowledge Acquisition in Cyc. En *Proceedings of the Workshop on Mixed Initiative Intelligent Systems*, páginas 138–145.

Witbrock, M. J. y V. O. Mittal. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. En *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, páginas 315–316.

Wu, Z. y M. Palmer. 1994. Verb Semantics and Lexical Selection. En *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, páginas 133–138.

Yarowsky, D. 1992. Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. En *Proceedings of the 15th International Conference of the Association for Computational Linguistics*, páginas 454–460.

Yoo, I., X. Hu, y I-Y. Song. 2007. A Coherent Graph-based Semantic Clustering and Summarization Approach for Biomedical Literature and a New Summarization Evaluation Method. *BMC Bioinformatics*, 8(9).

Zhang, Z., S. Blair-Goldensohn, y D. Radev. 2002. Towards CST-Enhanced Summarization. En *Proceedings of the Conference on Artificial Intelligence*.

Zhao, L., L. Wu, y X. Huang. 2009. Using Query Expansion in Graph-based Approach for Query-focused Multi-document Summarization. *Information Processing and Management*, 45:35–41.

Apéndice A

Publicaciones

A continuación se muestran las publicaciones que han resultado de la realización de esta tesis doctoral.

A.1. Generación de Resúmenes y Procesamiento de Información en el Dominio Biomédico

1. Plaza L, Díaz A, Gervás P. 2008. Concept-graph based Biomedical Automatic Summarization using Ontologies. En *Proceedings of the workshop "TextGraphs-3: Graph-based Algorithms for Natural Language Processing", held in conjunction with the International Conference on Computational Linguistics (COLING 2008)*, páginas 53-56. Manchester, Reino Unido.

2. Plaza L, Díaz A, Gervás P. 2008. Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina. En *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*, número 41, septiembre de 2008, páginas 191-198.

3. Plaza L, Carrillo de Albornoz J, Prados J. 2010. Sistemas de Acceso Inteligente a la Información Biomédica: una Revisión. En *Revista Internacional de Ciencias Podológicas*, volumen 4, número 1, páginas 7-15.

4. Plaza L, Díaz A. 2010. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. En *Proceedings of the International*

Conference on Applications of Natural Language to Information Systems. Lecture Notes in Computer Sciences, 6177, páginas 296-303. Cardiff, Reino Unido.

5. Plaza L, Stevenson M, Díaz A. 2010. Improving Summarization of Biomedical Documents using Word Sense Disambiguation. En *Proceedings of the workshop "BioNLP 2010" held in conjunction with the 48th Annual Meeting of the Association for Computational Linguistics*, páginas 55-63, Uppsala, Suecia.

6. Plaza L, Díaz A, Gervás P. A Semantic Graph-based Approach to Biomedical Summarization. En *Artificial Intelligence in Medicine*. Elsevier. En proceso de revisión tras ser aceptado con cambios.

7. Plaza L, Stevenson M, Díaz A. Resolving Ambiguity in Biomedical Text to Improve Summarization. En *Information Processing and Management*. En proceso de revisión.

A.2. Generación de Resúmenes de Noticias Periodísticas

1. Plaza L, Díaz A, Gervás P. 2009. Automatic Summarization of News using Wordnet concept graphs. En Proceedings of the *IADIS International Conference Informatics*. Algarve, Portugal. Best Paper Award.

2. Plaza L, Díaz A, Gervás P. 2010. Automatic Summarization of News using Wordnet concept graphs. En *IADIS International Journal on Computer Science and Information Systems*, volumen V, páginas 45-57.

A.3. Generación de Resúmenes Multi-Documento

1. Plaza L, Lloret E, Aker A. 2010. Improving Automatic Image Captioning Using Text Summarization Techniques. En *Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD 2010)*. Lecture Notes in Artificial Intelligence, 6231, páginas 165-172. Praga, República Checa.

2. Aker A, Plaza L, Lloret E, Gaizauskas R. 2010. Towards Automatic Image Description Generation using Multi-document Summarization Techniques. *Multi-source, Multi-lingual Information Extraction and Summarization (MMIES)*. Book Chapter. Springer Book. En proceso de publicación.

A.4. Aplicación de las Etapas de Identificación de Conceptos y Desambiguación Léxica a otras Tareas de Procesamiento del Lenguaje

1. Carrillo de Albornoz J, Plaza L, Gervás P. 2010. Improving Emotional Intensity Classification using Word Sense Disambiguation. En *Journal on Research in Computing Science, 46. Special issue: Natural Language Processing and its Applications*, páginas 131-142.

2. Carrillo de Albornoz J, Plaza L, Gervás P. 2010. A Hybrid Approach to Emotional Sentence Polarity and Intensity Classification. En *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010)*, páginas 153-161. Uppsala, Suecia.

A.5. Aplicación de las Etapas de Identificación y Clustering de Conceptos a otras Tareas de Procesamiento del Lenguaje

1. Carrillo de Albornoz J, Plaza L, Gervás P, Díaz A. 2011. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. En *Proceedings of the 33rd European Conference on Information Retrieval*. En proceso de publicación.

Apéndice B

Documentos Utilizados en los Casos de Estudio

B.1. Caso de Estudio: Artículos Científicos de Biomedicina

A continuación se especifica el documento utilizado como ejemplo en el caso de estudio dedicado a la generación de resúmenes de artículos científicos sobre biomedicina. Dicho documento forma parte del corpus de BioMed Central para investigación en minería de texto, y ha sido publicado en la revista *Current Controlled Trials in Cardiovascular Medicine*.

La versión online, en formato PDF, puede encontrarse en:
`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC64824/pdf/Cvm-2-6-254.pdf`
`(Consultada el 1 de noviembre de 2010)`

También puede descargarse en formato XML desde:
`http://www.biomedcentral.com/content/download/xml/cvm-2-6-254.xml`
`(Consultada el 1 de noviembre de 2010)`

B.2. Caso de Estudio: Noticias Periodísticas

A continuación se reproduce el documento utilizado como ejemplo en el caso de estudio de generación de resúmenes de noticias periodísticas. Dicho documento forma parte del corpus de evaluación elaborado para la conferencia DUC 2002. Puede obtenerse, previa petición, a través de la siguiente página web:
`http://www-nlpir.nist.gov/projects/duc/data.html`
`(Consultada el 1 de noviembre de 2010)`

```
AP880911-0016
AP-NR-09-11-88 0423EDT
r i BC-HurricaneGilbert 09-11 0339
BC-Hurricane Gilbert,0348
```

Hurricane Gilbert Heads Toward Dominican Coast
By RUDDY GONZALEZ
Associated Press Writer
SANTO DOMINGO, Dominican Republic (AP)

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. “There is no need for alarm, ”Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.

Cabral said residents of the province of Barahona should closely follow Gilbert’s movement. An estimated 100.000 people live in the province, including 70.000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a “broad area of cloudiness and heavy weather” rotating around the center of the storm.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m Sunday.

Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico’s south coast. There were no reports of casualties.

San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.

On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 80 mph winds and sheets of rain.

Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

B.3. Caso de Estudio: Páginas Web de Información Turística

A continuación se especifica el conjunto de documentos asociados al objeto *Acrópolis de Atenas* del corpus utilizado en el caso de estudio de generación de resúmenes multi-documento de páginas web de información turística (Aker y Gaizauskas, 2009). Dichos documentos han sido obtenidos automáticamente de las siguientes páginas web (consultadas el 1 de noviembre de 2010):

`http://en.wikipedia.org/wiki/Acropolis`
`http://en.wikipedia.org/wiki/Acropolis_of_Athens`
`http://travel.yahoo.com/p-travelguide-2780725-acropolis_athens-i`
`http://www.stoa.org/athens/sites/acropolis.html`
`http://www.answers.com/topic/acropolis`
`http://www.athensguide.com/acropolis.html`
`http://wikitravel.org/en/Acropolis`
`http://www.athensguide.com/athacrop.html`
`http://www.newworldencyclopedia.org/entry/Acropolis`
`http://www.gamehouse.com/download-games/acropolis`

The Use of Semantic Graphs in Automatic Summarization: Comparative Case Studies in Biomedicine, Journalism and Tourism

PhD Dissertation

Submitted by

Laura Plaza Morales

in partial fulfillment of the requirements for the degree of Doctor of Computer Science

Department of Software Engineering and Artificial Intelligence

Universidad Complutense de Madrid

Madrid, December 2010

Abstract

In recent years, with the increasing publication of online information, providing mechanisms to facilitate finding and presenting textual information has become a critical issue. New technologies, such as high-speed networks and massive storage, are supposed to improve work efficiency by assuring the availability of data everywhere at anytime. However, the exorbitant volume of data available threatens to undermine the convenience of information if no effective access technologies are provided. In this context, automatic text summarization may undoubtedly help to optimize the treatment of electronic documentation and to tailor it to the needs of users.

Automatic summarization is one of the most complex Natural Language Processing (NLP) tasks, and this is due to the number of other tasks that implicitly entails, such as topic detection, word sense disambiguation, anaphoric resolution, acronym expansion, sentence simplification and redundancy detection. In particular, this thesis studies a crucial issue that has been previously unexplored, as is the effect of lexical ambiguity in the knowledge source on semantic approaches to summarization, and demonstrates that using word sense disambiguation techniques leads to an improvement in summarization performance.

A controversial decision when designing a summarization system is whether it should be general (i.e. able to produce summaries for any type of document) or whether it should be changed by text types (i.e. be specific to documents of a given genre and structure). The advantage of the former is obvious, but the latter strategy has proved to be more effective and capable of improving the quality of the summaries. The main contribution of this thesis is the development of a generic summarization method that combines the advantages of both approaches, by taking into account the structure, genre and domain of the document to be summarized, but is easily configurable to work with new types of documents. The method proposed addresses

the problem of identifying salient sentences in a document by representing the text as a semantic graph, using concepts and relations from a knowledge source. This way it gets a richer representation than the one provided by traditional models based on terms. A degree-based clustering algorithm is then used to discover different themes or topics within the text. Different heuristics for sentence selection aiming to generate different types of summaries are tested.

The thesis also presents three case studies, in which the summarizer has been configured and used to generate summaries of texts from different domains and with very distinct structure and style: biomedical scientific articles, news items and tourism-related websites. The system is evaluated using the ROUGE metrics and the legibility criteria followed in the DUC conferences. It has been found that it compares favorably with existing approaches.

This document is a condensed translation from Spanish into English of the PhD dissertation "Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo".

Acronyms and Abbreviations

BE Basic Elements
DUC Document Understanding Conferences
GATE Generic Architecture for Text Engineering
HVS Hub Vertex Set
IHTSDO International Health Terminology Standards Development Organisation
IDF Inverse Document Frequency
LCS Least Common Subsumer
LKB Lexical Knowledge Base
MeSH Medical Subject Headings
NLM National Library of Medicine
NLP Natural Language Processing
PPR Personalized PageRank
PPR-W2W Personalized PageRank Word-to-Word
ROUGE Recall-Oriented Understudy for Gisting Evaluation
RST Rhetorical Structure Theory
SEE Summary Evaluation Environment
SNOMED-CT Systematized Nomenclature of Medicine-Clinical Terms
TF Term Frequency
TAC Text Analysis Conference
TSIN Text Semantic Interaction Network
UMLS Unified Medical Language System
WSD Word Sense Disambiguation

Index

Chapter 1

Introduction

1.1. Motivation

Nowadays, the rapidly increasing rate of new information being produced along with the easy transmission of data across the Internet has caused what is commonly called *information overload*. This situation is characterized by the presence of too much information that obstructs the decision-making process and diminishes work efficiency. It is often said that knowledge is power, but information is only valuable to the extent that it is accessible, it is easily retrieved and it concerns the personal interests of the user. Information overload threatens to undermine the convenience of this information if no easy and effective access technologies are provided.

In order to tackle this information overload, automatic text summarization can undoubtedly play a role. Automatic summaries may be used as substitutes for the original documents, or as a reference function to get a proper idea of what a document is about in just a few lines without having to read it all. Moreover, automatic summaries have been shown to improve other Natural Language Processing (NLP) tasks, such as indexing and categorization (Mani, 2001a). More specifically, according to Borko and Bernier (1975), an abstract or summary may produce the following benefit:

- It promotes current awareness.
- It saves reading time.
- It facilitates information selection.
- It facilitates literature searches.
- It improves indexing efficiency.

1.2. Exploring the Problem

Sparck-Jones (1999) defined a summary as *a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source.* According to this author (Sparck-Jones, 2007), there is no best summary of a document regardless of what summarizing is for, so that the design of any summarization system need to consider three classes of *context factors*: the *input factors* that characterize the properties of the source material, the *purpose* or task that the summaries are intended to serve, and the *output factors* that determine how the summaries have to be presented to the user.

Different context factors determine different types of summaries. According to them, Section 2.1 presents a well-accepted categorization of summaries. However, given the huge number of possible factor combinations, only a few group of cases have been explored so far. In particular, almost all research done so far either focuses on specific document types or intends to cover any kind of document but at the expense of reducing the quality of the summaries. However, genre and structure considerably differ across document types (news articles, scientific papers, web sites, brochures, books, etc.) and are extremely important variables to consider before facing the task, to the extent that the techniques that may prove useful in one case, may be useless in the other.

Beyond all this, automatic text summarization is one of the most complex NLP tasks, and this is due to the large number of other tasks that it implicitly entails:

- First, when elaborating a summary it is necessary to identify the topics or themes that are covered in the text (*Topic Detection*), in order to determine which of these topics are of interest to the readers.

- Second, to identify these topics and determine their salience, it is necessary to solve word ambiguity (*Word Sense Disambiguation*). Selecting the wrong meanings for the words in the document may bias the selection of salient information to sentences containing the wrong concepts, while discarding sentences containing the right ones. Consider, for instance, the two following statements from a survey on the effect of vitamin C on the common cold:

1. *More evidence is needed before the conclusion that ascorbic acid has value in providing protection against the* ***common cold****.*
2. *Vitamin C supplement to the diet may therefore be judged to give a "slight" advantage in reducing* ***cold****.*

Both sentences contain the ambiguous term *cold*. If this term is correctly understood as "common cold" in the first sentence, but it is incorrectly understood as "cold sensation" in the second one, the summarizer would regard both sentences as talking about different topics. Even if the summarizer successes in identifying that the meaning "common cold" is a central topic within the document, the second sentence could not be selected for the summary because it talks about a completely different meaning of *cold*. Moreover, if the wrong mapping for *cold* is selected repeatedly throughout the document, the summarizer may even fail to determine that the concept "common cold" represents a salient topic within it.

- Third, it is highly advisable to apply some kind of anaphora resolution to ensure the coherence of the summaries (*Anaphora and Co-reference Resolution*). Suppose, for example, that we aim to summarize a scientific paper that presents the two following sentences:

  1. *There are many viruses involved in the onset of the* ***common cold****.*
  2. *The treatment of* ***the disease*** *will depend on the virus that causes it.*

  If the system already knows that the information related to the concept "common cold" is salient, in order to determine that the information in the second sentence is also relevant, it should know that "the disease" is just another way of referring to "common cold". Moreover, if the second sentence is selected for the summary but not the first one, the result will be an incoherent summary.

- Forth, it may be necessary to identify acronyms and abbreviations in the document (*Acronym Resolution*) in order to substitute them with their corresponding expanded forms.

- Fifth, and especially if a high compression rate is needed, it may be mandatory to simplify and combine the sentences in order to create more space within which to capture important content (*Sentence Simplification or Compression* and *Sentence Fusion*).

- Finally, if a multi-document summarization task is faced, it is necessary to detect and remove the information that is duplicated across documents (*Redundancy Detection*).

Automatic text summarization is, therefore, a complex task, given the variety of questions to pay attention to when generating a summary. However, to the author's knowledge no previous work has addressed all of these problems. There are systems that performs anaphora resolution as a previous step to sentence selection (Steinberger et al., 2007). There are also a few works that use sentence simplification as a means of reducing the summary length (Barzilay y McKeown, 2005; Filippova y Strube, 2008). Redundancy detection is a common practice in multi-document summarization (Zhao, Wu, y Huang, 2009; Plaza, Lloret, y Aker, 2010). However, the author does not know of any work, except for the one developed as part of this thesis, which analyzes the effect of word disambiguation and acronym expansion in automatic summarization.

1.3. Main Contributions and Goals

Text summarization has been an important subject of study since pioneer works by Luhn (1958) and Edmundson (1969) in the 50s and 60s, and a number of different approaches have been proposed during this time. Summarization systems usually work with a representation of the document consisting of information that can be directly extracted from the document itself, ignoring the benefits of exploiting external knowledge sources to construct richer semantic representations. As a result, these systems usually exhibit important deficiencies which are consequences of not capturing the semantic relationships between terms (synonymy, hypernymy, homonymy and co-occurrence relations). The following sentences illustrate such problems:

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

2. ***Brain vascular disorders*** *during* ***gestation*** *results from any of three major mechanisms: arterial infarction, hemorrhage, or venous thrombosis.*

As both sentences present different terms, word-based approaches are unable to make use of the fact that both sentences have exactly the same meaning. They do not realize, for instance, that the terms *pregnancy* and *gestation* are synonyms. Consider, for example, a traditional approach based on term frequencies that aims to summarize a document where the term *pregnancy* appears with a high frequency. It will probably select for the summary the sentences containing this term, but will consider the sentences containing the synonym *gestation* to be less important. This problem may be solved dealing with concepts instead of terms, and semantic relations instead of lexical ones.

On the other hand, the need to consider the particular characteristics of the regarded domain and type of documents is becoming apparent. As already mentioned, it is not the same summarizing news articles than biomedical scientific literature, since both types of documents differ considerably in the structure and vocabulary used. For example, information in news items usually follows the *pyramid form*, which means that the most important information is placed in the first sentences of the document. In contrast, this property is not found in scientific papers, where the relevant information is spread throughout the document.

Therefore, a main contribution of this work is to show how the use of concepts and relations from an appropriated knowledge source, and the consideration of the structural properties of the documents provide additional knowledge that may benefit the summarization process and the quality of the summaries.

On the other hand, even though recent studies have demonstrated the benefit of summarization based on domain-specific knowledge (Reeve, Han, y Brooks, 2007; Fiszman, Rindesch, y Kilicoglu, 2004), summarization approaches that use domain knowledge and exploit the document structure have the disadvantage of being applicable only to documents of the regarded structure and domain. To overcome this limitation, we propose a generic method for the generation of domain-dependent summaries, but easily configurable to work with new types of documents. Only three requirements need to be met: modifying the knowledge source, using an appropriated method

for automatically identifying concepts within the text, and modeling the specific structure of the documents.

An important deficiency that, in the opinion of the author, presents most summarization systems, is the lack of a pre-processing step aimed at resolving lexical ambiguity. Even the approaches that make use of domain-specific knowledge do not attempt to disambiguate polysemous words. As already mention, this may affect negatively the quality of the automatic summaries. Hence, in this work we pay special attention to this issue, and study the effect of lexical ambiguity in automatic summarization.

Finally, it is worth noting that the main focus of this thesis is mono-document summarization, although the method proposed is designed to allow summarization of multiple documents with minor changes. Even though during the last years researchers have been mainly focused on the development of multi-document systems, the state of research in mono-document summarization is still far from satisfactory, as evidenced by the fact that the use of such systems in real environments is virtually nonexistent. In fact, recent studies have demonstrated that systems are further from human performance in the single document summarization task than in the multi-document task (Nenkova, 2005). This can be partly explained by the fact that repetition across input documents can be used as an indication of importance in multi-document summarization, while such cues are not available in single document summarization. In spite of this,

1.4. Thesis Road Map

This thesis is organized into nine chapters. Each chapter is briefly discussed below.

Chapter 1. Introduction. This chapter provides a global vision of the challenges faced in automatic summarization. It also discusses the main contributions of this work.

Chapter 2. Summary of Previous Work. This chapter presents an overview of a selection of representative systems and approaches to automatic summarization. It also introduces the problem of evaluating automatic summaries and describes the most widely accepted evaluation metrics.

Chapter 3. Tools and Resources. This chapter includes a brief description of the tools and resources used in the development of this thesis.

Chapter 4. Automatic Text Summarization Using Semantic Graphs. This chapter presents a generic graph-based method for knowledge-driven summarization.

Chapter 5. Case Study: Mono-document Summarization of Biomedical Scientific Literature. This chapter presents an application of the summarization method to biomedical scientific articles.

Chapter 6. Case Study: Mono-document Summarization of News. This chapter presents an application of the summarization method to news items.

Chapter 7. Case Study: Multi-document Summarization of Tourism Websites. This chapter presents an application of the summarization method to summarize multiple websites with information about tourist destinations.

Chapter 8. Evaluation. This chapter presents the evaluation accomplished to assess the summarization method in each of the three previous case studies, and compare it to other well-known commercial and research summarizers.

Chapter 9. Conclusion and Future Work. This chapter summarizes the main conclusions derived from this work and provides future lines of work.

Chapter 2

Summary of Previous Work

2.1. The Summary

Following Sparck-Jones (1999), a summary is *a reductive transformation of source text to summary text through content condensation by selection and/or generalization on what is important in the source.* Text summarization is not context-free, but strongly depends on what summarizing is for. That is, the context influences both the summarization process and result. Sparck-Jones (2007) defines three classes of *context factors*: the **input factors** that define properties of the source material (e.g. source material, language, genre, specificity, length or media); the **purpose factors** that characterize the intended situation, use and audience for the generated summary; and the **output factors** that denote the composition, style, format and reduction of the summary.

Ideally, a summarization system should take into consideration all these factors. However, given the huge number of possible factor combinations, only a few group of cases have been explored so far. Depending on the summarization factors taken into account, different taxonomies of textual summaries have been suggested. The most accepted classifications distinguish, at least, the following types of summaries:

- According to the scope, a summary may be restricted to a single document or to a set of documents about the same topic (*mono-document* summary *vs. multi-document* summary).

- According to their purpose, summaries are classified as *indicative*, if the aim is to anticipate for the user the content of the text and to help

him to decide on the relevance of the original document; *informative*, if they aim to substitute the original text by compiling all the new or relevant information; and *critical*, if they incorporate opinions or comments that do not appear in the original text.

- Finally, attending to their focus, we can distinguish between *generic* summaries, if they gather the main topics of the document and are addressed to a wide group of readers, and *user-adapted* summaries, if the summary is constructed according to the interests - i.e. previous knowledge, areas of interest, or information needs - of a particular reader or group of readers.

2.2. Summarization Techniques

Text summarization is the process of automatically creating a compacted version of a given text. Content reduction can be addressed by selection and/or by generalization of what is important in the source (Sparck-Jones, 1999). This definition suggests that two generic groups of summarization methods exist: those which generate extracts and those which generate abstracts. *Extractive summarization* produces summaries by selecting salient sentences from the original document, and therefore the summaries are integrally composed of material that is explicit in the source. In contrast, *abstractive summarization* constructs summaries in which the information from the source has been paraphrased. Although human summaries are typically abstracts, most existing systems produce extracts, which is motivated by the fact that extractive summarization has been demonstrated to report better results than abstractive summarization (Mani y Bloedorn, 1999).

This categorization of automatic summarization approaches (i.e. extractive approaches *vs.* abstractive approaches) is, to the author's opinion, excessively broad, due to the variety of techniques used within the extractive approaches. Consequently, it seems more appropriated to categorize the approaches according to the depth of the analysis performed on the text. Therefore, according to this criterion, summarization techniques may be classified into the following three families: surface techniques, discourse-level techniques and abstractive techniques.

- *Surface* or *shallow* techniques simply handle text as strings of symbols and typically construct summaries based on a superficial analysis of

the source. They are based on what Mani (2001a) called the *Edmundsonian paradigm* (Edmundson, 1969). In this paradigm, sentences are ranked using simple heuristic features, such as the position of the sentences in the document (Baxendale, 1958; Kupiec, Pedersen, y Chen, 1995; Teufel y Moens, 1997; Bawakid y Oussalah, 2008; Bossard, Généreux, y Poibeau, 2008), the frequency of their terms (Luhn, 1958; Edmundson, 1969; Kupiec, Pedersen, y Chen, 1995; Teufel y Moens, 1997; Hovy y Lin, 1999; Steinberger et al., 2007; Lloret et al., 2008), the presence of certain cue words and indicative phrases (Edmundson, 1969; Rush, Zamora, y Salvador, 1971; Aker y Gaizauskas, 2010), or the word overlap between the sentences and the document title and headings (Edmundson, 1969; Bawakid y Oussalah, 2008). These attributes are usually combined using a linear weighting function that assigns a single score to each sentence in the document, and the highest scoring sentences are extracted for the summary. More recent approaches also employ machine learning techniques to determine the best subset of features for extraction (Kupiec, Pedersen, y Chen, 1995; Lin y Hovy, 1997; Aone et al., 1999; Chuang y Yang, 2000; Zhou, Ticrea, y Hovy, 2004; Metzler y Kanungo, 2008; Binwahlan, Salim, y Suanmali, 2009).

Shallow summarization approaches present three main advantages: they are robust, simple and more general with regard to the application domain. However, they also present two important problems: the resulting summaries often suffer from lack of *cohesion* and *coherence*.

- A second family of techniques is that focused on *discourse* analysis, and it is based on the idea that a text is more than just the set of sentences which comprise it. Discourse-level techniques are often divided into those that analyze *text cohesion* and those that analyze *text coherence* (1985). Following Halliday and Hasan (1996), text cohesion involves relations between words, word senses or referring expressions, which determine how tightly connected the text is. It includes "gramatical cohesion", involving linguistic relations such as anaphora, ellipsis and conjunctions; and "lexical cohesion", which involves relations such as reiteration, synonymy, and hypernymy (Mani, 2001a). Text coherence, on the other hand, represents the overall structure of a text in

terms of macro-level relations between clauses or sentences. These relations determine the overall argumentative structure of the text (Mani, 2001a). The kind of discourse structure that emerges from text cohesion has to do with patterns of salience in text, while text coherence is related to the notion of *theme* and the patterns of reasoning expressed in the text. Discourse-level summarization approaches exploit the discourse structure that emerges from cohesion and coherence in order to discover patterns of salience and themes in the text.

Among the approaches that analyze the cohesion structure for summarization it is worth mentioning the work of Morris and Hirst (1991), where the notion of *lexical chain* is first defined. Morris and Hirst characterize a lexical chain as *a sequence of related words spanning a topical unit of text*. They argue that lexical chains may be useful in identifying topical segments. Barzilay and Elhadad (1997) implement the algorithm for computing such chains, using the results in summarization. They examine relationships of repetition, synonymy, hypernymy, antonymy and holonymy. Recently, several authors have adapted the original concept of lexical chains to work with highly specialized texts. For example, in the biomedical domain Reeve et al. (2007) adapt the lexical chaining approach to use UMLS concepts rather than terms, and apply it to single document summarization. They automatically identify UMLS concepts in the source and chain them so that each chain contains a list of concepts belonging to the same UMLS semantic type. The concept chains are then scored by multiplying the frequency of the most frequent concept in the chain by the number of distinct concepts in it, and these scores are used to identify the strongest concept chains. Finally, the sentences are scored based on the number of concepts from strong chains that they contain.

Concerning text coherence, a number of different theories have been proposed to provide an analysis of the argumentation structure, including the *Rhetorical Structure Theory, RST* (Mann y Thompson, 1988), the *Discourse Grammar* (Longacre, 1979), the *Macrostructure* (Van Dijk, 1988) and the *Coherence Relations* (Hobbs, 1985). Undoubtedly, the most outstanding is the Rhetorical Structure Theory. A fundamental point made by this theory is the concept of rhetoric relation to represent an asymmetric relation holding between two text

segments, one called *nucleus* and the other *satellite*. The nucleus contains information that is central within the text, while the satellite provides secondary information that complements the nucleus. Example of these relations are *circumstance*, *motivation*, *purpose* or *solution*. Marcu (1999; 2000) applies this theory to the automatic summarization task, by constructing a RST-tree for a text based on identifying cue phrases. He after computes the salience of the different terms represented by the vertices in the RST-tree to produce summaries at different levels of detail.

An interesting subgroup within the discourse-level approaches concerns the use of graphs. *Graph-based* methods usually represent the documents as graphs, where the nodes correspond to text units such as words, phrases, sentences or even paragraphs, and the edges represent cohesion relationships between these units or even similarity measures between them. Once the document graph is created, salient nodes in the graph are discovered and used to extract the corresponding units for the summary. LexRank (Erkan y Radev, 2004) is the best-known example of graph-based method for multi-document summarization. It assumes a fully connected and undirected graph for the set of documents to be summarized in which each node corresponds to a sentence represented by its TF-IDF vector, and the edges are labeled with the cosine similarity between the sentences. Only the edges connecting sentences with a similarity upper a predefined threshold are drawn in the graph. The sentences represented by the most highly connected nodes are selected for the summary. A very similar method, TextRank, is proposed by Mihalcea and Tarau (2004). TextRank differs from LexRank in three main aspects: first, it is intended for single document summarization; second, the similarity between sentences (i.e. the weight of the edges in the document graph) is measured as a function of their content overlap; and third, the PageRank algorithm (Brin y Page, 1998) is used to rank the nodes in the document graph. Most recently, Litvak and Last (2008) proposed a novel approach that makes use of a graph-based syntactic representation of textual documents for keyword extraction, which can be used as a first step in single document summarization. They represent the document as a directed graph, where the nodes represent single words found in the

text, and the edges (not labeled) represent precedence relations between words. The HITS algorithm (Kleinberg, 1999) is then run on the document graph under the assumption that the top-ranked nodes should represent the document keywords. In the biomedical domain, Yoo and colleagues (2007) use the Medical Subject Headings (MeSH) to represent a corpus of documents as a graph, where the nodes are the MeSH descriptors found in the corpus and the edges represent hypernymy and co-occurrence relations between them. The concepts are clustered to identify groups of documents dealing with the same topic using a degree-ranking method called Scale Free Graph Clustering (SFGC). Each cluster of documents is then used to produce a single summary by constructing Text Semantic Interaction Networks (TSIN) using only the semantic relations found in the document cluster. BioSquash (Shi et al., 2007) is a question-oriented extractive system for biomedical multi-document summarization. It constructs a graph that contains concepts of three types: ontological concepts (general ones from WordNet and specific ones from the UMLS), named entities and noun phrases. The edges of this graph represent semantic relationships between concepts, but nothing is said about the specific relationships that have been used.

- Finally, *abstractive* approaches involve making inferences about the content of the text, so that they can infer concepts that are not mentioned explicitly in the text. The abstraction process entails, in general, the identification of the most prevalent concepts in the source, the appropriate semantic representation of them, a minimum level of inference and the rewriting of the summary through Natural Language Generation techniques. Abstractive approaches were widely studied in the 80's and 90's, but have been given little attention in recent years.

  A variety of different approaches have been developed that may be broadly categorized in two groups. The first group uses *templates* to predefine the types of information that are likely to be included in the summary, and uses extraction methods to locate in the text the information that fills the corresponding slots of the template. Examples of template-based systems are FRUMP (DeJong, 1982), which defines templates for 50 different types of news articles, and the one presented in (Paice y Jones, 1993), which uses templates for summarizing

technical papers in the filed of crop agriculture.

The second group is abstraction by *term rewriting*. The approaches within this category perform various selection, aggregation and generalization operations in order to rewrite the summary. The SUSY summarizer (Fum, Gmda, y Tasso, 1985), for instance, produces summaries of technical articles on computer operating systems, using a small knowledge base of about 30 domain concepts to construct a sentential semantic where each sentence is represented as a list of logical terms, and then computes its salience based on a set of rules.

The main advantage of abstractive summarization is that it makes possible summarization at a much higher degree of compression. Its greatest disadvantage is that it is only feasible to work in very narrow domains.

2.3. Summarization Evaluation

Even though the evaluation of the automatically generated summaries is a critical issue, there is still controversy about what the evaluation criteria should be, mainly due to the necessary subjectivity of deciding whether or not a summary is of good quality (Radev et al., 2003).

Methods for summarization evaluation can be broadly divided in two categories, *intrinsic* and *extrinsic*, depending on whether or not the outcome is evaluated independently of the purpose that the summary is supposed to serve. As the method proposed in this work is not conceived for any specific task, the interest here is on intrinsic evaluation.

Intrinsic evaluation techniques tests the summarization itself, and measures mainly two desirable properties of the summary: *informativeness* and *readability*. Summary informativeness aims at measuring the summary's information content, while readability normally refers to text coherence, clarity and cohesion (Mani, 2001b).

A good number of metrics have been proposed to measure both informativeness and readability. Table 2.1 shows the most commonly used. Even though some informative evaluation metrics may be computed automatically, the measurement of readability requires a qualitative evaluation along with the participation of human subjects to deliver judgments.

Informativeness Evaluation
Precision and recall (Salton y McGill, 1983)
Utility Index (Radev, Jing, y Budzikowska, 2000)
Content Similarity (Donaway, Drummey, y Mather, 2000)
SEE (Summary Evaluation Environment) (Lin y Hovy, 2002)
BE (Basic Elements) (Hovy, Lin, y Zhou, 2005)
Pyramid (Passonneau et al., 2005; Nenkova, Passonneau, y McKeown, 2007)
Readability Evaluation
SEE (Summary Evaluation Environment) (Lin y Hovy, 2002)
DUC linguistic criteria (grammaticality, redundancy, clarity, focus and coherence) (Dang, 2005; Dang, 2006)

Table 2.1: Summarization evaluation metrics

2.3.1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) automatic summarization evaluation package (Lin, 2004b) is the most widely used tool for intrinsic informativeness evaluation, and has become the *de facto* standard since it was chosen as the official automatic evaluation package for the Document Understanding Conferences (DUC)[1] and the Text Analysis Conferences (TAC)[2] from 2004 onwards. ROUGE compares a generated summary from an automated system (called *peer*) with one or more ideal summaries (called *models*), usually created by humans, and computes a set of different measures to automatically determine the quality of the summary. The most widely used metrics are listed below:

- **ROUGE-N** evaluates n-grams occurrence, where N stands for the length of the n-gram.
- **ROUGE-L** computes the union of the longest common subsequences between the candidate and the model summary sentences.
- **ROUGE-W** is a refined version of ROUGE-L to take into account the presence of consecutive matches.
- **ROUGE-SN** evaluates "skip bigrams", that is, pairs of words having intervening word gaps no larger than N words.

[1] Document Understanding Conferences (DUC). http://duc.nist.gov/. Last accessed: 1 November 2010

[2] Text Analysis Conferences (TAC). http://www.nist.gov/tac/. Last accessed: 1 November 2010

Lin and Hovy (2003) demonstrated that the ROUGE metrics (specially ROUGE-2 and ROUGE-S4) show a high correlation with human judges, and that it is also possible to apply the methodology in a completely automatic way.

2.3.2. Readability Assessment: The DUC Criteria

The DUC and TAC conferences manually assess the quality of the automatically generated summaries by asking human judges to evaluate the summaries within a 5-point qualitative scale (1=Very poor; 2=Poor; 3=Barely acceptable; 4=Good; 5=Very good) according to the following criteria (Dang, 2005):

- **Grammaticality:** The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
- **Non-redundancy:** There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.
- **Referential clarity:** It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.
- **Focus:** The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- **Structure and Coherence:** The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

2.4. New Trends in Automatic Summarization

Automatic summarization, as it was originally intended, has gradually evolved and given rise to a broad spectrum of unconventional tasks and applications. This section briefly discusses those that, in the author's opinion, will have a role to play in future research on the field.

2.4.1. Multi-document Summarization

As stated in (Mani, 2001a) multi-document summarization is the extension of single-document summarization to collections of related documents. This technology is becoming increasingly important, since the continuing growth of online information demands improved mechanism to find and present textual information. Given a set of documents on the same topic, ideally multi-document summaries should contain the key information that is shared among the documents only once, plus other relevant information unique to some of the individual documents.

Although the techniques used in multi-document summarization are nearly the same than those used for mono-document summarization, there are three main differences among both tasks:

- Since there may be a lot of similar information repeated across documents, a mechanism for removing *redundant* information is needed.
- The compression rate will typically be much smaller than for mono-document summarization, so that the task becomes more difficult.
- Ordering the sentences in the final summary is problematic, since they come from different documents.

2.4.2. Query-Based and User-Oriented Summarization

Empirical studies have demonstrated that, when two persons are asked to summarize the same document, the information they include in the summaries highly depends on their previous knowledge and experiences, their interests and their information needs (Paice y Jones, 1993). Based on this observation, summarization systems may benefit from user queries that state what information they want to read in the summary. In other words, the output of the summarization process is adapted to suit the user's declared

information need (i.e. the query). This branch of summarization has been called *query-based* summarization.

Examples of summarization systems that make use of user queries to improve the quality of the summaries may be found in Carbonell et al. (1997), Sanderson (1998), Amini and Gallinari (2002) and Zhao et al. (2009).

2.4.3. Multimedia Summarization

A multimedia summarization is an application area where the summarizer's input and/or output consists of different media types instead of just text (e.g. video, speech, images, etc.) (Mani, 2001a). There exist systems able to generate summaries of dialogues (Xie et al., 2009; Hsueh y Moore, 2009), where the output of an automatic speech recognizer is used as the input to the summarization system. Other works address the problem of video summarization, aiming to produce summaries of broadcast news, films, TV programs or video games. Finally, it is worth mentioning the task of summarizing diagrams, figures and tables (Futrelle, 1999; Gao et al., 2009; Hong et al., 2009). These systems use selection, aggregation and fusion techniques on the text in the diagram, in the caption, as well as in the running text, to compute the salience of the information within it. However, research in this field is still highly preliminary.

Chapter 3

Tools and Resources

The purpose of this chapter is to present the various tools and linguistic resources that, not having been developed as part of this thesis, have been used in it.

3.1. UMLS, MetamorphoSys and MetaMap

3.1.1. The Unified Medical Language System

The Unified Medical Language System (UMLS)[1] is a collection of controlled vocabularies related to biomedicine that contains a wide range of information that can be used for Natural Language Processing (NLP). The UMLS consists of three main components: the Specialist Lexicon, the Metathesaurus and the Semantic Network.

- The **UMLS Specialist Lexicon**[2] is a database of lexicographic information specially conceived for NLP systems to address the high degree of variability in natural language words and terms. The lexicon entry for each word records syntactic, morphological and orthographic information. For example, Table 3.1 shows the information related to the entry *anaesthetic* in the Specialist Lexicon.

- The **UMLS Metathesaurus**[3] comprises a collection of biomedical and health related concepts derived from more than 100 different voca-

[1] National Library of Medicine (NLM). UMLS. http://www.nlm.nih.gov/research/umls. Last accessed: 1 November 2010

[2] National Library of Medicine (NLM). UMLS Specialist Lexicon fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. Last accessed: 1 November 2010

[3] National Library of Medicine (NLM). UMLS Metathesaurus fact sheet.

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

Table 3.1: Entry *anaesthetic* in the Specialist Lexicon

bulary sources, their various names, and the relationships among them. It is organized around concepts, each of which represents a meaning and is assigned a Concept Unique Identifier (CUI). The Metathesaurus comprises of several tables containing information about CUIs. These include the *MRCONSO*, *MRREL* and *MRHIER* tables. The concepts, their names and their source vocabularies are stored in the *MRCONSO* table. Table 3.2, for instance, shows the information associated to the concept *C0001175:AIDS* in the *MRCONSO* table.

CUI=C0001175	**AUI**=A2878223
language=ENG	**source**=SNOMEDCT
status=S	**string_type_source**=PT
LUI=L0001842	**code**=62479008
string_type=PF	**string**=AIDS
SUI=S0011877	**restriction_level**=4
preference=N	**suppress**=N

Table 3.2: Entry *AIDS* in the MRCONSO Metathesaurus table

The *MRREL* table lists relations between CUIs found in the various sources that are used to form the Metathesaurus. This table lists a range of different types of relations, including “child”, “parent”, “can be qualified by”, “related and possibly synonymous” and “other related”. For example, the *MRREL* table states that the concepts *C0009443:Common Cold* and *C0027442:Nasopharynx* are connected via the “other related” relation. The *MRHIER* table in the Metathesaurus lists the hierarchies in which each CUI appears, and lists the

http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. Last accessed: 1 November 2010

entire path to the root of each hierarchy for the CUI.

- The **UMLS Semantic Network**[4] consists of a set of categories (or *semantic types*) that provide a consistent categorization of the concepts in the Metathesaurus, along with a set of relationships (or *semantic relations*) that exist between the semantic types. For example, the concept *C0009443:Common Cold* is classified in the semantic type "Disease or Syndrome". The *SRSTR* table in the Semantic Network describes the structure of the network. This table lists a range of different relations between semantic types, including hierarchical relations ("is a") and non hierarchical relations (e.g. "result of", "associated with" and "co-occurs with"). For example, the semantic types "Disease or Syndrome" and "Pathologic Function" are connected through the "is a" relation in this table.

Using the UMLS for NLP tasks instead of another biomedical knowledge source (e.g. SNOMED-CT[5] or MeSH[6]) presents two main advantages: (1) a broader coverage, since it is a compendium of vocabularies including SNOMED-CT and MeSH, and (2) it is supported by a number of resources that aid developers of NLP applications, such as lexical tools, concept annotators and word sense disambiguation algorithms. Moreover, using the UMLS for concept annotation presents two further advantages: (1) it lists more than 15000 entries of ambiguous terms, and (2) it contains numerous entries for elisions and abbreviations.

3.1.2. MetamorphoSys

MetamorphoSys is a tool for configuring and personalizing the UMLS Metathesaurus. It is included in the standard UMLS distribution, and allows users to create subsets of the data in the Metathesaurus, excluding certain vocabulary sources or modifying the output format. The result of running MetamorphoSys on the Metathesaurus is a set of *ORF (Original Release Format)* or *RRF (Rich Release Format)* files containing the data subsets,

[4]National Library of Medicine (NLM). UMLS Semantic Network fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html. Last accessed: 1 November 2010

[5]International Health Terminology Standards Development Organisation (IHTSDO). SNOMED-CT. http://www.ihtsdo.org/snomed-ct/. Last accessed: 1 November 2010

[6]National Library of Medicine. Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/. Last accessed: 1 November 2010

along with a loading script for *Oracle* or *MySql* databases. Figure 3.1 shows MetamorphoSys user's interface for selecting UMLS vocabularies.

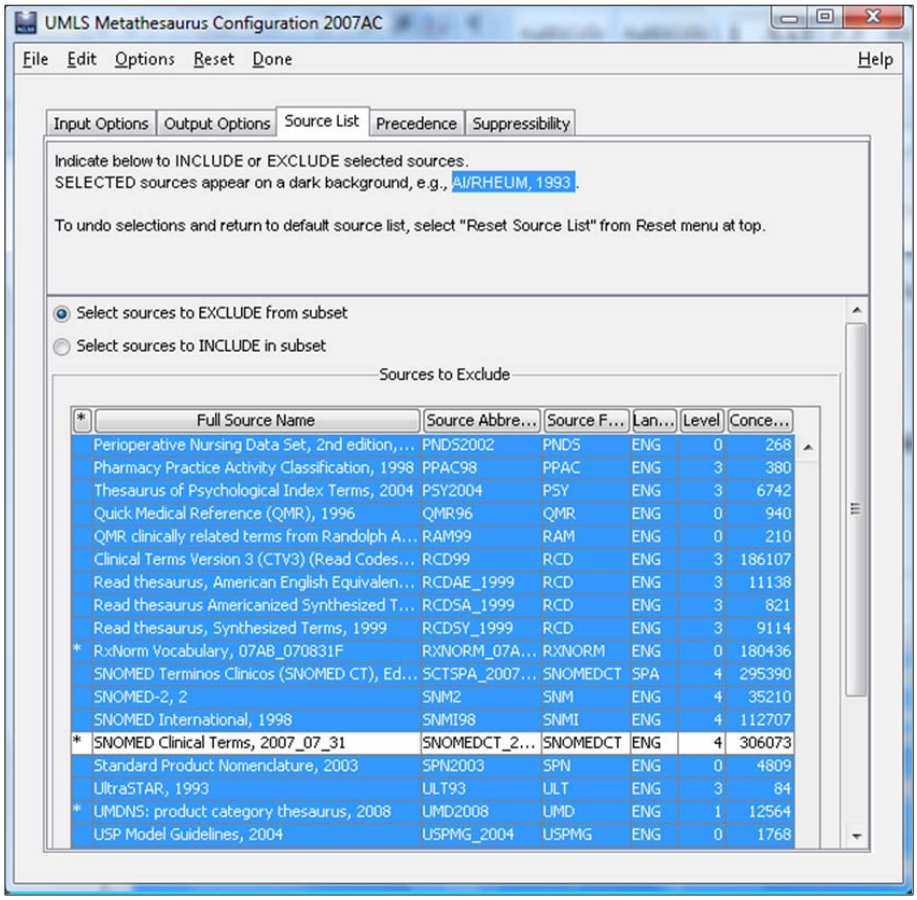


Figure 3.1: Selecting UMLS vocabulary sources using MetamorphoSys

3.1.3. MetaMap

The MetaMap program[7] maps biomedical text to concepts in the UMLS Metathesaurus. MetaMap employs a knowledge intensive approach that uses the Specialist Lexicon in combination with lexical and syntactic analysis to identify noun phrases in text. Matches between noun phrases and Metathesaurus concepts are computed by generating lexical variations and allowing partial matches between the phrase and concept. The possible UMLS concepts are assigned scores based on the closeness of the match between the

[7]National Library of Medicine (NLM). MetaMap. http://mmtx.nlm.nih.gov/. Last accessed: 1 November 2010

input noun phrase and the target concept. Figure 5.1 shows this mapping for the phrase "heart attack trial". The first section in the MetaMap response (Meta Candidates) shows the candidate concepts, while the second section (Meta Mapping) shows the highest scoring candidates. Each candidate is represented by its MetaMap score, its concept name and its semantic type.

Phrase: "Heart Attack Trial"
Meta Candidates (8):
827 C0008976:Trial (Clinical Trial) [Research Activity]
734 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
660 C0018787:Heart [Body Part, Organ, or Organ Component]
660 C0277793:Attack, NOS (Onset of illness) [Finding]
660 C0699795:Attack (Attack device) [Medical Device]
660 C1261512:attack (Attack behavior) [Social Behavior]
660 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
660 C1304680:Attack (Observation of attack) [Finding]
Meta Mapping (901):
734 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
827 C0008976:Trial (Clinical Trials) [Research Activity]

Figure 3.2: An example of MetaMap mapping for the syntagm *heart attack trial*. Each candidate mapping is given a score and is represented by its name in the Metathesaurus (in parentheses) and its semantic type in the Semantic Network (in brackets)

3.1.4. Using UMLS and MetaMap in Natural Language Processing Tasks

The UMLS and MetaMap have been used in a number of biomedical NLP applications, including machine translation (Eck, Vogel, y Waibel, 2004), question answering (Overby, Tarczy-Hornoch, y Demner-Fushman, 2009) and information retrieval (Aronson y Rindflesch, 1997; Plaza y Díaz, 2010). Erk et al. (2004), for instance, modify a simple statistical machine translation system to use information from the UMLS concepts and semantic types achieving significant improvement in translation performance. Overby et al. (2009) show that both the UMLS Metathesaurus and the MetaMap program are useful for extracting answers to translational research questions from biomedical text in the field of genomic medicine. Aronson and Rindflesch (1997) use MetaMap for expanding queries with UMLS Metathesaurus concepts. The authors conclude that query expansion based on UMLS improves retrieval performance and compares favorably with retrieval feedback. Plaza and Díaz (2010) propose a method for the retrieval of similar clinical cases based on mapping the text in electronic health records onto UMLS concepts

and representing the patient records as a set of semantic graphs. Each of these graphs corresponds to a different category of information (e.g. diseases, symptoms and signs, or medicaments) automatically derived from the UMLS semantic types to which the concepts in the records belong.

3.2. WordNet, WordNet::Similarity and WordNet::SenseRelate

3.2.1. WordNet

WordNet (Miller et al., 1990; Miller et al., 1998) is an electronic lexical database for English developed at Princeton University. WordNet structures lexical information in terms of word meanings. Words of the same syntactic category that can be used to express the same concept are grouped into a single synonym set, called *synset*. Each synset has a unique identifier and a *gloss* that defines the synset. Table 3.3 shows the glosses for the noun *pretty*.

(20) pretty – (pleasing by delicacy or grace; not imposing; "pretty girl"; "pretty song"; "pretty room") **(3) pretty** – ((used ironically) unexpectedly bad; "a pretty mess"; "a pretty kettle of fish")

Table 3.3: WordNet glosses for *pretty*

Most synsets are connected to other synsets via a number of semantic relations. These relations vary with the type of word, and include among others:

- **Synonymy:** The WordNet's basic relation. It is a symmetric relation between two words that share at least one sense in common (e.g. *dog* and *canine* are synonyms).
- **Antonymy:** It is also a symmetric relation between word forms that share an opposite meaning (e.g. *beautiful* and *ugly* are antonyms).
- **Hypernyms and Hyponyms:** Y is a hypernym of X if every X is a Y (e.g. *feline* is a hypernym of *cat*). Y is a hyponym of X if every Y is a X (e.g. *cat* is a hyponym of *feline*).

- **Holonym and Meronym:** Y is a holonym of X if X is a part of Y (e.g. *vehicle* is a holonym of *wheel*). Y is a meronym of X if Y is a part of X (e.g. *wheel* is a meronym of *vehicle*).
- **Troponym:** the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (e.g. *to lisp* is a troponym of *to talk*).
- **Entailment:** the verb Y is entailed by X if by doing X you must be doing Y (e.g. *to sleep* is entailed by *to snore*).

3.2.2. WordNet::Similarity

The WordNet::Similarity package[8] (Pedersen, Patwardhan, y Michelizzi, 2004) is a Perl module that implements a variety of semantic similarity and relatedness measures based on the information found in the WordNet lexical database.

In particular, WordNet::Similarity includes six similarity measures and three measures of relatedness. Three similarity measures are based on path lengths between concepts: **lch** (Leacock y Chodorow, 1998), which finds the shortest path between two concepts; **wup** (Wu y Palmer, 1994), which finds the path length to the root node from the least common subsumer (the most specific ancestor) of the two concepts; and **path**, which is equal to the inverse of the shortest path length between two concepts.

The three remaining similarity measures are based on information content, and include **res** (Resnik, 1995), which measures the information content of the *Least Common Subsumer (LCS)* of two concepts; and the **lin** (Lin, 1998) and **jcn** (Jiang y Conrath, 1997) measures, which augment the information content of the LCS with the sum of the information content of the individual concepts.

Finally, the three measures of relatedness include **hso** (Hirst y St Onge, 1998), which classifies relations in WordNet as having direction and tries to find a path between two concepts that is neither too long nor that changes direction too often; **lesk** (Lesk, 1986), which assigns relatedness by finding and scoring overlaps between the glosses of the two concepts; and **vector** (Patwardhan, 2003), which uses information about the co-occurrence of

[8] WordNet::Similarity. http://search.cpan.org/dist/WordNet-Similarity/. Last accessed: 1 November 2010

terms in a corpus made up of WordNet glosses to measure the relatedness between concepts.

Figure 3.5 shows the result of running WordNet::Similarity. The first command requests the *vector* similarity between the first noun senses of *car* and *bike*, while the second command returns this similarity between all posible noun senses of *car* and *bike*.

```
> similarity.pl --type WordNet::Similarity::vector car#n#1 bike#n#1
car#n#1 bike#n#1 0.756395613030899
> similarity.pl --type WordNet::Similarity::vector -allsenses car#n bike#n
car#n#1 bike#n#2 0.90618485736825
car#n#2 bike#n#2 0.877783971247943
car#n#1 bike#n#1 0.756395613030899
car#n#2 bike#n#1 0.701289428083391
car#n#5 bike#n#2 0.621345209159249
car#n#5 bike#n#1 0.50292923868496
car#n#4 bike#n#2 0.460159722727335
car#n#4 bike#n#1 0.407797080123658
car#n#3 bike#n#2 0.395296214612913
car#n#3 bike#n#1 0.342766679044078
```

Table 3.4: WordNet::Similarity running example

3.2.3. WordNet::SenseRelate

WordNet::SenseRelate[9] is a Perl package that performs Word Sense Disambiguation (WSD) by measuring the semantic similarity between a word and its neighbors using the different similarity and relatedness metrics implemented in the WordNet::Similarity package (see Section 3.2.2). The "all words" version assigns a meaning (as found in WordNet) to each word in a text.

Figure 3.5 shows the result of running WordNet::SenseRelate having the sentence *The red car is parked near the supermarket* as input and the lesk algorithm as similarity measure. The output is a list of the concepts discovered in the text, along with their grammatical role and their sense or meaning in WordNet.

```
> wsd.pl --type WordNet::Similarity::lesk
--context The red car is parked near the supermarket
--format tagged --stoplist config/SRStopWord.txt
The red#n#4 car#n#1 be#v#1 parked#a#1 near#a#2 the supermarket#n#1
```

Table 3.5: WordNet::SenseRelate running example

[9] WordNet::SenseRelate. http://www.d.umn.edu/~tpederse/senserelate.html. Last accessed: 1 November 2010

3.3. GATE

GATE (Generic Architecture for Text Engineering)[10] is a well-known framework for developing software components for text processing applications. The core functions include parsing, information retrieval tools, information extraction components, and many more. The set of resources included in GATE is known as CREOLE (a Collection of REusable Objects for Language Engineering). Users may also develop their own resources that can be embebed in the GATE framework to build different NLP applications.

GATE is distributed with an information extraction system called *ANNIE* (*A Nearly-New Information Extraction System*). ANNIE incorporates a wide range of resources to solve language analysis tasks at different levels. For our purpose, the most relevant components (i.e. those which have been used in this work) are listed below:

- The **Tokenizer** splits the text into simple tokens, such as words, numbers and punctuation marks.
- The **Gazetteer** identifies entity names in the text based on lists. Each list represents a name category such as countries, organizations, months, etc., and also includes common abbreviations and acronyms for such names.
- The **Sentence Splitter** consists of a set of finite state transducers which segment the text into sentences.
- The **Part of Speech Tagger** is a modified version of the Brill tagger which tags each word or symbol with its part of speech.

3.4. Word Sense Disambiguation Using Personalized PageRank

The aim of Word Sense Disambiguation (WSD) is to resolve lexical ambiguities by identifying the correct meaning of a word based on its context (Navigli, 2009), and it is regarded as an important stage in text processing (Ide y Véronis, 1998; Agirre y Edmonds, 2006). The most popular approaches to WSD are based on supervised learning (e.g. McInnes et al. (2007),

[10]GATE. http://gate.ac.uk/. Last accessed: 1 November 2010

Stevenson et al. (2008) and Savova et al. (2008)), since previous evaluations (Mihalcea y Tarau, 2004; Pradhan et al., 2007) have suggested that supervised approaches perform better than unsupervised ones. However, they require labeled examples which are often unavailable and can be impractical to create.

Knowledge-based approaches to WSD are an alternative to supervised learning that do not require manually-tagged data and have recently been shown to compete with supervised systems in terms of performance (Ponzetto y Navigli, 2010). Graph-based methods are now widely used for knowledge-based WSD (Sinha y Mihalcea, 2007; Navigli y Lapata, 2007; Tsatsaronis, Vazirgiannis, y Androutsopoulos, 2007; Agirre y Soroa, 2009). These methods represent the knowledge base as a graph which is then analyzed to identify the meanings of ambiguous words. An advantage of this approach is that the entire knowledge base can be used during the disambiguation process by propagating information through the graph.

One such method is Personalized PageRank[11] (Agirre y Soroa, 2009) which makes use of the PageRank algorithm used by internet search engines (Brin y Page, 1998). PageRank assigns weight to each node in a graph by analyzing its structure and prefers ones that are linked to by other nodes that are highly weighted. Agirre and Soroa (2009) use WordNet as the lexical knowledge base and create graphs using the entire WordNet hierarchy. The ambiguous words in the document are added as nodes to this graph and directed links created from them to each of their possible meanings. These nodes are assigned weight in the graph and the PageRank algorithm is applied to distribute this information through the graph. The meaning of each word with the highest weight is chosen.We refer to this approach as *PPR*. It is efficient since it allows all ambiguous words in a document to be disambiguated simultaneously using the whole lexical knowledge base, but can be misled when two of the possible senses for an ambiguous word are related to each other in WordNet since the PageRank algorithm assigns weight to these senses rather than transferring it to related words. Agirre and Soroa (2009) also describe a variant of the approach, referred to as "word to word" (*PPR-w2w*), in which a separate graph is created for each ambiguous word. In these graphs no weight is assigned to the word being disambiguated so

[11] UKB: Graph Based Word Sense Disambiguation and Similarity.
http://ixa2.si.ehu.es/ukb/. Last accessed: 1 November 2010

that all of the information used to assign weights to the possible senses of the word is obtained from the other words in the document. The PPR-w2w is more accurate but less efficient due to the number of graphs that have to be created and analyzed. Agirre and Soroa (2009) show that the Personalized PageRank approach performs well in comparison to other knowledge-based approaches to WSD and report an accuracy of around 58 % on standard evaluation data sets.

Chapter 4

Automatic Text Summarization Using Semantic Graphs

This section presents a generic graph-based method for knowledge-driven summarization. The method presents the document as a semantic graph using concepts and relations from a knowledge base, and a degree-based clustering algorithm is used to discover different themes or topics within the text. The selection of sentences for the summary is based on the presence in them of the most representative concepts for each topic. The method may be easily configured to work with documents with different structure and from different domains. Only three requirements need to be met: modifying the knowledge base (KB), using an appropriated method for automatically identifying concepts within the text, and modeling the specific structure of the new type of documents.

The system accomplishes the task of identifying the n most relevant sentences in a document through seven steps: (1) document pre-processing, (2) concept recognition, (3) sentence representation, (4) document representation, (5) concept clustering and subtheme recognition, (6) sentence-to-cluster assignment and (7) sentence selection. Figure 4.1 illustrates the summarization architecture. Each step is discussed in detail in the following sections.

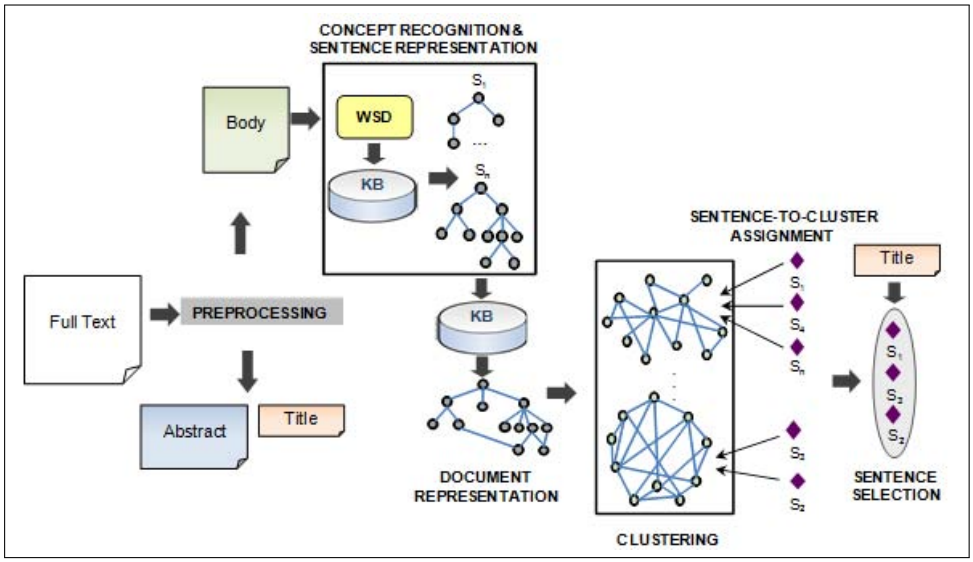


Figure 4.1: Generic summarizer architecture. The figure shows the seven steps involved in the algorithm: (1) Preprocessing, (2) Concept recognition, (3) Sentence representation, (4) Document representation, (5) Concept clustering, (6) Sentence-to-cluster assignment and (7) Sentence selection

4.1. Step I: Document Pre-processing

A preliminary step is undertaken in order to prepare the document for the subsequent steps. This pre-processing involves the following actions:

1. First, sections of the document that are considered irrelevant for including in the summary are removed.

2. Second, if the document presents a title and/or abstract section, the title, abstract and body sections are separated.

3. Third, generic terms (e.g. prepositions and pronouns) are removed, since they are not useful in discriminating between relevant and irrelevant sentences, using a stop list.

4. Finally, the text in the body section is split into sentences using the Tokenizer, Part of Speech Tagger and Sentence Splitter modules of GATE (see Section 3.3).

The preprocessing step can be easily configured to deal with documents of different structures, as well as with unstructured documents. A *config.xml* file allows users to specify, for instance, if the document is not structured and

so the entire text should be considered for the purpose of summarization, the document sections that have to be ignored, the XML tags (if any) that enclose the title, abstracts and body sections, the stop list to be used, etc.

4.2. Step II: Concept Recognition

The next stage is to map the text in the document to concepts from the knowledge source. To this end, it is needed:

1. An appropriated domain knowledge base (KB) that covers the vocabulary in the input documents.

2. A word sense disambiguation (WSD) algorithm capable of selecting the correct concept mapping for each term or phrase in the input text, according to the context in which the term or phrase appears (Navigli, 2009). This algorithm must be chosen having in mind the characteristics and domain of the document to be summarized.

Each sentence in the input text is then analyzed by the WSD algorithm, and the result is a list of concepts from the knowledge base that are found in the text. This process is illustrated in Figure 4.2.

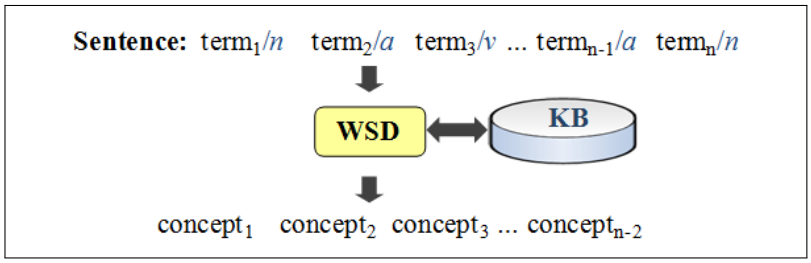


Figure 4.2: Sentence-to-concepts translation

Finally, depending on the regarded domain and the information provided by the knowledge base, it might be interesting to ignore certain concepts that due to their meaning or grammatical role, are found to be excessively broad and useless for determining the salience of the sentences.

4.3. Step III: Sentence Representation

The next step consists of building a graph representation for each sentence in the document. To this end, the domain concepts identified in the previous

step are expanded with their complete hierarchy of hypernyms from the knowledge base (*is a* relations). All the hierarchies for each sentence are merged creating a *sentence graph* where the edges (temporally unlabeled) represent semantic relations, and only a single vertex is created for each distinct concept in the text. Finally, the n upper levels of this hierarchy are removed, once again because they may represent very general concepts. The n parameter must be empirically determined since it depends on the document type and domain. Figure 4.3 shows the graph for a generic example sentence.

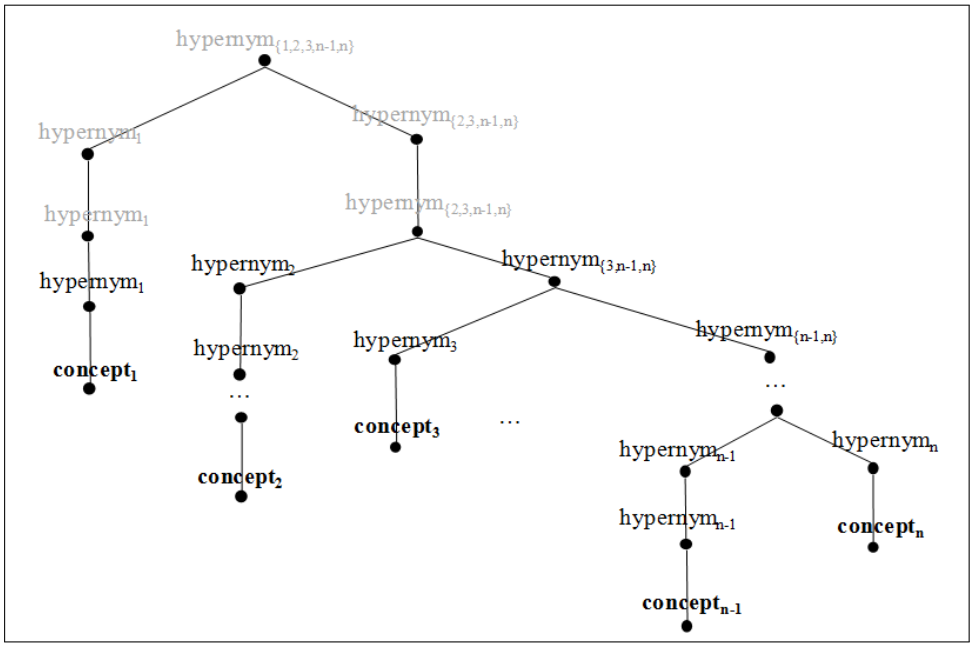


Figure 4.3: Sentence graph

4.4. Step IV: Document Representation

Next, all the sentence graphs are merged to a single *document graph*. This graph can be extended with more specific relationships between nodes (apart from hypernymy) to obtain a more complete representation of the document. These relations may be defined in the *config.xml* file according to the application domain and the possibilities offered by the knowledge base.

Next, each edge is assigned a weight, which is directly proportional to the depth in the hierarchy at which the concepts lies. This way a greater

importance is attached to specific concepts than to general ones. In order to compute these weights, the summarization architecture provides two alternative weighting functions:

- The **Jaccard similarity coefficient** (Jaccard, 1901). The Jaccard similarity coefficient between two sample sets is defined as the size of the intersection divided by the size of the union of the sample sets. Following this definition, the weight of an edge representing a relation between two non-leaf concepts A and B is calculated using Equation 4.1, where α is the set of all the parents of the concept A, including A, and β is the set of all the parents of the concept B, including B. The weight of an edge representing any other relation between a pair of leaf vertices is always '1'.

$$weight(A, B) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \tag{4.1}$$

Figure 4.4 shows the graph for a generic document where the edges have been labeled using the Jaccard similarity coefficient.

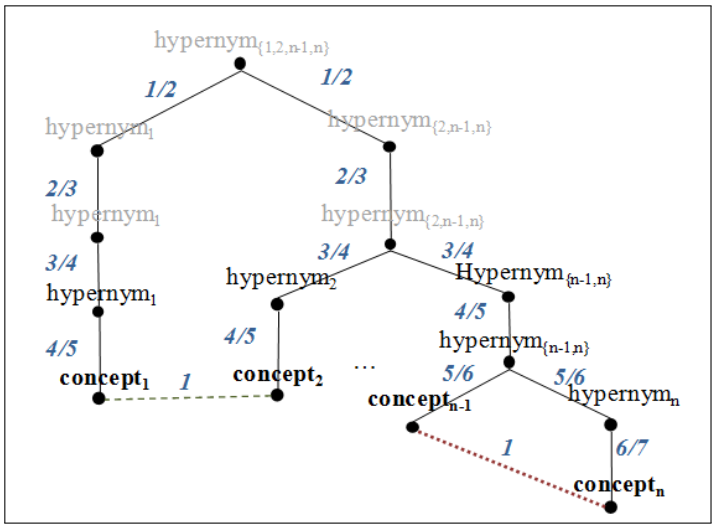


Figure 4.4: Document graph (Jaccard coefficient)

- The **Dice-Sorensen similarity coefficient** (Legendre y Legendre, 1998). It is a similarity measure related to the Jaccard coefficient which attaches double importance to those elements belonging to both sample sets. Following this definition, the weight of an edge representing

a relation between two non-leaf concepts A and B is calculated using Equation 4.2, where α is the set of all the parents of the concept A, including A, and β is the set of all the parents of the concept B, including B. The weight of an edge representing any other relation between a pair of leaf vertices is always '1'.

$$weight(A, B) = \frac{2 \times |\alpha \cap \beta|}{2 \times |\alpha \cap \beta| + |\alpha - \beta|} \tag{4.2}$$

Figure 4.5 shows the graph for a generic document where the edges have been labeled using the Dice-Sorensen similarity coefficient.

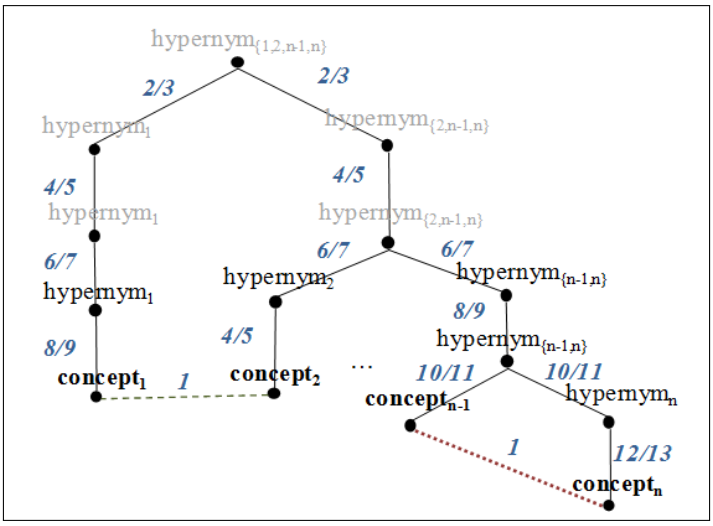


Figure 4.5: Document graph (Dice-Sorensen coefficient)

4.5. Step V: Concept Clustering and Subtheme Recognition

The following step consists of clustering the concepts in the document graph using a *degree-based clustering algorithm* (Erkan y Radev, 2004). The aim is to construct sets or clusters of concepts that are closely related in meaning, under the assumption that each set represents a different subtheme in the document and that the most central concepts in the cluster (the *centroids*) give the necessary and sufficient information related to its subtheme.

The working hypothesis is that the document graph is an instance of

a scale-free network (Barabási y Albert, 1999). A scale-free network is a complex network in which some nodes are highly connected to other nodes in the network, while the remaining nodes are relatively quite unconnected. The salience of each vertex in the graph is then computed. Following (Yoo, Hu, y Song, 2007), the *salience* or *prestige* of a vertex (v_i) is defined as the sum of the weights of the edges (e_j) that are connected to it, as shown in Equation 4.3.

$$salience(v_i) = \sum_{\substack{\forall e_j | \exists v_k \\ \wedge e_j connect(v_i, v_k)}} weight(e_j) \tag{4.3}$$

The n vertices with a highest salience are named *Hub Vertices*, and they represent the most connected nodes in the graph, taking into account both the number and the weight of the edges. The n parameter must be empirically determined, since depends on the type and length of the documents to be summarized. The clustering algorithm starts grouping the hub vertices into *Hub Vertex Sets (HVS)*. These can be interpreted as set of concepts strongly related in meaning and represent the centroids of the clusters. To construct the HVS, the clustering algorithm first identifies the pairs of hub vertices that are most closely connected and merges them into a HVS. Then, for each pair of HVS, the algorithm checks if the internal connectivity of the vertices they contain is lower than the connectivity between them. If it is the two HVS are merged. This decision is encouraged by the assumption that the clustering should show maximum intra-cluster connectivity but minimum inter-cluster connectivity. Intra-cluster connectivity for a HVS is calculated as the sum of the weights of all edges connecting two vertices within the HVS, as shown in Equation 4.4. Inter-cluster connectivity for two HVS is computed as the sum of the weights of all edges connecting two vertices, each vertex belonging to one of the HVS, as shown in Equation 4.5.

$$Intra - connectivity(HVS_i) = \sum_{\substack{\forall e_j | \exists v,w \in HVS_i \\ \wedge e_j connect(v,w)}} weight(e_j) \tag{4.4}$$

$$Inter - connectivity(HVS_i, HVS_j) = \sum_{\substack{\forall e_j | \exists v \in HVS_i, w \in HVS_j \\ \wedge e_k connect(v,w)}} weight(e_k) \tag{4.5}$$

Finally, the remaining vertices (i.e. those not included in the HVS) are iteratively assigned to the cluster to which they are more connected. This connectivity is computed as the sum of the weights of the edges that connect the target vertex to the other vertices in the cluster, as shown in Equation 4.6.

$$connectivity(v, HVS_i) = \sum_{\substack{\forall e_j | \exists w \in HVS_i \\ \wedge e_j connect(v,w)}} weight(e_j) \tag{4.6}$$

4.6. Step VI: Sentence-to-Cluster Assignment

Once the concept clusters have been created, the aim of this step is to compute the semantic similarity between each sentence graph and cluster. Thus, a measure of the similarity between a cluster and a sentence graph is needed. As both representations are quite different in size, traditional graph similarity metrics (e.g. the edit distance (Levenshtein, 1966)) are not appropriated. Instead, a non-democratic voting mechanism is used. Each vertex, v_k, within a sentence graph, S_j, assigns a vote $w_{k,j}$ to a cluster C_i if the vertex belongs to that cluster's HVS, half a vote it the vertex belongs to the cluster but not to its HVS, and no votes otherwise. The similarity between the sentence and the cluster is then computed as the sum of the votes assigned by all vertices in the sentence to the cluster, as shown in Equation 4.7. It should be noted that a sentence may assign votes to several clusters (i.e. it may contain information about different themes). The result of this process is a sentence ranking for each cluster in decreasing order of semantic similarity.

$$Semantic\_Similarity(C_i, S_j) = \sum_{v_k | v_k \in S_j} w_{k,j} \tag{4.7}$$

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

The values for parameters γ and δ have been empirically set to 1.0 and 0.5 respectively, which means that the concepts belonging to the HVS are attached double importance.

4.7. Step VII: Sentence Selection

At this point in the exposition, it is important to remind that extractive summarization works by choosing salient sentences in the original document. In this work, sentence selection is assessed based on the similarity between sentences and clusters as defined in Equation 4.7. The number of sentences to be selected (N) varies on the desired summary compression. The system allows the user to specify the compression rate for the summary, both in terms of sentences and words ($\pm$X %). Three different heuristics have been investigated, each of them aiming to produce a different type of summary:

- **Heuristic 1:** Under the hypothesis that the cluster with more concepts represents the main theme or topic in the document, the top ranked N sentences from this cluster are selected. The aim of this heuristic is to include in the summary just the information related to the main topic of the document.
- **Heuristic 2:** All clusters contribute to the summary proportionally to their sizes. Therefore, for each cluster, the top ranked n_i sentences are selected, where n_i is proportional to the size of the cluster. The aim of this heuristic is to include in the summary information about all different topics covered in the source, regardless of their salience.
- **Heuristic 3:** Halfway between the two heuristics above, this one modifies Equation 4.7 to computes a single score for each sentence as the sum of the votes assigned to each cluster adjusted to their sizes, as shown in Equation 4.8. Then, the N highest scoring sentences are selected. The aim of this heuristic is to select most of the sentences from the main topic of the document but also include other dependent or secondary information that might be relevant to the user.

$$Semantic\_Similarity(S_j) = \sum_{C_i} \frac{similarity(C_i, S_j)}{|C_i|} \tag{4.8}$$

Two additional features, apart from the semantic similarity, have been tested when computing the salience of sentences: *sentence location* and *similarity with the title*. Despite their simplicity, these features are commonly used in the most recent works on extractive summarization (Bossard, Généreux, y Poibeau, 2008; Bawakid y Oussalah, 2008).

- **Sentence location:** The position of the sentences in the document has been traditionally considered an important factor in finding the sentences that are most related to the topic of the document (Brandow, Mitze, y Rau, 1995; Bossard, Généreux, y Poibeau, 2008; Bawakid y Oussalah, 2008). Following this assumption, sentences close to the beginning and the end of the document are supposed to deal with the main theme of the document, and so more weight is assigned to them. In this work, a score $Location \in \{0, 1\}$ is calculated for each sentence as shown in Equation 4.9, where M represents the number of sentences in the document and m_j represents the position of the sentence, S_j, within the document.

$$Location(S_j) = max\{\frac{1}{m_j}, \frac{1}{M - m_j + 1}\} \tag{4.9}$$

- **Similarity with the title:** The title given to a document by its author is intended to comprise the most significant information in the document, and so it is frequently used to quantify the relevance of a sentence (Bawakid y Oussalah, 2008). In this work, the similarity of a sentence to the title is computed as the proportion of common concepts between the sentence and the title, as shown in Equation 4.10.

$$Title(S_j) = \frac{Concepts_{S_j} \cap Concepts_{title}}{Concepts_{S_j} \cup Concepts_{title}} \tag{4.10}$$

The final selection of the sentences for the summary is based on the weighted sum of these feature values, as stated in Equation 4.11. The values for the parameters λ, θ and χ strongly depend on the application domain, and so need to be empirically determined.

$$Score(S_j) = \lambda \times Sem\_Sim(S_j) + \theta \times Location(S_j) + \chi \times Title(S_j) \tag{4.11}$$

It should be noted that, since a sentence may assign votes to several clusters or themes, the Heuristic 2 may include repeated sentences in summary. To avoid this, the system makes sure not to add to the summary any sentence that is already part of it.

Chapter 5

Case Study: Mono-document Summarization of Biomedical Scientific Literature

The first case study aims to configure the method proposed in the previous chapter to produce summaries of scientific articles in the biomedical domain. The chapter is organized as follows. Section 5.1 studies the peculiarities of the biomedical language and the structure of scientific articles in this field. Section 5.2 details the process performed to adapt the generic summarization algorithm to work with biomedical scientific literature.

5.1. The Biomedical Language and the Scientific Papers

Biomedical texts exhibit certain unique attributes that must be taken into account in the development of a summarization system. First, medical information arises in a wide range of document types (Afantenos, Karkaletsis, y Stamatopoulos, 2005): electronic medical records, scientific articles, semi-structured databases, X-ray images or even videos. Each document type presents very distinct characteristics that should be considered in the summarization process. We focus on scientific articles, which are mainly composed of text but usually contain tables and images that may enclose important information that ought to appear in the summary. Biomedical papers often present the *IMRAD* structure (*Introduction*, *Method*, *Results*

And Discussion), but sometimes present additional sections such as *Abbreviations*, *Limitations of the Study* and *Competing Interests*. In most cases, depending on the summarization task, this knowledge about the article layout can be exploited in order to improve the automatic summaries.

Second, the peculiarities of the terminology make it difficult to automatically process biomedical information (Nadkarni, 2000). The first challenge is the problem of **synonyms** (the use of different terms to designate the same concept) and **homonyms** (the use of words/phrases with multiple meanings). For instance, the syntagms *coronary failure* and *heart attack* stand for the same concept, while the term *anaesthesia* may refer to either the loss of sensation or the procedure for pain relief. Another handicap to automatic concept recognition is the presence of **neologisms**, which are newly coined words that are not likely to be found in a dictionary (e.g. the term *coumadinize* for the administration of coumadin). Finally, **elisions** and **abbreviations** complicate the automatic processing of medical text. Elision is the omission of words or sounds in a word or phrase. An example of elision is *white count*, understood by physicians as *the count of white blood cells*. An abbreviation is a shortened form of a word or phrase, for example, the use of *OCP* to refer to *oral contraceptive pills*.

5.2. Method Specification for Biomedical Literature Summarization

The aim of this section is to specify the process and resources needed to adapt the generic method for summarizing biomedical scientific articles. Moreover, in order to clarify how the algorithm works, a complete document example from the BioMed Central[1] corpus (document *cvm-2-6-254.xml*) is elaborated throughout the summarization process. The document body presents 58 sentences and is attached in Appendix B.1 of the Spanish version of this document.

5.2.1. Step I: Document Pre-processing

The generic document pre-processing step (Section 4.1) is adapted to reflect the peculiarities of the biomedical domain and the scientific articles:

[1]BioMed Central Corpus. http://www.biomedcentral.com/info/about/datamining/. Last accessed: 1 November 2010

- First, the following sections of the document that are considered irrelevant for inclusion in the summary are removed: *Authors*, *Institutions*, *Source*, *Year*, *ISSN*, *Volume*, *Issue*, *Url*, *Competing interests*, *Acknowledgments*, *References* and section headings. Tables and figures are extracted.

- Second, if the document includes an *Abbreviations* section, the abbreviations and their expansions are extracted from it. This information is next used to replace these shortened forms in the document body. For example, if the Abbreviations section defines *embryonic submandibular* as the expansion of *SMG* for a particular document and that document contains the phrase "survivin may be a key mediator of SMG epithelial cell survival" then that phrase would become "survivin may be a key mediator of embryonic submandibular epithelial cell survival".

- Third, in order to expand the acronyms and abbreviations not defined in the Abbreviations section, the BioText[2] software for abbreviation definition recognition (Schwartz y Hearst, 2003) is used. This software is publicly available and allows the identification of abbreviations and their expansions in biomedical text with an average precision of 95 %. Abbreviations are then substituted with their expansions in the document body.

- Finally, a stop list from Medline[3] is used to remove generic terms.

5.2.2. Step II: Concept Recognition

We use the Unified Medical Language System (Section 3.1.1) as the domain knowledge base and the MetaMap program (Section 3.1.3) for translating the biomedical documents to domain concepts.

We first run the MetaMap program over the text in the body section of the document. MetaMap identifies all the phrases that could be mapped onto a UMLS CUI, retrieves and scores all possible CUI mappings for each phrase, and returns all the candidates along with their score. The semantic

[2]BioText. http://biotext.berkeley.edu/software.html. Last accessed: 1 November 2010

[3]PubMed StopWords.
http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords. Last accessed: 1 November 2010

type for each concept mapping is also returned. Table 5.1 shows this mapping for the phrase *tissues are often cold*.

Phrase: "Tissues"
Meta Mapping (1000):
1000 C0040300:Tissues (Body tissue)
Phrase: "are"
Phrase: "often cold"
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0234192:Cold (Cold Sensation)
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0009443:Cold (Common Cold)
MetaMapping (888):
694 C0332183:Often (Frequent)
861 C0009264:Cold (Cold Temperature)

Figure 5.1: MetaMap mapping for the phrase *Tissues are often cold*

It is important to note that, in presence of lexical ambiguity, MetaMap may fail to identify a unique mapping for a given phrase (Aronson y Lang, 2010). This occurs, for instance, for the phrase *Tissues are often cold*, where MetaMap returns three candidate concepts with equal scores for "cold" (*Cold Sensation*, *Common Cold* and *Cold Temperature*). To select the correct mapping according to the context in which the phrase appears, the summarizer allows the user to select one of the following WSD algorithms:

- The Personalized PageRank algorithm (Section 3.4), which has been adapted to use the UMLS Metathesaurus as knowledge base. The UMLS is converted into a graph in which the CUIs are the nodes and the edges are derived from the *MRREL* Metathesaurus table. All relations in this table are included in the graph. The output from MetaMap is used to provide the list of possible CUIs for each term in a document and these are passed to the disambiguation algorithm. We use both the standard (*ppr*) and "word to word" (*ppr-w2w*) variants of the Personalized PageRank approach for disambiguation.

- The Journal Descriptor Indexing methodology (JDI) (Humphrey et al., 2006) provided by MetaMap, which is invoked using the *-y* flag. This algorithm is based on semantic type indexing, which resolves Metathesaurus ambiguity by choosing a concept having the most likely semantic type for a given context. Using the *-y* flag forces MetaMap to choose a single mapping if there is more than one candidate for a

given phrase. However, when the candidate concepts share the same semantic type the JDI algorithm may fail to return a single mapping. When this happens the first mapping returned by MetaMap is selected.

Concepts from very generic UMLS semantic types are discarded, since they have been found to be excessively broad. These semantic types are Quantitative Concept, Qualitative Concept, Temporal Concept, Functional Concept, Idea or Concept, Intellectual Product, Mental Process, Spatial Concept and Language. These types were empirically determined by evaluating the summaries generated using UMLS concepts from different combinations of semantic types. Examples of concepts from the example document that belong to these semantic types are given below:

- **Quantitative Concept:** Lowered, Two, Four, Several.
- **Qualitative Concept:** Firstly, Initial, Possibly, Definite.
- **Temporal Concept:** Previous, Year, Seconds, Frequent.
- **Functional Concept:** Purpose, Designate, Treat, Lead
- **Idea or Concept:** Reasons, Complete, Goal, Accepted.
- **Intellectual Product:** Class, Groups, Agencies, Reports.
- **Spatial Concept:** Upper, Separate, Address, Over.

Table 5.1 shows the UMLS concepts for the sentence *S1* from the example document: *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.*

Concept	MetaMap Score	Semantic Type
~~Goals~~	~~1000~~	~~Intellectual Product~~
Clinical Trials	1000	Research Activity
Cardiovascular system	694	Body System
~~Mortality vital statistics~~	~~861~~	~~Quantitative Concept~~
~~Morbidity - disease rate~~	~~1000~~	~~Quantitative Concept~~
Cerebrovascular accident	1000	Disease or Syndrome
Coronary heart disease	1000	Disease or Syndrome
Congestive heart failure	1000	Disease or Syndrome
~~Evidence of~~	~~660~~	~~Functional Concept~~
~~Basis~~	~~660~~	~~Functional Concept~~
Clinicians	1000	Prof. or Occup. Group
~~Treatment intent~~	~~1000~~	~~Functional Concept~~
Hypertensive disease	1000	Disease or Syndrome

Table 5.1: UMLS concepts for the sentence *S1*. Ignored concepts of generic semantic types appear crossed out

5.2.3. Step III: Sentence Representation

For each sentence in the document, the hypernyms of the UMLS concepts returned by MetaMap are retrieved from the *MRHIER* Metathesaurus table. All the hierarchies for each sentence are merged creating a sentence graph, as explained in Section 4.3. The two upper levels of this hierarchy are removed, once again because they represent very general concepts. Figure 5.2 shows the graph for the example sentence used in the previous section.

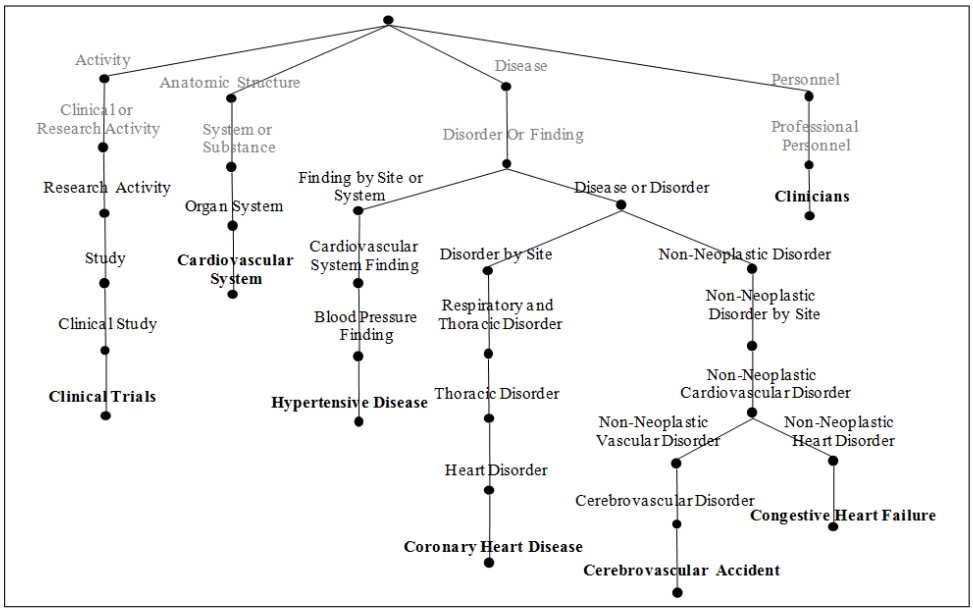


Figure 5.2: An example of sentence graph for the sentence *S1* from document *cvm-2-6-254.xml*. Very general concepts that are ignored appear softened. Final concepts are shown in bold type

5.2.4. Step IV: Document Representation

The sentence graphs are then merged to create a single document graph. This graph is extended with more semantic relations to obtain a more complete representation of the document. Various types of information from the UMLS can be used to extend the graph. We experimented with different sets of relations: (1) no relation (apart from hypernymy), (2) the *associated with* relation between semantic types from the UMLS Semantic Network, (3) the *related to* relation between concepts from the UMLS Metathesaurus and (4) both the *associated with* and *related to* relations. *Related to* relations are extracted from the *MRREL* Metathesaurus table, and *associated with* relations

are extracted from the *SRSTR* Semantic Network table. Finally, the edges of this document graph are labeled, as explained in Section 4.4. Figure 5.3 shows an example of document graph for a simplified document composed of two sentences, *S2* and *S3*, extracted from document *cvm-2-6-254.xml* from the BioMed Central corpus:

S2 *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.*

S3 *While event rates for fatal cardiovascular disease were similar, there was a disturbing tendency for stroke to occur more often in the doxazosin group, than in the group taking chlorthalidone.*

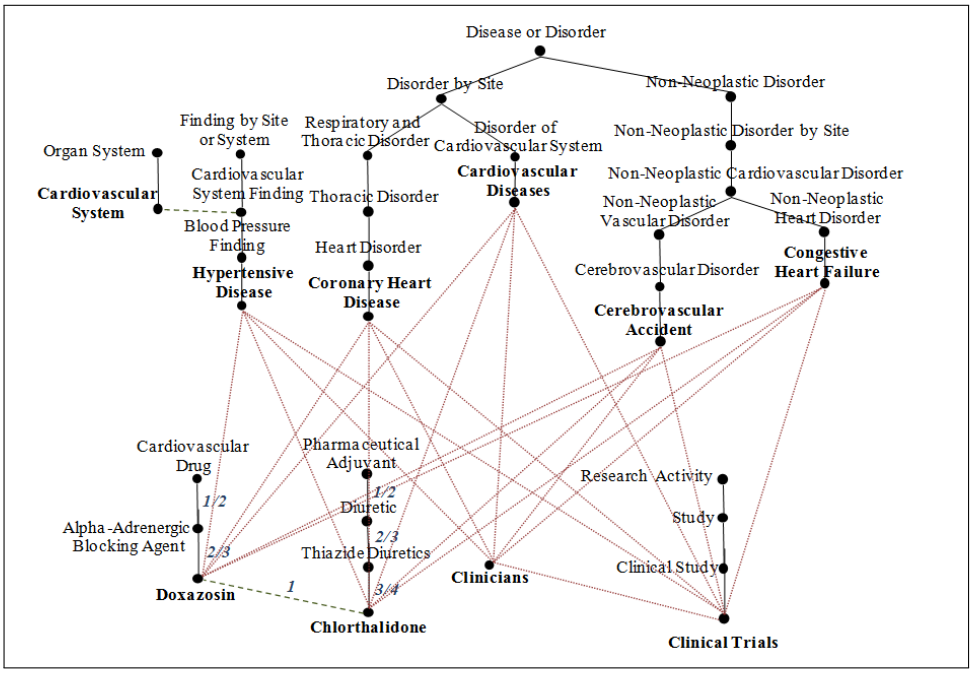


Figure 5.3: An example of simplified document graph from sentences *S2* and *S3*. The edges of a portion of this graph have been labeled with their weights using the Jaccard similarity coefficient

5.2.5. Step V: Concept Clustering and Subtheme Recognition

In this step, the algorithm does not require any modification with respect to what is explained in Section 4.5. Following with the working example, Table 5.2 shows the HVS or centroids of the clusters generated by the algorithm for the document *cvm-2-6-254.xml* from the BioMed Central corpus, when 20 % of the concepts in the document graph are used as hub vertices and this graph is built using *is a* and *associated with* relations.

HVS 1 (3 concepts)		
Study	Provide	View
HVS 2 (28 concepts)		
Analysis of substances	Hepatic	Audiological observations
Blood	Entire upper arm	Age
receptor	Base	Very large
Other therapy NOS	Reserpine	Related personal status
Reduction - action	Agent	In care
Therapeutic procedure	Discontinued	Cardiovascular event
Finding	Admin. occup. activities	Cardiovascular system
Primary operation	Descriptor	heart rate
Entire lung	Guide device	Adverse reactions
Entire heart		
HVS 3 (4 concepts)		
Systolic hypertension	May	Person
Blood Pressure		
HVS 4 (32 concepts)		
Duplicate concept	Lower	Clonidine
Articular system	Support, device	Hydralazine
Entire hand	Antihypertensive Agents	Adrenergic beta-Antag.
Body system structure	Falls	Diuretics
Diastolic blood pressure	Angiotensin-Conv. Enz. Inh.	PREVENT
Expression procedure	Prazosin	CONCEPT Drug
Prevention	Immune Tolerance	Calcium Channel Blockers
Chlorthalidone	Tissue damage	Assessment procedure
Hopelessness	Ramipril	Qualifier value
Lisinopril	Amlodipine	Unapproved attribute
Doxazosin	Reporting	

Table 5.2: HVS for the *cvm-2-6-254.xml* document graph

5.2.6. Step VI: Sentence-to-Cluster Assignment

In this step, the algorithm does not require any adaptation with respect to what is explained in Section 4.6. Following with our example, Table 5.3 shows the scores assigned by each sentence in the *cvm-2-6-254.xml* document to each concept cluster.

Sentence	C.1	C.2	C.3	C.4	Sentence	C.1	C.2	C.3	C.4
1	98.0	172.0	74.0	208.0	30	6.0	7.0	4.0	7.0
2	13.0	21.0	9.0	18.0	31	9.0	14.0	5.0	12.0
3	8.0	13.0	4.0	17.0	32	4.0	6.0	2.0	5.0
4	8.0	19.0	4.0	16.0	33	3.0	3.0	2.0	2.0
5	9.0	13.0	4.0	16.0	34	18.0	18.0	9.0	20.0
6	6.0	4.0	4.0	8.0	35	8.0	15.0	5.0	17.0
7	1.0	1.0	1.0	3.0	36	1.0	6.0	1.0	2.0
8	8.0	15.0	4.0	25.0	37	8.0	12.0	4.0	12.0
9	3.0	9.0	2.0	10.0	38	6.0	6.0	4.0	9.0
10	9.0	10.0	6.0	10.0	39	0.0	2.0	0.0	6.0
11	4.0	9.0	2.0	14.0	40	2.0	2.0	1.0	2.0
12	5.0	7.0	3.0	8.0	41	1.0	4.0	1.0	8.0
13	16.0	29.0	11.0	25.0	42	9.0	14.0	6.0	17.0
14	7.0	5.0	5.0	15.0	43	14.0	14.0	8.0	26.0
15	3.0	10.0	1.0	9.0	44	11.0	12.0	7.0	22.0
16	11.0	11.0	7.0	15.0	45	3.0	6.0	2.0	9.0
17	4.0	5.0	2.0	6.0	46	5.0	16.0	2.0	18.0
18	5.0	11.0	3.0	17.0	47	3.0	5.0	2.0	11.0
19	5.0	6.0	2.0	4.0	48	4.0	11.0	3.0	20.0
20	17.0	24.0	10.0	25.0	49	7.0	9.0	4.0	12.0
21	16.0	31.0	12.0	27.0	50	5.0	10.0	4.0	13.0
22	11.0	23.0	8.0	27.0	51	3.0	9.0	3.0	6.0
23	3.0	8.0	4.0	14.0	52	2.0	6.0	1.0	6.0
24	14.0	17.0	9.0	27.0	53	2.0	2.0	2.0	4.0
25	7.0	27.0	6.0	20.0	54	0.0	0.0	0.0	0.0
26	5.0	7.0	3.0	12.0	55	0.0	0.0	0.0	0.0
27	5.0	6.0	2.0	8.0	56	2.0	5.0	2.0	4.0
28	2.0	5.0	1.0	2.0	57	12.0	14.0	7.0	22.0
29	10.0	22.0	8.0	18.0	58	4.0	7.0	2.0	7.0

Table 5.3: Sentence-to-clusters similarity

5.2.7. Step VII: Sentence Selection

Once again, since this step does not need to be modified with respect to the generic algorithm, we just present here the result of this step for the example document. Table 5.4 shows the sentences selected for each heuristic along with their scores. To produce this extract, the compression rate is set to 15 % and no positional or similarity with the title criteria are used. The tables and figures from the source that are referred to from any sentence belonging to the summary are also included in it.

Finally, Tables 5.5, 5.6 and 5.7 show the extracts generated by each heuristic.

Heuristic 1		Heuristic 2		Heuristic 3	
Sentence	**Score**	**Sentence**	**Score**	**Sentence**	**Score**
1	208.0	1	208.0	1	101.33
22	27.0	21	31.0	21	16.47
24	27.0	13	29.0	13	15.81
43	26.0	22	27.0	31	15.53
8	25.0	24	27.0	8	15.06
20	25.0	25	27.0	10	13.2
44	22.0	29	26.0	25	12.67
3	22.0	8	25.0	20	12.5
34	20.0	20	25.0	2	11.53

Table 5.4: Sentences selected by each heuristic and their scores

	Heuristic 1
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 ñull'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.

Table 5.5: Extract generated using Heuristic 1

	Heuristic 2
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 ñull'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.

Table 5.6: Extract generated using Heuristic 2

	Heuristic 3
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.
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.
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.

Table 5.7: Extract generated using Heuristic 3

Chapter 6

Case Study: Mono-document Summarization of News

The second case study aims to configure the method proposed in Chapter 4 to produce summaries of news articles. The chapter is organized as follows. Section 6.1 studies the characteristics of the journalism language and the structure of news articles. Section 6.2 details the process performed to adapt the generic summarizer to work with news articles.

6.1. The Language of Journalism and the News Articles

The journalistic writing presents certain important defining attributes. It arises in a wide rage of document types, such as news articles, editorials, feature articles or columns, that differ in their structure and content. In this work, we focus on the news article. A news article usually presents the following elements:

- The **headline** or text at the top of the article, indicating the nature of the article. The headline should not be a summarization of the article; instead it should serve the purpose of getting the reader's attention.
- The **lead** paragraph, which sums up the focus of the story.
- The **body**, which details and elaborates the news story. The information in the article body is usually presented according to the *inverted pyramid* form, so that the most important information is placed first

within the article followed by the remaining material in order of diminishing importance.

On the other hand, journalistic writing is expected to be **concise**, which means that a news article should not contain redundant information. The writer can also give facts and detailed information following answers to general questions like who, what, when, where, why and how. Besides, the thematic scope is very wide, as is the vocabulary used. Finally, news articles tend to be riddled with cliches and metaphors.

6.2. Method Specification for News Articles Summarization

We next describe the process and resources used to adapt the generic algorithm for summarizing news articles. To illustrate how the algorithm works, a complete document example from the DUC 2002 corpus (document *AP880911-0016*) is drawn. This document presents 16 sentences and is attached in Appendix B.2 of the Spanish version of this document.

6.2.1. Step I: Document Pre-processing

The pre-processing step explained in Section 4.1 is configured to consider the characteristics of the journalism domain and the news articles:

- First, the following sections of the document that are considered irrelevant for inclusion in the summary are removed: *Lead*, *Publication date*, *Authors* and *Publisher*.
- Second, since abbreviations and acronyms are quite frequent in this type of documents, we use the abbreviation lists from the *ANNIE Gazetteer* module in GATE to replace these shortened forms in the document body with their corresponding expansions. Examples of abbreviations that are very common in news articles are NY (*New York*), Al (*Alabama*), Co (Company), etc.
- Finally, the WordNet stop list[1] is used to remove generic terms.

[1] WordNet Stop List.
http://www.d.umn.edu/~tpederse/Group01/WordNet/wordnet-stoplist.html. Last accessed: 1 November 2010

6.2.2. Step II: Concept Recognition

Since the vocabulary in news articles is relatively broad and little specialized, we use the WordNet lexical database (Section 3.2.1) as the knowledge source, and the WordNet::SenseRelate program (Section 3.2.3) for translating the news articles to WordNet concepts. Thus, we do not use here domain-specific resources, but general-purpose ones.

We run WordNet::SenseRelate over the text in the body section of the document. WordNet::SenseRelate executes a WSD algorithm, specified by the user in the *config.xml* file, to assign a sense or meaning (as found in WordNet) to each word in the text. As the result, the list of WordNet concepts that are found within the text is returned. Each concept is described by its name or literal and its sense number.

Figure 1 shows the result of applying WordNet::SenseRelate to the sentence *S4* from document *AP880911-0016*: *Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.*

Term	WN Sense	Term	WN Sense
Hurricane	1	**populate**	2
Gilbert	2	**south**	1
sweep	1	**coast**	1
Dom. Rep	1	**prepare**	4
Sunday	1	**high**	2
civil	1	**wind**	1
defense	9	**heavy**	1
alert	1	**rain**	1
heavily	2	**sea**	1

Table 6.1: WordNet::SenseRelate mapping for sentence *S4*

6.2.3. Step III: Sentence Representation

For each sentence in the document, the hypernyms of the concepts for nouns are retrieved from WordNet, using the *JWI* API[2], and the hierarchies of all the concepts in the same sentence are merged to build a graph representing it. Our experimental results have shown that the use of verbs in these graphs decreases the quality of the summaries, while adjectives and adverbs are not included because they do not present the hypernymy relation in WordNet. Finally, the three upper levels of this hierarchy are removed, since they

[2]JWI (the MIT Java WordNet Interface). http://projects.csail.mit.edu/jwi/. Last accessed: 1 November 2010

represent concepts with a excessively broad meaning. Figure 6.1 shows the graph for the sentence *S4* presented in the previous section.

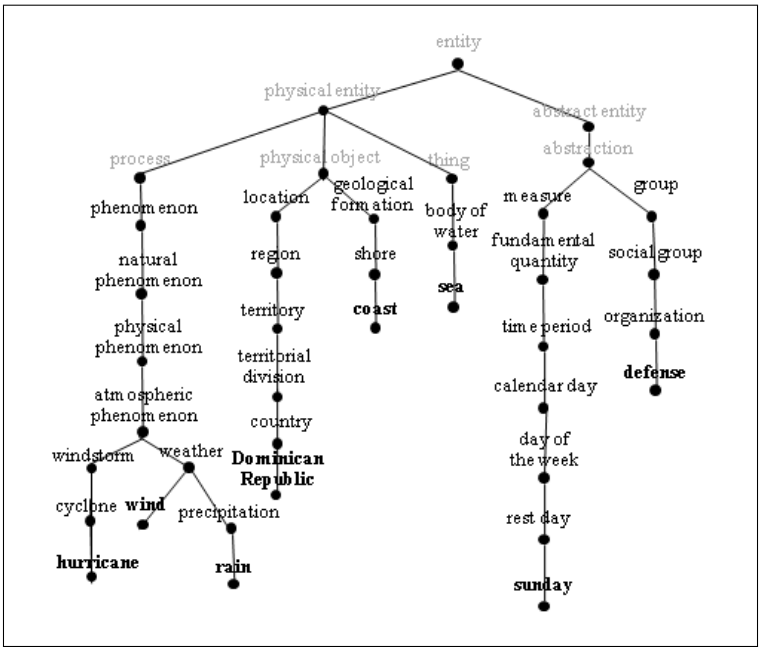


Figure 6.1: An example of sentence graph for the sentence *S4* from document *AP880911-0016*

6.2.4. Step IV: Document Representation

The sentence graphs are then merged to create a single document graph, which is extended with further relations to obtain a more accurate document representation. We have conducted several experiments using a semantic similarity relation apart from the *is a* relation previously mentioned. To this end, we compute the similarity between every pair of leaf concepts in the graph, using the WordNet::Similarity package (Section 3.2.2). This package implements a variety of semantic similarity and relatedness measures based on the information found in WordNet. The summarization system allows the user to specify the similarity measure to be used in the *config.xml* file. To expand the document graph with these additional relations, a new edge is added between two leaf nodes if the similarity between the underlying concepts exceeds a configurable *similarity threshold*. Figure 6.2 shows an example of an extended document graph for a fictitious document that consists solely of the sentence S4. For this example, we used the Lesk similarity

measure (Lesk, 1986) and 0.01 as similarity threshold.

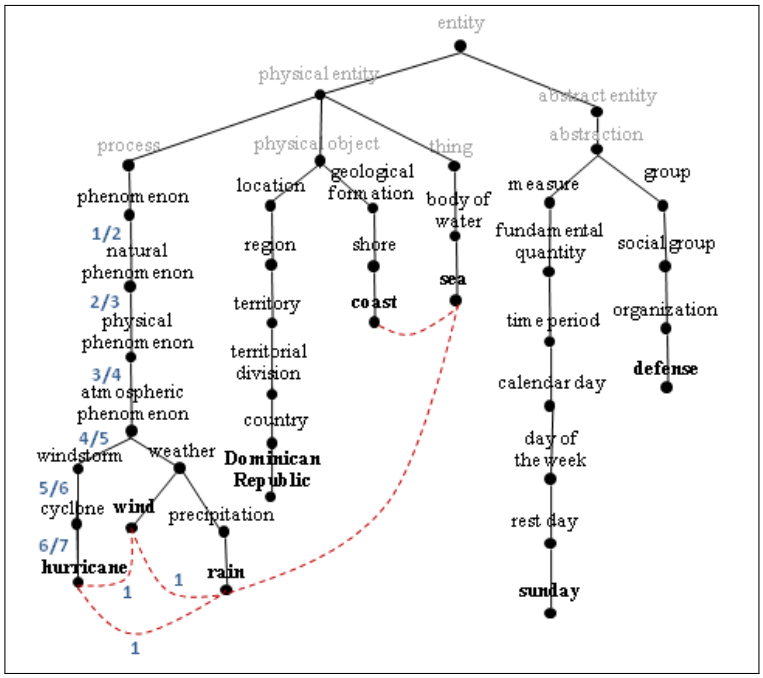


Figure 6.2: An example of simplified document graph from sentence S4. The edges of a portion of this graph have been labeled with their weights using the Jaccard similarity coefficient

6.2.5. Step V: Concept Clustering and Subtheme Recognition

In this step, the algorithm does not require any modification with respect to what is explained in Section 4.5. Following with the working example, Table 6.2 shows the HVS of the clusters generated by the algorithm for the document *AP880911-0016* from the DUC 2002 corpus, when 10 % of the concepts in the document graph are used as hub vertices.

6.2.6. Step VI: Sentence-to-Cluster Assignment

In this step, the algorithm does not require any modification with respect to what is explained in Section 4.6. Following with our example, Table 6.3 shows the scores assigned by each sentence in the *AP880911-0016* document to each concept cluster.

HVS 1 (7 concepts)				
atmospheric condition cyclone	wind hurricane	flood	precipitation	weather
HVS 2 (3 concepts)				
people	resident	inhabitant		
HVS 3 (3 concepts)				
casualty	fatality	accident		
HVS 4 (5 concepts)				
island	gulf	region	republic	Caribbean
HVS 5 (3 concepts)				
Saturday	Sunday	weekday		

Table 6.2: HVS for the *AP880911-0016* document graph

Sent.	**C.1**	**C.2**	**C.3**	**C.4**	**C.5**	**Sent.**	**C.1**	**C.2**	**C.3**	**C.4**	**C.5**
1	54.0	37.0	34.5	49.5	19.5	9	21.5	6.5	18.0	29.5	17.0
2	27.0	3.5	4.5	20.5	0.0	10	46.5	9.0	20.5	41.5	1.5
3	18.0	15.5	48.0	5.0	32.5	11	1.0	0.5	5.5	0.0	0.0
4	17.5	14.5	12.0	38.5	2.0	12	41.5	25.0	20.5	38.5	28.0
5	5.0	38.0	21.0	40.0	2.0	13	40.0	21.0	18.0	35.5	16.5
6	48.0	10.0	22.5	45.5	19.5	14	35.5	28.0	20.5	21.5	2.5
7	29.5	9.0	14.5	48.0	18.0	15	48.5	16.0	22.5	18.0	0.0
8	50.0	8.0	15.0	51.5	3.5	16	30.5	12.5	18.0	39.0	9.0

Table 6.3: Sentence-to-clusters similarity

6.2.7. Step VII: Sentence Selection

Since this step does not need to be modified with respect to what is explained in Section 4.7, we just present here the result of this step for the example document. Table 6.4 shows the sentences selected for each heuristic.

Heuristic 1		**Heuristic 2**		**Heuristic 3**	
Sentence	**Score**	**Sentence**	**Score**	**Sentence**	**Score**
1	54.0	1	54.0	1	9.33
8	50.0	8	50.0	3	8.11
15	48.5	15	48.5	12	7.99
6	48.0	7	48.0	6	6.31
10	46.5	3	48.0	13	6.13

Table 6.4: Sentences selected by each heuristic and their scores

Finally, Tables 6.5, 6.6 and 6.7 show the extracts produced by each of the three heuristics. To produce these extracts, the compression rate is set to 30 % and no positional or similarity with the title criteria are used.

	Heuristic 1
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
6	Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
8	The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a broad area of cloudiness and heavy weather rotating around the center of the storm.
10	Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico s south coast.
15	Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.

Table 6.5: Extract generated using Heuristic 1

	Heuristic 2
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
3	There is no need for alarm, Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.
7	The National Hurricane Center in Miami reported its position at 2 a.m Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico and 200 miles southeast of Santo Domingo.
8	The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a broad area of cloudiness and heavy weather rotating around the center of the storm.
15	Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.

Table 6.6: Extract generated using Heuristic 2

	Heuristic 3
1	Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
3	There is no need for alarm, Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.
6	Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
12	San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
13	On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

Table 6.7: Extract generated using Heuristic 3

Chapter 7

Case Study: Multi-document Summarization of Tourism Websites

The third and last case study aims to configure the generic summarization method to summarize multiple web pages that contain information related to a tourist destination. The chapter is organized as follows. First, we present a brief survey of the structure and language characteristics of tourism websites. Secondly, we describe the process accomplished to configure the generic algorithm presented in Chapter 4 to generate summaries from multiple documents with information about tourist attractions.

7.1. The Language of Tourism and the Tourism Information Websites

The documents to summarize present several properties that make them particularly interesting as a case study for automatic summarization. Firstly, the tourism language uses a very rich terminology which involves concepts and ideas from very distinct fields, such as economy, geography, art and history. Second, web pages on tourist destinations are characterized by the wide variety of information they gather. When describing a monument or city, it is usual to include information about where the object is located, when it was built, what other interesting places can be found in the surroundings, visiting information, etc. However, websites frequently contain

a good deal of information which is secondary or unrelated to the object being described that poses difficulty when generating automatic summaries (e.g. nearby hotels and other tourist services, advertisements from the website that hosts the information, personal opinions and experiences of users and so on). Third, when summarizing multiple documents about the same topic, it is expected that the information is repeated across different documents, so that avoiding redundancy in the automatic summary is a key issue.

7.2. Method Specification for Tourism Websites Summarization

As in the previous case studies, in this section we aim to describe the process undertaken in order to adapt the generic summarization algorithm to generate summaries from multiple tourism websites.

The main difference between this case study and the previous ones is that now we face a multi-document summarization problem, which means, as explained in Section 2.4.1, that we need to overcome two major problems: (1) grouping the documents within the corpus that refer to the same topic (or tourist place) and (2) detecting and removing redundancy. In our case study, we do not address the first problem, but consider that the documents are already clustered, so that each document cluster groups together all the documents dealing with the same place. In contrast, we focus on the second problem.

In order to adapt the summarizer to generate summaries from tourism web pages, we follow a very similar process to that described in Chapter 6 for summarizing news articles. Therefore, we use the WordNet stop list to remove generic terms and the abbreviation lists from the *ANNIE Gazetteer* to expand acronyms and abbreviations. We also use the WordNet lexical database as the knowledge source, the WordNet::SenseRelate package to solve word ambiguity and the WordNet::Similarity package to derive the relationships among the vertices in the document graph.

Since the summarizer has not been designed to deal with multi-document summarization, to overcome this shortcoming, we simply merge all documents about the same topic or place into a single document, and run the summarizer over it. After producing the summary, we apply the textual en-

tailment (TE) module describen in (Ferrández et al., 2007) to detect and remove redundancy. This module computes TE between every pair of sentences in the partial summary to determine if one sentence can be deduced from another, thus meaning the information from one sentence is already contained in the other. In such a case, the second sentence is considered as redundant, and consequently discarded from the summary.

Therefore, since the configuration is exactly the same than in the previous case study, we just reproduce in Tables 7.1, 7.2 and 7.3 the summaries generated for a set of 10 web pages with information about the Acropolis of Athens. To produce these extracts, the summary length is set to 200 ± 10 words and 10 % of vertices in the document graph are used as hub vertices. We use the *jcn* algorithm to compute the semantic similarity between WordNet concepts (see Section 3.2.2), and the similarity threshold is set to 0.25. No other criterion for sentence selection (i.e. sentence position and similarity with the title) is used. The full text of these web pages can be found in Appendix B.3 of the Spanish version of this document.

Heuristic 1
Acropolis (Gr akros, akron, edge, extremity + polis, city, pl. acropoleis) literally means city on the edge (or extremity)
In Greek, Acropolis means Highest City .
For purposes of defense, early settlers naturally chose elevated ground, frequently a hill with precipitous sides.
Although originating in the mainland of Greece, use of the acropolis model quickly spread to Greek colonies such as the Dorian Lato on Crete during the Archaic Period.
Because of its classical Greco-Roman style, the ruins of Mission San Juan Capistrano's Great Stone Church in California, United States has been called the American Acropolis.
The word Acropolis, although Greek in origin and associated primarily with the Greek cities Athens, Argos, Thebes, and Corinth (with its Acrocorinth), may be applied generically to all such citadels, including Rome, Jerusalem, Celtic Bratislava, many in Asia Minor, or even Castle Rock in Edinburgh.
The term acropolis is also used to describe the central complex of overlapping structures, such as plazas and pyramids, in many Mayan cities, including Tikal and Copán.
In Central Italy, many small rural communes still cluster at the base of a fortified habitation known as La Rocca of the commune.

Table 7.1: Extract generated using Heuristic 1

Heuristic 2
Acropolis (Gr akros, akron, edge, extremity + polis, city, pl acropoleis) literally means city on the edge (or extremity).
In Greek, Acropolis means Highest City .
For purposes of defense, early settlers naturally chose elevated ground, frequently a hill with precipitous sides.
I think it would be good for school children to learn to think and enjoy words.
I love words and this is a great game.
In many parts of the world, these early citadels became the nuclei of large cities, which grew up on the surrounding lower ground, such as modern Rome.
The word Acropolis, although Greek in origin and associated primarily with the Greek cities Athens, Argos, Thebes, and Corinth (with its Acrocorinth), may be applied generically to all such citadels, including Rome, Jerusalem, Celtic Bratislava, many in Asia Minor, or even Castle Rock in Edinburgh.
By: arizonalady on 06 february 09 Easy Challenging Relaxing Fast Paced Clicky Thinky This game is: Addictive , Good Replay , Original , Involved , Good Value , Kid Friendly I just love all word games.
The most famous example is the Acropolis of Athens, which, by reason of its historical associations and the several famous buildings erected upon it (most notably the Parthenon), is known without qualification as the Acropolis.

Table 7.2: Extract generated using Heuristic 2

Heurística 3
Acropolis (Gr. akros, akron, edge, extremity + polis, city, pl. acropoleis) literally means city on the edge (or extremity).
The Acropolis was designated as a UNESCO World Heritage site in 1987, for its, illustrating the civilizations, myths, and religions that flourished in Greece over a period of more than 1,000 years, the Acropolis, the site of four of the greatest masterpieces of classical Greek art - the Parthenon, the Propylaea, the Erechtheum, and the Temple of Athena Nike-can be seen as symbolizing the idea of world heritage.
The Acropolis of Athens, a hill c.260 ft (80m) high, with a flat oval top c.500 ft (150m) wide and 1,150 ft (350m) long, was a ceremonial site beginning in the Neolithic Period and was walled before the 6th cent. B.C. by the Pelasgians.
Devoted to religious rather than defensive purposes, the area was adorned during the time of Cimon and Pericles with some of the world's greatest architectural and sculptural monuments.
This temple is the first building visitors see as they make their way up the Acropolis.
The first stone temple to Athena, the patron goddess and protector of the city, was built on the Acropolis at the beginning of the 6th century B.C.

Table 7.3: Extract generated using Heuristic 3

Chapter 8

Evaluation

The purpose of the experimentation is to evaluate the adequacy of semantic graphs for extractive summarization, as well as the viability of the method proposed to work with different types of documents with minor changes. To this aim, the summaries generated using the configurations explained in the three previous case studies are compared to those produced by other well-known research and commercial summarizers. The evaluation is accomplished in two phases: (1) a preliminary experimentation to find out the best values for the different parameters involved in the algorithm; and (2) a large-scale evaluation using the ROUGE metrics and following the guidelines observed in the DUC and TAC conferences.

8.1. Evaluation Methodology

8.1.1. Evaluation Metrics

In this work, the ROUGE package (see Section 2.3.1) is used to evaluate the informativeness of the automatic summaries. ROUGE compares an automatic summary (called *peer*) with one or more human-made summaries (called *models* or *reference* summaries), and uses the proportion of n-grams in common between the peer and model summaries to estimate the content that is shared between them. The following ROUGE metrics are used in this work: ROUGE-1, ROUGE-2, ROUGE-W-1.2 and ROUGE-S4.

On the other hand, we also evaluate the readability of the summaries generated for the case study of tourism websites summarization. To this end, we asked three different persons to evaluate a set of 50 randomly-selected

summaries within a 5-point qualitative scale according to the criteria followed in the DUC and TAC conferences (see Section 2.3.2).

8.1.2. Evaluation Collections

In order to measure the system performance in each application domain, we use the following evaluation collections:

- For evaluating the system in the biomedical domain, a collection of 150 biomedical scientific articles randomly selected from the BioMed Central full-text corpus for text mining research[1] was used. This corpus contains approximately 58000 papers of peer-reviewed biomedical research, available in XML structured format. As stated in (Lin, 2004a), the document sample size is large enough to allow significant evaluation results. To evaluate the automatically generated summaries, medical students in their fifth year were asked to generate extractive summaries from papers by selecting between 20-30 % of the sentences within the papers. The summary size was chosen based on the well-accepted affirmation that a summary should not be shorter than 15 % but no longer than 35 % of the source text (Hovy, 2005). Taken into account the characteristics of the type of documents (scientific papers), it has been preferred to give to the domain experts some freedom to decide what is or not important within a margin. As the participation was quite reduced, only ten valid extracts were obtained. These extracts have been used as model summaries in the parametrization process. For the final evaluation, the abstracts of the papers were used as model summaries.

- For evaluating the system in the news domain, the evaluation corpus of DUC 2002, was used. This corpus is the most recent one for single document summarization. The collection is composed of 567 news articles in English, grouped in 59 clusters. All the documents in the same cluster deal with the same topic. Each document comes with one or more human-made abstractive model summaries, which are approximately 100 words long. Since the news items have been selected from different sections of different newspapers and news agencies (e.g.

[1] BioMed Central Corpus. http://www.biomedcentral.com/info/about/datamining/. Last accessed: 1 November 2010

the *Financial Times* and the *Associated Press*), the topics covered in the collection are diverse. As in the biomedical domain, only 10 documents were used for the parametrization. To obtain the extractive model summaries for this parametrization, a person was asked to select the most relevant sentences from each document until the extract reaches a length of 100 words.

- Finally, for evaluating the system in the tourism domain, we used the image-summary pairs collection described in Aker and Gaizauskas (2010). The collection contains 310 images with manually assigned place names. The images feature static places or objects such as *Eiffel Tower* or *London Eye* and are manually categorized by object/scene types such as *bridge, church, river, mountain, etc.* Each image is described by the top 10 web-documents retrieved from the Internet using the Yahoo! search engine, and has up to 4 model summaries which are collected from an online social site, *VirtualTourist.com*. The model summaries were created manually and contain a minimum of 190 and a maximum of 210 words.

8.1.3. Algorithm Parametrization

For each case study, a preliminary experimentation is performed in order to determine the optimal values for the parameters involved in the algorithm. This means studying the following research questions:

1. What percentage of vertices should be considered as hub vertices by the clustering method? (see Section 4.5).

2. Which set of semantic relations should be used to construct the document graph? (see Section 4.4).

3. What similarity coefficient (Jaccard *vs.* Dice-Sorensen) is more appropriated to weight the edges of the document graph? (see Section 4.4).

4. In the news domain, if the semantic similarity relation is used, what similarity threshold should be considered? (see Section 6.2.4).

5. Does the use of traditional criteria (i.e. the position of the sentences and their similarity with the title) improve the quality of the summaries? (see Section 4.7).

6. Which of the three heuristics for sentence selection produces the best summaries? (see Section 4.7).

8.1.4. Comparison with Others Summarizers

The summaries generated in each case study are compared to those obtained using different research and commercial summarizers:

- *SUMMA* (Saggion, 2008), a single and multi-document summarizer that provides several customizable statistical and similarity-based features to score the sentences for extraction. It is one of the most popular research systems and it is publicly available. The features used for this evaluation include the position of the sentences within the document and within the paragraph, their overlap with the title and abstract sections, their similarity to the first sentence, and the frequency of their terms.

- The *LexRank* summarizer, which has been already presented in Section 2.2. Comparison with LexRank will allow us to evaluate whether semantic information provides benefits over merely lexical information in graph-based summarization approaches.

- *Microsoft Autosummarize*[2], a feature of the Microsoft Word software based on a word frequency algorithm.

- Besides, we compare the summaries generated for the second case study (news articles) to those produced by a lexical summarizer improved with anaphoric information (*LeLSA+AR*) (Steinberger et al., 2007), a term frequency summarizer improved with textual entailment (*Freq+TextEnt*) (Lloret et al., 2008) and the five systems which participated in DUC 2002 and achieved the best results (in terms of the ROUGE metric). In short, *System 19* uses topic representation templates to extract salient information; *System 21*, *System 27* and *System 28* employ machine learning techniques to determine the best set of attributes for extraction (word frequency, sentence position, etc.); and *System 29* uses lexical chains.

[2]Microsoft Corporation. Microsoft Office online: automatically summarize a document. http://office.microsoft.com/en-us/word/HA102552061033.aspx. Last accessed: 1 November 2010

- Finally, the summaries generated for the third case study (tourism websites) are compared to those produced by *COMPENDIUM*, a statistical summarizer which relies on different word or noun phrases frequency counting for identifying pieces of information relevant to a location, the *Language Models* summarizer, which uses n-gram language models derived from corpora to capture salient features regarding a certain object type (e.g. church, bridge, etc.) and MEAD[3](Radev, BlairGoldensohn, y Zhang, 2001), a toolkit for multi-document and multi-lingual summarization that produces extractive summaries using a linear combination of features such as the sentence position and length or the centrality of the sentences to the overall topic of the cluster of documents. The results of COMPENDIUM and Language Models have been previously published in (Plaza, Lloret, y Aker, 2010).

- Two baseline summarizers have been also implemented. The first, *Lead baseline*, generates summaries by selecting the first N sentences from the document. The second, *Random baseline*, randomly selects N sentences.

8.2. Case Study: Mono-document Summarization of Biomedical Scientific Literature

We first evaluate the system performance in the task of generating summaries of biomedical scientific articles. To this end, the summarization method is configured as explained in Chapter 5. As already mentioned, the evaluation includes determining the values of the parameters involved in the algorithm, and comparing the results of a large-scale evaluation on 150 documents from the BioMed Central corpus to those obtained by other summarizers facing the same task.

8.2.1. Algorithm Parametrization

In order to answer the questions raised in Section 8.1.3, different experiments have been performed on a collection of ten documents from the evaluation

[3]MEAD. http://www.summarization.com/mead/. Last accessed: 1 November 2010. We run MEAD using its default parametrization.

corpus. It is important to remind that the model summaries for this parametrization are extracts constructed by domain experts.

8.2.1.1. Determining the Optimal Percentage of Hub Vertices and the Best Set of Semantic Relations

The first group of experiments is directed to find out the best combination of semantic relations for building the document graph (Section 4.4), along with the best percentage of hub vertices for the clustering method (Section 4.5). Note that both parameters need to be evaluated together, as the relations have an influence on the connectivity of the document graph, and so on the optimum percentage of hub vertices. The results of these experiments are presented in Table 8.1. For these experiments, no positional or similarity with the title criteria were used, and the Jaccard similarity coefficient was employed to label the edges in the document graph.

It may be observed from Table 8.1 that both the best set of semantic relations and the percentage of hub vertices depend on the heuristic. Concerning Heuristic 1, it behaves similar when all three semantic relations (i.e. hypernymy, *associated with* and *related to*) are used to build the document graph and the percentage of hub vertices is set to 2 %, to when only the hypernymy and *related to* relations are used along with 5 % of hub vertices. Heuristics 2 and 3 perform better when all three semantic relations are used, but Heuristic 2 obtains the best ROUGE scores when 10 % of vertices are used as hubs, while Heuristic 3 registers the best outcome when the percentage of hub vertices is set to 5 %.

The best overall results are reported by the third heuristic. It may be also observed that, in average, the *associated with* relationship is more effective than the *related to* relation, and this is due to the fact that the *related to* relation links together a relatively low number of concepts, and so produces a quite unconnected document graph. Another interesting result is that, in general, the optimum percentage of hub vertices increases with the number of relations (i.e. with the connectivity of the document graph).

		Heuristic 1				**Heuristic 2**				**Heuristic 3**			
		R-1	**R-2**	**R-W**	**R-S4**	**R-1**	**R-2**	**R-W**	**R-S4**	**R-1**	**R-2**	**R-W**	**R-S4**
Hypernymy	2%	0.7770	*0.6229*	0.2339	*0.5812*	0.7673	*0.6179*	0.2303	*0.5821*	0.7752	*0.6208*	0.2333	*0.6016*
	5%	0.7739	0.5802	0.2328	0.5625	*0.7770*	0.5909	*0.2339*	0.5728	0.7745	0.6093	0.2332	0.5924
	10%	*0.7794*	0.5721	*0.2354*	0.5528	0.7686	0.5916	0.2282	0.5711	*0.7817*	0.6048	*0.2363*	0.5853
	20%	0.7730	0.4791	0.2307	0.4615	0.7636	0.4984	0.2262	0.4794	0.7752	0.5931	0.2309	0.5759
Hypernymy + *Associated with*	2%	*0.7810*	*0.6188*	*0.2355*	*0.6011*	0.7758	0.6012	**0.2348**	0.5977	*0.7796*	0.6148	*0.2355*	0.5960
	5%	0.7129	0.6129	0.2110	0.5950	*0.7589*	*0.6168*	0.2247	0.5831	0.7761	*0.6251*	0.2329	0.6068
	10%	0.7422	0.6168	0.2215	0.5977	0.7564	0.6035	0.2257	*0.6137*	0.7728	0.5425	0.2295	*0.6137*
	20%	0.6680	0.6014	0.1910	0.5858	0.6851	0.5952	0.1970	0.5784	0.7623	0.6070	0.2283	0.5891
Hypernymy + *Related to*	2%	*0.7829*	**0.6275**	*0.2366*	**0.6088**	*0.7734*	0.6143	*0.2332*	*0.5960*	0.7810	0.6223	0.2357	0.6041
	5%	0.7476	0.5730	0.2220	0.5578	0.7585	*0.6151*	0.2260	0.5806	*0.7827*	*0.6234*	*0.2360*	*0.6062*
	10%	0.7474	0.5804	0.2246	0.5635	0.6955	0.5193	0.1993	0.4989	0.7594	0.5888	0.2248	0.5698
	20%	0.6961	0.5181	0.1977	0.5012	0.7296	0.5615	0.2082	0.5391	0.7625	0.5991	0.2295	0.5819
Hypernymy + *Associated with* + *Related to*	2%	0.7752	0.6148	0.2333	0.5960	0.7752	0.6148	*0.2333*	0.5960	0.7749	0.6146	0.2334	0.5962
	5%	**0.7853**	*0.6250*	**0.2371**	*0.6068*	0.7751	0.6099	*0.2333*	0.5941	**0.7886**	**0.6324**	**0.2388**	**0.6151**
	10%	0.7558	0.5960	0.2220	0.5776	**0.7795**	**0.6189**	0.2329	**0.6135**	0.7830	0.6277	0.2338	0.6086
	20%	0.7666	0.6042	0.2317	0.5870	0.7777	0.6166	0.2326	0.5997	0.7791	0.6151	0.2346	0.5981

Table 8.1: ROUGE scores for different combinations of semantic relations and percentages of hub vertices. The best results for each heuristic and set of relations are shown in italic, while the scores in bold indicate the best overall results for each heuristic

8.2.1.2. Determining the Best Criteria for Sentence Selection

The aim of the second group of experiments is to learn if the use of the positional and similarity with the title criteria to select sentences for the summaries helps to improve the content quality of these summaries (see Section 4.7). For these experiments, the percentage of hub vertices was set to 5 % for Heuristics 1 and 3, and to 10 % for Heuristic 2. All semantic relations were used to construct the document graph, and the Jaccard similarity coefficient was used to label the edges in this graph. The ROUGE scores for these tests are presented in Table 8.2, along with the values for parameters λ, θ and χ that define the weight of each criterion in the linear function presented in Equation 4.11.

Heuristic 1							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	**0.7853**	**0.6250**	**0.2371**	**0.6068**
Semantic Graphs + Position	0.9	0.1	0.0	0.7826	0.6202	0.2371	0.5998
Semantic Graphs + Title Similarity	0.9	0.0	0.1	0.7817	0.6233	0.2370	0.6056
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	0.7823	0.6198	0.2371	0.6029
Heuristic 2							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	0.7795	0.6189	0.2329	**0.6135**
Semantic Graphs + Position	0.9	0.1	0.0	0.7804	0.6113	0.2364	0.5967
Semantic Graphs + Title Similarity	0.9	0.0	0.1	0.7786	0.6201	0.2354	0.6048
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	**0.7823**	**0.6225**	**0.2375**	0.5956
Heuristic 3							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	**0.7886**	**0.6324**	**0.2388**	**0.6151**
Semantic Graphs + Position	0.9	0.1	0.0	0.7860	0.6149	0.2385	0.5984
Semantic Graphs + Title Similarity.	0.9	0.0	0.1	0.7801	0.6173	0.2362	0.6038
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	0.7808	0.6171	0.2366	0.6005

Table 8.2: ROUGE scores for different combinations of sentence selection criteria. The best results for each heuristic are shown in bold type

It may be seen from Table 8.2 that the use of the positional and similarity with the title criteria does not benefit Heuristics 1 and 3, but slightly improves the results obtained by the second heuristic. Again, Heuristic 3 behaves better than the other heuristics. Table 8.2 also shows that the similarity with the title criteria contributes more to the quality of the summaries than the positional one for all three heuristics.

Therefore, it may be concluded from Tables 8.1 and 8.2 that the best configuration for Heuristics 1 and 3 implies using the three semantic relations along with a 5 % of hub vertices and no other criterion for sentence selection (i.e. λ =1.0, θ =0.0 and χ =0.0); while the best configuration for Heuristic 2 suggests using the three semantic relations along with a 10 % of hub vertices, and both the sentence position and similarity with the title criteria with weights λ =0.8, θ =0.1 and χ =0.1, respectively. Consequently, these parameter values are used for the remaining experiments.

8.2.1.3. Determining the Best Similarity Coefficient

The aim of the third experiment is to determine which of the two similarity coefficients for weighting the edges in the document graph explained in Section 4.4 (i.e. the Jaccard coefficient and the Dice-Sorensen coefficient) contributes more to the quality of the summaries. To this end, we repeated the best score experiment shown in Table 8.2 using the Dice-Sorensen coefficient. Table 8.3 reports the result of this experiment. It may be observed that the ROUGE scores for all heuristics are lower than those obtained using the Jaccard coefficient. However, the differences are not significant, since both coefficients produce relatively similar weights.

	Criteria	**R-1**	**R-2**	**R-W**	**R-S4**
Heuristic 1	Semantic Graphs	0.7728	0.6143	0.2292	0.5981
Heuristic 2	Semantic Graphs + Position + Title Sim.	0.7704	0.6146	0.2264	0.5928
Heuristic 3	Semantic Graphs	**0.7823**	**0.6285**	**0.2352**	**0.6123**

Table 8.3: ROUGE scores for the optimal combination of sentence selection criteria, using the Dice-Sorensen coefficient to label the edges in the document graph

8.2.2. Evaluating the Effect of Word Ambiguity

We next aim to measure the effect of lexical ambiguity in biomedical text on the quality of the automatic summaries. The need to resolve ambiguity in automatic summarization was introduced in Section 1.2. We improve the summarization system by incorporating different strategies for mapping documents onto concepts in the UMLS Metathesaurus. These strategies have been described in detail in Section 5.2.2 but are recalled here for completeness:

1. Selecting the first candidate mapping returned by MetaMap (*1st Candidate*). Thus, no attempt to solve ambiguity is made.
2. Using the Journal Descriptor Indexing methodology (JDI) (Humphrey et al., 2006) provided by MetaMap, which is invoked using the -y flag (*MetaMap -y*).
3. Using "standard" Personalized PageRank (*PPR*).
4. Using "word-to-word" Personalized PageRank (*PPR-w2w*).

We perform a new evaluation to test the four disambiguation strategies above. For this experiment, we use the best parametrization determined in previous sections for each heuristic and the 150 documents from the BioMed Central corpus. The results are shown in Table 8.4.

Summarizer	**R-1**	**R-2**	**R-W**	**R-S4**
Heuristic 1 - *1st Candidate*	0.7514	0.3304	0.1921	0.3128
Heuristic 2 - *1st Candidate*	0.7305	0.3093	0.1811	0.2856
Heuristic 3 - *1st Candidate*	0.7504	0.3283	0.1915	0.3117
Heuristic 1 - *MetaMap -y*	**0.7724**	**0.3453**	**0.1936**	**0.3189**
Heuristic 2 - *MetaMap -y*	**0.7772**	0.3421	**0.1969**	0.3205
Heuristic 3 - *MetaMap -y*	**0.7845**	**0.3538**	**0.1983**	**0.3267**
Heuristic 1 - *PPR*	0.7692	0.3383	0.1933	0.3150
Heuristic 2 - *PPR*	0.7718	0.3380	0.1935	0.3145
Heuristic 3 - *PPR*	0.7737	0.3419	0.1937	0.3178
Heuristic 1 - *PPR-w2w*	0.7704	0.3379	0.1926	0.3108
Heuristic 2 - *PPR-w2w*	0.7751	**0.3438**	0.1965	**0.3210**
Heuristic 3 - *PPR-w2w*	0.7804	0.3530	0.1966	0.3262

Table 8.4: ROUGE scores for different word sense disambiguation strategies

It may be seen from this table that using WSD improves the average ROUGE scores for the summarizer when compared against the "first candidate" baseline. This improvement is observed for all approaches to WSD

and it is more obvious for Heuristic 2. According to a Wilcoxon Signed Ranks Test ($p < 0.01$), the "standard" version of the Personalized PageRank disambiguation algorithm significantly improves ROUGE-1 and ROUGE-2 metrics for Heuristics 1 and 3, and all ROUGE metrics for Heuristic 2, compared with no WSD (i.e. *1st Candidate*), while the "word-to-word" PPR variant significantly improves all ROUGE metrics for the three heuristics. Results using the JDI algorithm (*MetaMap -y*) are also significantly better than those of the *1st Candidate* strategy for all ROUGE metrics and heuristics. The best WSD strategy for Heuristics 1 and 3 is *MetaMap -y*, whose performance is higher than *PPR* and *PPR-w2w* for all ROUGE metrics. Regarding Heuristic 2, the best strategy is *PPR-w2w*, whose performance is higher than *PPR* and *MetaMap-y* for ROUGE-2 and ROUGE-S4 metrics.

8.2.3. Evaluating the Effect of Expanding Acronyms

We next examine how the presence of non-resolved acronyms and abbreviations in the documents affects to the quality of the automatic summaries. In spite of the possibility of using an *Abbreviations* section in most biomedical scientific publications, it has been observed that most authors define their acronyms and abbreviations *ad hoc* in the document body and do not include them in a separate section. Besides, as these contracted forms are usually non-standard, they do not exist in the UMLS Metathesaurus. As a consequence, when mapping the document to UMLS concepts, MetaMap does not find the occurrences of the concepts associated to these acronyms and abbreviations. It has been empirically observed that the terms (or phrases) represented in an abbreviated form frequently correspond to central concepts in the documents. Thus, when they appear in a sentence, as they cannot be mapped onto UMLS concepts, the sentence "misses" the scores given by those concepts and the opportunities of being selected for the summary are reduced.

To quantify the extent of this problem, we have repeated the best score experiments from Table 8.4, but using the BioText software to replace all abbreviations in the document body with their expansions, as explained in Section 5.2.1. As a consequence, it has been found that resolving these shortened forms improves the quality of the summaries for all three heuristics (see Table 8.5). However, this improvement is not statistically significant.

	R-1	R-2	R-W	R-S4
Heuristic 3	**0.7874**	**0.3560**	**0.2017**	**0.3300**
Heuristic 2	0.7800	0.3440	0.1996	0.3228
Heuristic 1	0.7754	0.3476	0.1953	0.3232

Table 8.5: ROUGE scores for the three heuristics after solving acronyms and abbreviations

8.2.4. Comparison with Other Summarizers

To evaluate the summarization performance, eight different types of summaries have been generated using the three heuristics for sentence selection with their best configurations concluded in Sections 8.2.1, 8.2.2 and 8.2.3, the SUMMA, LexRank and Microsoft Autosummarize systems, and the Lead and Random baselines, as explained in Section 8.1.4. The ROUGE results for all summarizers are presented in Table 8.6.

Summarizer	R-1	R-2	R-W	R-S4
Heuristic 3	**0.7874**	**0.3560**	**0.2017**	**0.3300**
Heuristic 2	0.7800	0.3440	0.1996	0.3228
Heuristic 1	0.7754	0.3476	0.1953	0.3232
LexRank	0.7317	0.3248	0.1873	0.3097
SUMMA	0.7123	0.3187	0.1812	0.2989
AutoSummarize	0.5994	0.2446	0.1380	0.2318
Lead	0.6483	0.2566	0.1621	0.2646
Random	0.4998	0.1777	0.1207	0.2315

Table 8.6: ROUGE scores for different summarizers. The best score for each metric is indicated in bold font. Systems are sorted by decreasing R-2 score

Table 8.6 shows that the three heuristics report significantly better results than the other summarizers and baselines for all ROUGE metrics (Wilcoxon Signed Ranks Test, $p < 0.01$). These results seem to indicate that the use of domain-specific concepts along with a WSD algorithm to solve lexical ambiguity improves the quality of the automatic summaries according to the ROUGE metrics. Concerning comparison between the three heuristics, the performance of Heuristic 3 is better than that of Heuristics 1 and 2 for all ROUGE metrics, but the differences are not statistically significant.

8.2.5. Discussion

The results in Table 8.6 demonstrate that using domain-specific knowledge improves summarization performance compared to traditional word-based

approaches, in terms of the informative content quality of the summaries that are generated. The use of concepts instead of terms, along with the semantic relations that exist between them, allows the system to identify the different topics covered in the text more accurately, and with relative independence of the vocabulary used to describe them. As a consequence, the information in the sentences selected for the summaries is closer to the model abstracts.

On the other hand, a close study of the abstracts of 50 documents from the evaluation set has revealed that the information considered as important by the authors of these documents may be classified in three main sections or categories: (1) the background of the study, (2) the method or case presentation and (3) the results and conclusions of the study. The method presentation section includes approximately 58 % of the information in the abstract, the results and conclusions section comprises around 25 %, and the background section involves less than 17 %. It has been also observed that the clustering method usually produces a single large cluster together with a variable number of small clusters. Even though some of the concepts within the large cluster may be found in all three sections of the abstract, the majority of the concepts in this cluster are found in the section describing the method. Therefore, it seems clear that any heuristic for sentence selection aiming to compare well with the authorsábstracts should include mainly information related to the concepts within this large cluster, but also some other secondary information. Hence, Heuristic 3 is, by definition, at advantage compared to Heuristics 1 and 2 when the abstracts of the documents are used as model summaries.

In spite of this, the differences among the heuristics are not as remarkable as expected. A careful analysis of the summaries generated by the three heuristics suggests that the explanation for this finding is that, given the larger size of the main cluster, the three heuristics extract most of their sentences from this cluster, and hence the summaries generated have most of the sentences in common. Nevertheless, the best results are reported by Heuristic 3. It has been checked that this heuristic selects most of the sentences from the most populated cluster, but it also includes some sentences from other clusters when such sentences assign high scores to them. Thus, in addition to the information related to the central topic, this heuristic also includes other secondary or “satellite” information that might be relevant to

the user. On the contrary, Heuristic 1 fails to present secondary information; while Heuristic 2 includes more secondary information, but may leave out some of the core information.

Finally, an important research question that arises when examining these results is why the ROUGE scores differ so much across different documents. This is not shown in the tables (as they present the average results) but has been observed during the experimentation and can be appreciated in Table 8.7. This table shows the standard deviation of the different ROUGE scores for the summaries generated by the third heuristic.

Metric	**Standard Deviation**
ROUGE-1	0.0813
ROUGE-2	0.1228
ROUGE-W-1.2	0.0497
ROUGE-S4	0.1074

Table 8.7: Standard deviation of ROUGE scores for the summaries generated using Heuristic 3

In order to clarify the reasons for these differences, the two extreme cases (that is, the two documents with the highest and lowest ROUGE scores respectively) were carefully examined. The best case turned out to be one of the largest document in the corpus, while the worst case was one of the shortest (six pages *vs.* three pages). According to the starting hypothesis (i.e. the document graph is an instance of scale-free network) it is not surprising that the summarization algorithm works better with larger documents, since the scale-free distribution is more likely to emerge in large networks. A second interesting difference between both documents is their underlying subject matters. The best case is published in the *BMC Biochemistry* journal, and concerns the reactions of some kind of proteins over the brain synaptic membranes. In contrast, the worst case is published in the *BMC Bioinformatics* journal, and regards the use of pattern matching for database searching. It has been verified that the UMLS covers better the vocabulary in the first document than in the second one, in terms of both concepts and relations, which leads to a more accurate graph that better reflects the content of the document. It has also been observed that the best case contains a significant number of bigrams, trigrams and even 4-grams that map to single concepts in the UMLS, both in the document body and in the abstract, which obviously has a positive influence on its ROUGE results.

Finally, in the worst-case document, the use of synonyms is quite frequent, which does not occur in the best-case document. For instance, the same concept is referred to in the document body as *string searching*, while it is always referred to as *pattern matching* in the abstract. Since the ROUGE metrics are based on the number of word overlaps, the summaries containing synonyms of the terms in the abstracts are unreasonably penalized.

8.3. Case Study: Mono-document Summarization of News Articles

We next assess the system in the task of generating summaries of news articles. To this end, the summarization algorithm is configured as explained in Chapter 6.

8.3.1. Algorithm Parametrization

We first accomplish a preliminary evaluation aiming at determining the optimal values for the different parameters of the algorithm (Section 8.1.3). This evaluation is performed on a collection of 10 documents from the DUC 2002 corpus, as explained in Section 8.1.2. Again, the model summaries for this parametrization are human-made extracts.

8.3.1.1. Determining the Optimal Percentage of Hub Vertices and the Best Similarity Threshold

The first group of experiments is directed to determine the best percentage of hub vertices for the clustering method (Section 4.5), along with the optimal similarity threshold to build the document graph (Section 6.2.4). Thus, all semantic relations (i.e. hypernymy and semantic similarity) are used to build this graph. For these experiments, the *jcn* algorithm (Section 3.2.2) was used to compute the semantic similarity between WordNet concepts, no positional or similarity with the title criteria were used, and the Jaccard coefficient was employed to label the edges in the document graph. These results are shown in Table 8.8.

		Heuristic 1				Heuristic 2				Heuristic 3			
		R-1	**R-2**	**R-W**	**R-S4**	**R-1**	**R-2**	**R-W**	**R-S4**	**R-1**	**R-2**	**R-W**	**R-S4**
0.01	2 %	**0.5562**	**0.2565**	**0.1898**	**0.2261**	0.5362	0.2365	0.1798	0.2061	0.5405	0.2360	0.1850	0.2079
	5 %	0.5307	0.2288	0.1816	0.2054	0.5231	0.2206	0.1788	0.1982	**0.5608**	**0.2565**	**0.1910**	**0.2260**
	10 %	0.5307	0.2288	0.1816	0.2054	0.5231	0.2206	0.1788	0.1982	0.5405	0.2360	0.1850	0.2079
	20 %	0.4811	0.1914	0.1661	0.1691	*0.5455*	*0.2415*	*0.1841*	*0.2171*	0.5464	0.2509	0.1906	0.2200
0.05	2 %	*0.5438*	*0.2513*	*0.1868*	*0.2232*	0.5403	0.2422	0.1885	0.2132	0.5400	0.2324	0.1842	0.2058
	5 %	0.5305	0.2285	0.1805	0.2064	0.5279	0.2225	0.1812	0.1968	*0.5428*	*0.2475*	*0.1861*	*0.2219*
	10 %	0.4915	0.1921	0.1633	0.1715	0.5391	0.2060	0.1784	0.1848	0.5346	0.2320	0.1826	0.2024
	20 %	0.4426	0.1628	0.1543	0.1499	*0.5463*	**0.2455**	*0.1895*	**0.2184**	0.5285	0.2387	0.1822	0.2067
0.1	2 %	0.5205	0.2244	0.1775	0.1983	0.5485	0.2141	*0.1881*	0.1956	0.5438	0.2414	0.1875	0.2133
	5 %	*0.5347*	*0.2353*	*0.1852*	*0.2081*	0.5125	0.2097	0.1745	0.1865	*0.5544*	*0.2472*	*0.1885*	*0.2164*
	10 %	0.4592	0.1682	0.1517	0.1547	0.5421	0.2378	0.1874	0.2084	0.5216	0.2384	0.1808	0.2076
	20 %	0.5104	0.1950	0.1718	0.1783	*0.5488*	*0.2405*	0.1865	*0.2131*	0.5271	0.2305	0.1822	0.2007
0.2	2 %	*0.5419*	*0.2384*	*0.1872*	*0.2116*	0.5419	0.2336	0.1822	0.2060	0.5419	0.2384	0.1872	0.2116
	5 %	0.4723	0.1638	0.1581	0.1571	0.5371	0.2179	0.1760	0.1872	*0.5500*	*0.2539*	*0.1891*	*0.2251*
	10 %	0.4723	0.1638	0.1581	0.1571	**0.5524**	*0.2437*	**0.1950**	0.2161	*0.5500*	*0.2539*	*0.1891*	*0.2251*
	20 %	0.4723	0.1638	0.1581	0.1571	*0.5483*	*0.2204*	0.1911	0.1948	*0.5500*	*0.2539*	*0.1891*	*0.2251*

Table 8.8: ROUGE scores for different combinations of similarity thresholds and percentages of hub vertices. The best results for each heuristic and set of relations are shown in italic, while the scores in bold indicate the best results for each heuristic

It may be observed from Table 8.8 that the best performance configuration for Heuristic 1 implies using 2 % of vertices in the graph as hubs and 0.01 as similarity threshold. In average, raising the number of *hub vertices* decreases the ROUGE scores. The reason seems to be that, since news articles usually include little redundancy, when the number of concepts used as centroids is excessively high, the clustering method attaches importance to concepts which are in fact secondary or unrelated to the main topic. Concerning Heuristic 2, the best results are achieved using 20 % of vertices as hubs, and 0.05 as similarity threshold. Thus, this heuristic prefers a larger number of hub vertices. Finally, Heuristic 3 performs better when an intermediate percentage of hub vertices is used (5 %) and the similarity threshold is set to 0.01.

8.3.1.2. Determining the Best Set of Semantic Relations

The aim of the second experiment is to find the best combination of relations for building the document graph (Section 4.4). Since the use of both relations (ie. hypernymy and semantic similarity) has been already evaluated in the previous section, we just repeat here the experiments in Table 8.8 but using just the hypernymy relation. The results are presented in Table 8.9.

Heuristic 1					
		R-1	**R-2**	**R-W**	**R-S4**
Hypernymy	2 %	0.5387	0.2209	0.1840	0.2102
	5 %	0.4748	0.1883	0.1635	0.1746
	10 %	0.5235	0.2410	0.1836	0.2084
	20 %	0.5176	0.2045	0.1760	0.1804
Hypernymy + Sem. Sim.		**0.5562**	**0.2565**	**0.1898**	**0.2261**
Heuristic 2					
		R-1	**R-2**	**R-W**	**R-S4**
Hypernymy	2 %	0.5186	0.2237	0.1793	0.1984
	5 %	0.5034	0.2143	0.1746	0.1902
	10 %	0.5590	0.2291	0.1884	0.1970
	20 %	*0.6063*	*0.2405*	*0.1949*	*0.2095*
Hypernymy + Sem. Sim.		**0.5463**	**0.2455**	**0.1895**	**0.2184**
Heuristic 3					
		R-1	**R-2**	**R-W**	**R-S4**
Hypernymy	2 %	*0.5357*	*0.2479*	*0.1824*	*0.2181*
	5 %	0.5152	0.2181	0.1769	0.1937
	10 %	0.5326	0.2312	0.1832	0.2026
	20 %	0.5411	0.2395	0.1860	0.2082
Hypernymy + Sem. Sim.		**0.5608**	**0.2565**	**0.1910**	**0.2260**

Table 8.9: Best combination of semantic relations for each heuristic

It can be seen from this table that, for the three heuristics, the ROUGE scores decrease when only the hypernymy relation is used, and this is due to the fact that this relation links together a quite low number of concepts, producing an excessively unconnected document graph.

8.3.1.3. Determining the Best Criteria for Sentence Selection

The third experiment aims to learn if the use of the positional and similarity with the title criteria for sentence selection helps to improve the content quality of the summaries (see Section 4.7). For these experiments, the percentage of hub vertices was set to 2 % for Heuristic 1, 20 % for Heuristic 2 and 5 % for Heuristics 3, and the similarity threshold was set to 0.01 for Heuristics 1 and 3, and to 0.05 for Heuristic 2. The Jaccard coefficient was used to label the edges in the document graph.

Heuristic 1							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	0.5562	0.2565	0.1898	0.2261
Semantic Graphs + Position	0.9	0.1	0.0	**0.5593**	**0.2585**	**0.1921**	**0.2332**
Semantic Graphs + Title Similarity	0.9	0.0	0.1	0.5359	0.2358	0.1816	0.2089
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	0.5526	0.2548	0.1906	0.2241
Heuristic 2							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	0.5463	0.2455	0.1895	0.2184
Semantic Graphs + Position	0.9	0.1	0.0	**0.5592**	**0.2596**	**0.1921**	**0.2286**
Semantic Graphs + Title Similarity	0.9	0.0	0.1	0.5248	0.2320	0.1779	0.2054
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	0.5383	0.2481	0.1850	0.2214
Heuristic 3							
	λ	θ	χ	**R-1**	**R-2**	**R-W**	**R-S4**
Semantic Graphs	1.0	0.0	0.0	0.5608	0.2565	0.1910	0.2260
Semantic Graphs + Position	0.9	0.1	0.0	**0.5612**	**0.2596**	**0.1921**	**0.2286**
Semantic Graphs + Title Similarity.	0.9	0.0	0.1	0.5359	0.2358	0.1816	0.2089
Semantic Graphs + Position +Title Similarity	0.8	0.1	0.1	0.5526	0.2548	0.1906	0.2271

Table 8.10: ROUGE scores for different combinations of sentence selection criteria. The best results for each heuristic are shown in bold type

The ROUGE scores for these tests are presented in Table 8.10, along with the values for the parameters λ, θ and χ that define the weight of each criterion. It may be seen from this table that the use of the similarity with the title criterion does not benefit any heuristic. In contrast, the use of the positional criterion together with the semantic graph-based approach slightly improves the results obtained by all heuristics and achieves better ROUGE scores than any other combination of sentence selection criteria. This result was expected since, as studied in Section 6.1, the information in news articles is usually presented according to the *inverted pyramid* form, so that the most important information is placed first.

8.3.1.4. Determining the Best Similarity Coefficient

We next study which of the two similarity coefficients for weighting the edges in the document graph (i.e. the Jaccard coefficient and the Dice-Sorensen coefficient) contributes more to the quality of the summaries. To this end, we repeat the experiment shown in Table 8.8 using the Dice-Sorensen coefficient rather than the Jaccard one.

Table 8.11 reports the ROUGE scores for this experiment. It may be observed that the results for all heuristics are lower than those obtained using the Jaccard coefficient. However, as in the previous case study, the differences are not statistically significant.

	Criteria	**R-1**	**R-2**	**R-W**	**R-S4**
Heuristic 1	Semantic Graphs + Position	0.5523	0.2534	0.1896	0.2277
Heuristic 2	Semantic Graphs + Position	0.5517	0.2522	0.1888	0.2265
Heuristic 3	Semantic Graphs + Position	**0.5586**	**0.2543**	**0.1904**	**0.2265**

Table 8.11: ROUGE scores for the optimal combination of sentence selection criteria, using the Dice-Sorensen coefficient to label the edges in the document graph

8.3.2. Evaluating the Effect of Word Ambiguity

A further experiment has been conducted to examine the effect of lexical ambiguity on the results reported by the three heuristics. To do this, we improve the summarization system by incorporating different strategies for mapping the text onto WordNet concepts. These strategies have been described in Section 6.2.2, but are recalled here for completeness:

1. No attempt is made to solve ambiguity. Instead, the WordNet sense for each term is selected randomly (*Random*).

2. Assigning to each word its first sense in WordNet, which is also its most frequent sense (*1st Sense*).

3. Using the Lesk WSD algorithm (Lesk, 1986) (*Lesk*)

Table 8.12 shows that, as in the biomedical domain, the use of word disambiguation improves the quality of the automatic summaries. According to the Wilcoxon Signed Ranks Test (p <0.05) both the *1st Sense* and *Lesk* disambiguation strategies produce significantly better results than the *Random* strategy for all ROUGE metrics and heuristics. However, when comparing the results using *Lesk* to those obtained using the *1st Sense* strategy, the improvement is less than expected. The reason seems to be that the first WordNet sense criterion is a quite pertinent one, since the senses of the words in WordNet are ranked according to their frequency. Besides, the Lesk algorithm introduces some noise (according to Banerjee and Pedersen (2002), it presents an average accuracy of 32 %), and it is biased toward the first WordNet sense. In fact, it has been checked that the percentage of concepts in the DUC 2002 corpus that Lesk labels with the first sense is above 61 %. Therefore, the difference among the disambiguation performed by both criteria is not too marked.

Summarizer	**R-1**	**R-2**	**R-W**	**R-S4**
Heuristic 1 - *Random*	0.4214	0.1932	0.1503	0.1691
Heuristic 2 - *Random*	0.4253	0.1972	0.1555	0.1713
Heuristic 3 - *Random*	0.4322	0.2001	0.1576	0.1780
Heuristic 1 - *1st Sense*	0.4584	0.2057	0.1626	0.1794
Heuristic 2 - *1st Sense*	0.4594	0.2074	0.1631	0.1810
Heuristic 3 - *1st Sense*	0.4619	0.2104	0.1643	0.1838
Heuristic 1 - *lesk*	0.4641	0.2191	0.1647	0.1919
Heuristic 2 - *lesk*	**0.4651**	0.2193	0.1650	0.1927
Heuristic 3 - *lesk*	0.4648	**0.2196**	**0.1652**	**0.1928**

Table 8.12: ROUGE scores for different word sense disambiguation strategies

8.3.3. Comparison with Other Summarizers

Table 8.13 shows the ROUGE scores for the summaries generated using the three versions of our system with their best configurations concluded in

Sections 8.3.1 and 8.3.2, LeLSA+AR, Freq+TextEnt, LexRank, SUMMA, Microsoft AutoSummarize and the 5 systems which participated in DUC-2002 and achieved the best results (in terms of the ROUGE metric). We also show two baselines (Lead and Random). All these systems were explained in Section 8.1.4. Automatic summaries are generated by selecting sentences until the summary length reaches 100 words.

Summarizer	**R-1**	**R-2**	**R-L**	**R-S4**
Heuristic 3	0.4648	**0.2196**	**0.4277**	**0.1928**
Heuristic 2	**0.4651**	0.2193	0.4276	0.1927
Heuristic 1	0.4641	0.2191	0.4268	0.1919
LexRank	0.4558	0.2115	0.4173	0.1846
Freq+TextEnt	0.4518	0.1942	0.4104	-
LeLSA+AR	0.4228	0.2074	0.3928	0.1661
DUC 28	0.4278	0.2177	0.3865	0.1732
SUMMA	0.4217	0.1952	0.3876	0.1516
AutoSummarize	0.4216	0.1887	0.3671	0.1429
Lead Baseline	0.4113	0.2108	0.3754	0.1660
DUC 19	0.4082	0.2088	0.3735	0.1638
DUC 27	0.4052	0.2022	0.3691	0.1600
DUC 21	0.4149	0.2104	0.3754	0.1655
DUC 29	0.3993	0.2006	0.3617	0.1576
Random Baseline	0.2996	0.1110	0.2795	0.0900

Table 8.13: ROUGE scores for different summarizers. The best score for each metric is indicated in bold font. Systems are sorted by decreasing R-2 score

According to a Wilcoxon Signed Ranks Test (p <0.05), the three heuristics report significantly better results than both baselines, LexRank and all systems participating in the DUC 2002 conference. In contrast, no significant differences exist among the three heuristics. Regarding the anaphoric and textual entailment approaches, as we only know their average ROUGE scores, we could not apply the Wilcoxon test for these systems. However, the three versions of our summarizer outperform LeLSA+AR in all ROUGE scores, and Freq+TextEnt in ROUGE-1, ROUGE-2 and ROUGE-W scores (the ROUGE- S4 score is not available).

8.3.4. Discussion

Once again, the results in Table 8.13 demonstrate that the use of concept graphs improves the quality of automatic summaries compared to traditional term-based approaches. As in the previous case study (biomedical scientific literature summarization), the best performance is obtained when the

Heuristic 3 is used to select the sentences for the summary.

However, in contrast to biomedical summarization, the use of the positional criterion together with our graph-based approach has a positive impact on the quality of the summaries that are generated. This result was expected, since the information in news articles usually follows the *inverted pyramid* structure. This property is not usually satisfied in scientific papers, where the first sentences usually introduce the problem or motivation, but the most important information is presented in the middle sentences of the document, as part of the *Method* and *Discussion* sections. On the other hand, as in the biomedical case study, the similarity with the title criterion does not benefit the selection of sentences.

We have also found that applying word sense disambiguation improves the performance of our summarization method but, as already mentioned, the disambiguation algorithm introduces some noise in the concept recognition, which in turns affects the subtheme identification step. As a result, the improvement achieved by using a WSD algorithm is less than expected.

Another coincidence with the previous case study is that the differences between the three heuristics (both with and without WSD) are insignificant. In order to understand the reason, we examined the intermediate results of our algorithm and found, once again, that the clustering method usually produces one big cluster along with a variable number of small clusters. As a consequence, the three heuristics extract most of their sentences from this large cluster, and therefore the summaries generated are quite similar.

Finally, once again the ROUGE scores differ importantly across different documents. The reason seems to be the variety in size of the different news articles, the worst cases being the shortest in the corpus.

8.4. Case Study: Multi-document Summarization of Tourism Websites

We finally evaluate the method in the task of generating summaries from multiple web pages describing a same tourist destination (see Section 7.1). As already mentioned, in this case study we do not perform any process to determine the appropriated values for the parameters of the algorithm, but evaluate the method using the best parametrization obtained for the news summarization case study. The reason is that we aim to study whether the

algorithm may be applied to new (but relatively similar) types of documents without having to accomplish a new parametrization process.

Like in previous case studies, we use ROUGE to evaluate the informativeness of the automatic summaries, but here this evaluation is completed with a manual readability assessment.

8.4.1. Comparison with Other Summarizers

We compare our system to other well-known multi-document summarizers. Table 8.14 shows the ROUGE scores for the summaries generated using Heuristics 1, 2 and 3, SUMMA, MEAD, COMPENDIUM and Language Models. All these systems were explained in Section 8.1.4. Automatic summaries are generated by selecting sentences until the summary length reaches 200 words. We only show the results for ROUGE-2 and ROUGE-S4 metrics, since only these metrics are available for COMPENDIUM and Language Models.

Summarizer	**R-2**	**R-S4**
Heuristic 3	**0.090**	**0.143**
Heuristic 1	0.089	0.139
MEAD	0.089	0.138
COMPENDIUM	0.086	0.134
Language Models	0.071	0.119
Heuristic 2	0.069	0.117
SUMMA	0.064	0.109

Table 8.14: ROUGE scores for different summarizers. The best score for each metric is indicated in bold font. Systems are sorted by decreasing R-2 score

It can be observed from Table 8.14 that Heuristic 3 behaves better than the remaining systems for all ROUGE metrics. According to a pairwise Wilcoxon Signed Ranks Test (p <0.01), MEAD and Heuristics 1 and 3 produce significantly better summaries than SUMMA, Heuristic 2, COMPENDIUM and Language Models for both ROUGE-2 and ROUGE-SU4. However, no differences exist between Heuristics 1 and 3 and with respect to MEAD. In contrast, Heuristic 2 behaves considerably worse than in the other case studies. The reason seems to be that the documents to summarize contain a wide variety of information regarding very distinct topics. The result is that the clustering algorithm usually produce a relatively large number of clusters, and, since Heuristic 2 selects the sentences for the summary from all the clusters, it includes a lot of secondary information.

8.4.2. Readability Evaluation

We also perform a manual readability evaluation, following the guidelines observed in DUC 2005 and 2006, as explained in Section 8.1.1. Table 8.15 reports the average results for the summaries generated using the third heuristic and regarding each readability criterion. We also show the results of the systems that participated in the DUC 2006 competition and obtained the best result in each criterion.

Criteria	**Heuristic 3**	**DUC 2006**
Grammaticality	4.11	4.62
Redundancy	3.8	4.0
Clarity	3.72	4.66
Focus	4.1	4.28
Coherence	3.15	3.28

Table 8.15: Average result for each readability criterion

As it may be seen from Table 8.15, although our results are poorer than those obtained by the DUC 2006 systems, it is worth mentioning that these systems were evaluated in a query-oriented summarization task, which means that they had information about what the users expected to find in the summaries. This kind of information has been shown to improve significantly the quality of automatic summaries. On the other hand, the worst results are obtained for the clarity and coherence criteria.

8.4.3. Discussion

The results of the evaluation demonstrate, first, that the summarization method may be easily adapted to summarize multiple documents dealing with the same theme or topic, since it only requires the use of a redundancy detection algorithm in order to remove the information that is repeated across documents; and second, that the summarizer may be applied to new (but similar) domains and document types without accomplishing a new parametrization process.

As in the previous case studies, it has been observed that the ROUGE scores differ significantly across documents. In particular, the main problem is directly related to the type of the documents to summarize: in most of these documents, the salient information is concerned with proper nouns describing monuments, cities, beaches, etc., that are not likely to be found

in WordNet (e.g. *Sacre Coeur*, *Santorini* or *Ipanema*). If no concept is found in the database for these terms, the document graph will be inevitably losing essential information to identify the topics covered in the document.

Concerning the readability assessment, we found that our approach performs good in almost all criteria except for clarity and coherence. These two criteria should be improved in future work by using anaphora resolution (Steinberger et al., 2007) and sentence simplification (Barzilay y McKeown, 2005; Filippova y Strube, 2008) techniques.

Chapter 9

Conclusion and Future Work

This work presents a generic graph-based method for generating textual summaries. The main contribution is the use of specific knowledge about the domain of the documents and their structure to improve the quality of the final summaries. In order to make the system general with regard to the application domain, the summarization algorithm may be easily configured to work on new types of documents (both in terms of their structure and domain). The summarizer is intended for single-document summarization, but may be adapted to summarize multiple documents on the same topic by adding a post-processing step to remove the redundancy in the summary.

The method proposed represents the document as a semantic graph, using concepts and relations from a knowledge database. This way it gets a richer representation than the one provided by traditional models based on terms. A degree-based clustering algorithm is then used to discover different themes or topics within the text. Three different heuristics for sentence selection have been proposed, each of them aiming to construct a different type of summary according to the type of information in the source that is likely to be included in the summary.

This work also deals with two problems that should be taken into account when generating automatic summaries, but have been little explored to date: the presence of lexical ambiguity and acronyms within the text to be summarized. As studied in the introductory section, these problems may reduce the quality of the automatic summaries. In fact, the results presented in the evaluation section demonstrate that using word sense disambiguation improves the summarization performance, particularly when dealing with highly specialized domains. The WSD algorithms are able to more accura-

tely identify the concepts that are being referred to in the document than when the typical first mapping approach is used and this leads to the creation of a graph that more accurately reflects the content of the document. As a result, the clustering method is better able to identify the topics covered in the document, and the information in the sentences selected for the summary is closer to that presented in the model summaries. However, this improvement is less than expected and this is probably due to errors made by the disambiguation algorithms. On the other hand, we have also examined the effect of non-resolved acronyms and abbreviations in summarization and found that using a simple software to automatically identify these shortened forms and their expansions from the document, and replace any occurrences of the abbreviations in the document with their corresponding expansions slightly improves the summarization performance.

However, the readability assessment performed for the summaries generated in the third case study has shown that the automatic summaries suffer from a lack of referential clarity and coherence. The first problem is related to the occurrence of pronouns and noun phrases for which the entities being referenced or their relation to the story remains unclear. The second problem is related to the lack of structure and organization of the summary, which is usually consequence of gaps or lacking information between sentences. To solve both aspects, future work will concentrate on resolving anaphoric references as a previous step to summarization (Steinberger et al., 2007), and revising the summary to ensure that, if a sentence containing an anaphora is added to the summary, then the sentence containing the referent of such anaphora is also included. On the other hand, we also plan to study the viability of applying sentence compression and aggregation techniques to reduce the length of the summaries that are generated, while retaining the relevant information content (Barzilay y McKeown, 2005; Filippova y Strube, 2008).

Finally, future work will also study the production of query-oriented summaries. Whereas generic summaries address a broad readership, query-based summaries are preferred by specific groups of people aiming for quick knowledge gain about specific topics (Mani, 2001a). This is because the query itself may contain important hints as to what the user is interested in. The summarization method presented in this thesis may be easily adapted to deal with query-based summarization. It suffices to modify the function

for computing the weight of the edges in the document graph, so that if an edge is linked to a vertex that represents a concept which is also present in the query, then the weight of the edge is increased. This weight is distributed through the graph and the vertices representing concepts from the query and those other concepts connected to them are assigned a higher salience and ranked higher. As a result, the sentences containing concepts closely related in meaning to those in the query increase their probability to be selected for the summary.

References

Afantenos, S.D., V. Karkaletsis, and P. Stamatopoulos. 2005. Summarization from Medical Documents: A Survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

Agirre, E. and P. Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Aplications*. Springer.

Agirre, E. and A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Aker, A. and R. Gaizauskas. 2010. Generating Image Descriptions using Dependency Relational Patterns. In *Proceedings of the Association of Computational Linguistic*, pages 1250–1258.

Amini, M-R. and P. Gallinari. 2002. The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–112.

Aone, C., M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1999. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automated Text Summarization*. The MIT Press, pages 71–80.

Aronson, A. R. and F-M. Lang. 2010. An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association*, 17:229–236.

Aronson, A.R. and T.C. Rindflesch. 1997. Query Expansion Using the UMLS Metathesaurus. In *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 485–489.

Banerjee, S. and T. Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145.

Barabási, A.L. and R. Albert. 1999. Emergence of Scaling in Random Networks. *Science*, 268:509–512.

Barzilay, R. and M. Elhadad. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the Association for Computational Linguistics, Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Barzilay, R. and K. R. McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–327.

Bawakid, A. and M. Oussalah. 2008. A Semantic Summarization System: University of Birmingham at TAC 2008. In *Proceedings of the First Text Analysis Conference.*

Baxendale, P.B. 1958. Man-Made Index for Technical Literature: An Experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Binwahlan, M. S., N. Salim, and L. Suanmali. 2009. Swarm Based Features Selection for Text Summarization. *International Journal of Computer Science and Network Security*, 9(1):175–179.

Borko, H. and C. Bernier. 1975. *Abstracting Concepts and Methods.* Academic Press, New York.

Bossard, A., M. Généreux, and T. Poibeau. 2008. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In *Proceedings of the 1st Text Analysis Conference.*

Brandow, R., K. Mitze, and L. F. Rau. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, 5(31):675–685.

Brin, S. and L. Page. 1998. The Anatomy of a Largescale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30:1–7.

Carbonell, J., Y. Geng, and J. Goldstein. 1997. Automated Query-relevant Summarization and Diversity-based Reranking. In *Proceedings of the International Joint Conferences on Artificial Intelligence, Workshop on Artificial Intelligence in Digital Libraries*, pages 12–19.

Chuang, W. T. and J. Yang. 2000. Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 454–457.

Dang, H.T. 2005. Overview of DUC 2005. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Document Understanding Conference 2005 Workshop.*

Dang, H.T. 2006. Overview of DUC 2006. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Document Understanding Conference 2006 Workshop.*

DeJong, G.F. 1982. An Overview of the FRUMP System. In *Strategies for Natural Language Processing.* Lawrence Erlbaum, pages 149–176.

Donaway, R. L., K. W. Drummey, and L. A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic Summarization*, pages 69–78.

Eck, M., S. Vogel, and A. Waibel. 2004. Improving Statistical Machine Translation in the Medical Domain Using the Unified Medical Language System. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 792–798.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 2(16):264–285.

Erkan, G. and D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Ferrández, O., D. Micol, R. Muñoz, and M. Palomar. 2007. A Perspective-Based Approach for Solving Textual Entailment Recognition. In *Proceedings of the Association for Computational Linguistics, PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 66–71.

Filippova, K. and M. Strube. 2008. Sentence Fusion via Dependency Graph Compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 177–185.

Fiszman, M., T. C. Rindesch, and H. Kilicoglu. 2004. Abstraction Summarization for Managing the Biomedical Research Literature. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83.

Fum, D., G. Gmda, and C. Tasso. 1985. Evaluating Importance: A Step Towards Text Summarization. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 840–844.

Futrelle, R., 1999. *Summarization of Diagrams in Documents*, chapter in Advances in Automatic Text Summarization, pages 403–421. The MIT Press.

Gao, Y., W-B. Wang, J-H. Yong, and H-J. Gu. 2009. Dynamic Video Summarization Using Two-level Redundancy Detection. *Multimedia Tools and Application*, 42:233–250.

Halliday, M. 1985. *An Introduction to Functional Grammar*. Edward Arnold.

Halliday, M. and R. Hasan. 1996. *Cohesion in English*. Longsmans.

Hirst, G. and D. St Onge, 1998. *Lexical Chains as Representation of Context for the Detection and Correction Malapropisms*. The MIT Press.

Hobbs, J. 1985. On the Coherence and Structure of Discourse. *CSLI Technical Report*, pages 85–37.

Hong, R., J. Tang, H-K. Tan, S. Yan, C. Ngo, and T-S. Chua. 2009. Event Driven Summarization for Web Videos. In *Proceedings of the 1st Conference of the Special Interest Group on Multimedia, Workshop on Social Media*, pages 43–48.

Hovy, E. 2005. Automated Text Summarisation. In *Handbook of Computational Linguistics*. Oxford University Press.

Hovy, E. and C-Y. Lin. 1999. *Automated Text Summarization in SUMMARIST*. MIT Press.

Hovy, E., C-Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 Using Basic Elements. In *Proceedings of Document Understanding Conference*.

Hsueh, P-Y. and J. D. Moore. 2009. Improving Meeting Summarization by Focusing on User Needs: A Task-oriented Evaluation. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, pages 17–26.

Humphrey, S. M., W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch. 2006. Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113.

Ide, N. and J. Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40.

Jaccard, P. 1901. Étude Comparative de la Distribution Florale Dans une Portion des Alpes et des Jura. *Bulletin de la Societé Vaudoise des Sciences Naturelles*, 37:547–579.

Jiang, J. J. and D. W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference Research on Computational Linguistics*, pages 19–33.

Kleinberg, J. 1999. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632.

Kupiec, J., J. O. Pedersen, and F. Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73.

Leacock, C. and M. Chodorow, 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, chapter in 11, pages 265–283. The MIT Press.

Legendre, P. and L. Legendre. 1998. *Numerical Ecology*. The Netherlands, Amsterdam.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from a Ice Cream Cone. In *Proceedings of Special Interest Group on Design of Communication*, pages 24–26.

Levenshtein, V. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Lin, C-Y. 2004a. Looking for a Few Good Metric: Automatic Summarization Evaluation - How Many Samples are Enough? In *Proceedings of the NII Test Collection for Information Retrieval Systems, Workshop 4*.

Lin, C-Y. 2004b. Rouge: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Association for Computational Linguistics, Workshop: Text Summarization Branches Out*, pages 74–81.

Lin, C-Y. and E. Hovy. 1997. Identifying Topic by Position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 283–290.

Lin, C-Y. and E. Hovy. 2002. Manual and Automatic Evaluation of Summaries. In *Proceedings of the Document Understanding Conference, Workshop on Automatic Summarization*, pages 45–51.

Lin, C-Y and E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 71–78.

Lin, D. 1998. An Information-theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

Litvak, M. and M. Last. 2008. Graph-based Keyword Extraction for Single-document Summarization. In *Proceedings of the International Conference on Computational Linguistics, Workshop on Multi-source Multilingual Information Extraction and Summarization*.

Lloret, E., O. Ferrández, R. Muñoz, and M. Palomar. 2008. A Text Summarization Approach under the Influence of Textual Entailment. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science in Conjunction with the 10th International Conference on Enterprise Information Systems*, pages 22–31.

Longacre, R. 1979. The Discourse Structure of the Flood Narrative. *Journal of the American Academy of Religion*, 47(1):89–133.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Mani, I. 2001a. *Automatic Summarization*. Jonh Benjamins Publishing Company.

Mani, I. 2001b. Summarization Evaluation: An Overview. In *Proceedings of the Second NII Test Collection for Information Retrieval Systems Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Mani, I. and E. Bloedorn. 1999. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1(1-2):35–67.

Mann, W. and S. Thompson. 1988. Rethorical Structure Theory: Towards a Functional Theory of Text Organisation. *Text*, 8(3):243–281.

Marcu, D., 1999. *Advances in Automatic Text Summarization*, chapter in Discourse Trees Are Good Indicators of Importance in Text, pages 123–136. The MIT Press.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

McInnes, B., T. Pedersen, and J. Carlis. 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 533–537.

Metzler, D. and T. Kanungo. 2008. Machine Learned Sentence Selection Strategies for Query-Biased Summarization. In *Proceedings of the Special Interest Group on Information Retrieval Conference, Learning to Rank for Information Retrieval Workshop*.

Mihalcea, R. and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 404–411.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1998. Five Papers on WordNet. In *WordNet: An Electronic Lexical Database*. The MIT Press.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Morris, J. and G. Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

Nadkarni, P.M. 2000. Information Retrieval in Medicine: Overview and Applications. *Journal of Postgraduate Medicine*, 46(2):122–166.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R. and M. Lapata. 2007. Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1683–1688.

Nenkova, A. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proceedings of the 20th National Conference on Artificial Intelligence*, volume 3, pages 1436–1441.

Nenkova, A., R. Passonneau, and K. McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Overby, C. L., P. Tarczy-Hornoch, and D. Demner-Fushman. 2009. The Potential for Automated Question Answering in the Context of Genomic Medicine: An Assessment of Existing Resources and Properties of Answers. *BMC Bioinformatics*, 10 (Suppl 9):S8.

Paice, C. D. and P. A. Jones. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78.

Passonneau, R.J., A. Nenkova, K. McKeown, and S. Sigelman. 2005. Applying the Pyramid Method in DUC 2005. In *Proceedings of the 5th Document Understanding Conference*.

Patwardhan, S. 2003. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota.

Pedersen, T., S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 1024–1025.

Plaza, L. and A. Díaz. 2010. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. In *Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems*, pages 296–303.

Plaza, L., E. Lloret, and A. Aker. 2010. Improving Automatic Image Captioning Using Text Summarization Techniques. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue*, pages 165–172.

Ponzetto, S. P. and R. Navigli. 2010. Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531.

Pradhan, S., E. Loper, D. Dligach, and M. Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92.

Radev, D., S. BlairGoldensohn, and Z. Zhang. 2001. Experiments in Single and MultiDocument Summarization Using MEAD. In *Proceedings of the Document Understanding Conference*.

Radev, D., W. Lam, A. C. Elebi, S. Teufel, J. Blitzer, D. Liu, H. Saggion, H. Qi, and E. Drabek. 2003. Evaluation Challenges in Large-scale Document Summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 375–382.

Radev, D. R., H. Jing, and M. Budzikowska. 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In *Proceeding of the North American Chapter of the Association for Computational Linguistics, Workshop on Automatic Summarization*, pages 21–30.

Reeve, L. H., H. Han, and A. D. Brooks. 2007. The Use of Domain-specific Concepts in Biomedical Text Summarization. *Information Processing and Management*, 43:1765–1776.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Rush, J. E., A. Zamora, and R. Salvador. 1971. Automatic Abstracting and Indexing II. Production of Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. *Journal of the American Society for Information Science*, 22(4):260–274.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Revue Traitement Automatique des Langues*, 49(2):103–125.

Salton, G. and M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Sanderson, M. 1998. Accurate User Directed Summarization from Existing Tools. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, pages 45–51.

Savova, G. K., A. Coden, I. L. Sominsky, R. Johnson, P. V. Ogren, P. C. de Groen, and C. G. Chute. 2008. Word Sense Disambiguation Across Two Domains: Biomedical Literature and Clinical Notes. *Journal of Biomedical Informatics*, 41(6):1088–1100.

Schwartz, A. and M. Hearst. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 451–462.

Shi, Z., G. Melli, Y. Wang, Y. Liu, B. Gu, M. M. Kashani, A. Sarkar, and F. Popowich. 2007. Question Answering Summarization of Multiple Biomedical Documents. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 284–295.

Sinha, R. and R. Mihalcea. 2007. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In *Proceedings of the IEEE International Conference on Semantic Computing*.

Sparck-Jones, K. 1999. *Automatic Summarising: Factors and Directions*. The MIT Press.

Sparck-Jones, K. 2007. Automatic Summarising: A Review and Discussion of the State of the Art. Technical Report 679, University of Cambridge.

Steinberger, J., M. Poesio, M. A. Kabadjov, and K. Jezek. 2007. Two Uses of Anaphora Resolution in Summarization. *Information Processing and Management*, 43(6):1663–1180.

Stevenson, M., Y. Guo, R. Gaizauskas, and D. Martinez. 2008. Disambiguation of Biomedical Text Using Diverse Sources of Information. *BMC Bioinformatics*, 9(Suppl 11):S7.

Teufel, S. and M. Moens. 1997. Sentence Extraction as a Classification Task. In *Proceedings of the Association for Computational Linguistics, Workshop on Intelligent Scallable Text Summarization*, pages 58–65.

Tsatsaronis, G., M. Vazirgiannis, and I. Androutsopoulos. 2007. Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1725–1730.

Van Dijk, T. 1988. *News as Discourse*. Erlbaum Associates.

Wu, Z. and M. Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Xie, S., B. Favre, D. Hakkani-Tur, and Y. Liu. 2009. Leveraging Sentence Weights in a Concept-based Optimization Framework for Extractive Meeting Summarization. In *Proceedings of Interspeech*, pages 1503–1506.

Yoo, I., X. Hu, and I-Y. Song. 2007. A Coherent Graph-based Semantic Clustering and Summarization Approach for Biomedical Literature and a New Summarization Evaluation Method. *BMC Bioinformatics*, 8(9).

Zhao, L., L. Wu, and X. Huang. 2009. Using Query Expansion in Graph-based Approach for Query-focused Multi-document Summarization. *Information Processing and Management*, 45:35–41.

Zhou, L., M. Ticrea, and E. Hovy. 2004. Multi-document Biography Summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 434–441.

Apendix A

Publications

A.1. Biomedical Summarization and Other Biomedical NLP Task

1. Plaza L, Díaz A, Gervás P. 2008. Concept-graph based Biomedical Automatic Summarization using Ontologies. In *Proceedings of the workshop "TextGraphs-3: Graph-based Algorithms for Natural Language Processing", held in conjunction with the International Conference on Computational Linguistics (COLING 2008)*, pages 53-56. Manchester, United Kingdom.
2. Plaza L, Díaz A, Gervás P. 2008. Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina. *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*, no. 41, pages 191-198.
3. Plaza L, Carrillo de Albornoz J, Prados J. 2010. Sistemas de Acceso Inteligente a la Información Biomédica: una Revisión. *Revista Internacional de Ciencias Podológicas*, vol. 4, no. 1, pages 7-15.
4. Plaza L, Díaz A. 2010. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems*. Lecture Notes in Computer Sciences, 6177, pages 296-303. Cardiff, United Kingdom.
5. Plaza L, Stevenson M, Díaz A. 2010. Improving Summarization of Biomedical Documents using Word Sense Disambiguation. In *Proceedings of the workshop "BioNLP 2010" held in conjunction with the 48th An-*

nual Meeting of the Association for Computational Linguistics, pages 55-63, Uppsala, Sweden.

6. Plaza L, Díaz A, Gervás P. A Semantic Graph-based Approach to Biomedical Summarization. *Artificial Intelligence in Medicine*. Elsevier. In press.
7. Plaza L, Stevenson M, Díaz A. Resolving Ambiguity in Biomedical Text to Improve Summarization. *Information Processing and Management*. Under review.

A.2. News Summarization

1. Plaza L, Díaz A, Gervás P. 2009. Automatic Summarization of News using WordNet concept graphs. In *Proceedings of the IADIS International Conference Informatics*. Algarve, Portugal. Best Paper Award.
2. Plaza L, Díaz A, Gervás P. 2010. Automatic Summarization of News using WordNet concept graphs. *IADIS International Journal on Computer Science and Information Systems*, vol. V, pages 45-57.

A.3. Multi-document Summarization

1. Plaza L, Lloret E, Aker A. 2010. Improving Automatic Image Captioning Using Text Summarization Techniques. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD 2010)*. Lecture Notes in Artificial Intelligence, 6231, pages 165-172. Brno, Czech Republic.
2. Aker A, Plaza L, Lloret E, Gaizauskas R. 2010. Towards Automatic Image Description Generation using Multi-document Summarization Techniques. *Multi-source, Multi-lingual Information Extraction and Summarization (MMIES)*. Book Chapter. Springer Book. In press.

A.4. Application of the Concept Identification and Disambiguation Step to Other NLP Tasks

1. Carrillo de Albornoz J, Plaza L, Gervás P. 2010. Improving Emotional Intensity Classification using Word Sense Disambiguation. *Journal on Research in Computing Science, 46. Special Issue: Natural Language*

Processing and its Applications, pages 131-142.

2. Carrillo de Albornoz J, Plaza L, Gervás P. 2010. A Hybrid Approach to Emotional Sentence Polarity and Intensity Classification. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010)*, pages 153-161. Uppsala, Sweden.

A.5. Application of the Concept Identification and Clustering Steps to Other NLP Tasks

1. Carrillo de Albornoz J, Plaza L, Gervás P, Díaz A. 2011. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. In *Proceedings of the 33rd European Conference on Information Retrieval*. In press.