Procesamiento del lenguaje Natural, nº 41 (2008), pp. 191-198 recibido 7-05-2008; aceptado 16-06-2008

Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina*

Concept-graphs based Biomedical Automatic Summarization using UMLS

Laura Plaza Morales
Alberto Díaz Esteban
Pablo Gervás
Universidad Complutense de Madrid
C/Profesor José García Santesmases, s/n, Madrid 28040, Spain
lplazam@pas.ucm.es,albertodiaz@fdi.ucm.es, pgervas@sip.ucm.es

Resumen: Uno de los principales problemas en la investigación sobre generación automática de resúmenes (GAR) es la falta de utilización de conocimiento de dominio, que se refleja en la incorrecta interpretación semántica del documento y la baja calidad de los resúmenes obtenidos. En este trabajo se propone un método de extracción de oraciones para la GAR de artículos biomédicos, mediante el mapeo del documento a los conceptos de la ontología UMLS, y la representación del documento y de sus oraciones como grafos. La selección de las oraciones relevantes se realiza a partir de la conectividad de los conceptos que contienen en el grafo del documento. Se muestran los resultados empíricos preliminares de la aplicación de distintas heurísticas para la selección de las oraciones del resumen, y se identifican algunos problemas y líneas de trabajo futuras.
Palabras clave: Generación automática de resúmenes, Unified Medical Language System(UMLS), redes libres de escala, artículo biomédico, ontología.

Abstract: One of the main problems in research on automatic summarization is the inaccurate semantic interpretation of the source, which is reflected in the deficiencies shown by the resulting summaries. Using specific domain knowledge, as that supplied by ontologies, can considerably alleviate the problem. In this paper, we introduce an ontology-based extractive method for summarization. It is based on mapping the text to concepts in the ontology and representing the document as a scale-free graph. To assess the importance of the sentences we compute the centrality of their concepts in the text. We have applied our approach to summarize scientific biomedical literature, taking advantage from free resources as UMLS. Preliminary empirical results are presented and pending problems are identified.
Keywords: automatic summarization, degree-based methods,Unified Medical Language System(UMLS), scale-free network, biomedical article, ontology.

1. Introducción

En la sociedad actual, la cantidad de documentación electrónica accesible desde cualquier lugar y cualquier dispositivo crece de manera exponencial. Gracias a los avances tecnológicos de las últimas décadas, su almacenamiento y acceso ya no suponen un problema, pero el tiempo sigue siendo un bien valioso y limitado. Esta realidad afecta especialmente al campo de la medicina, en el que los recursos digitales son muchos y muy variados. Es obvio que en este contexto, la generación automática de resúmenes (en adelante, GAR), ya sean informativos o meramente indicativos, puede ser de gran utilidad. Durante el ejercicio de la medicina, disponer de resúmenes de los historiales de los pacientes puede ayudar a los profesionales a actuar con mayor celeridad en el tratamiento de urgencias sanitarias; mientras que, durante la formación y la investigación, los resúmenes pueden ser útiles para determinar si un documento resulta de interés, y si merece o no

* Esta investigación está financiada por el Ministerio de Educación y Ciencia (TIN2006-14433-C02-01) y la Universidad Complutense de Madrid y la Dirección General de Universidades e Investigación de la Comunidad Autónoma de Madrid (CCG07-UCM/TIC-2803).

 © 2008 Sociedad Española para el procesamiento del Lenguaje Natural

una lectura exhaustiva.

Durante los últimos años, y como respuesta a esta explosión de información, han aparecido nuevos recursos lingüísticos para su tratamiento. Diccionarios, tesauros, bases de datos léxicas y grandes bases de conocimiento biomédico, muchos de ellos de disponibilidad pública, facilitan la construcción de sistemas de procesamiento de lenguaje natural y les confieren mayores garantías de éxito.

Por otra parte, construir resúmenes genéricos y totalmente independientes del contexto es un ideal aún lejos de alcanzar. Restringir el problema a un dominio concreto, la biomedicina, y un tipo de documentos específico, el artículo científico, sin duda reduce la complejidad del proceso y redunda en una mayor calidad de los resúmenes.

En este trabajo se propone un método de extracción de oraciones para la GAR de artículos biomédicos, mediante el mapeo del documento a los conceptos de la ontología biomédica UMLS, y la representación del documento y de sus oraciones como grafos. La selección de las oraciones relevantes se realiza a partir de la conectividad de los conceptos que contienen en el grafo del documento.

El resto del documento se organiza como sigue. En el apartado 2 se ofrece una panorámica general de la problemática de la GAR y del estado del arte. En el apartado 3 se describen algunas de las bases de conocimiento biomédico más populares y se justifica la elección de UMLS para el trabajo que nos ocupa. El apartado 4 presenta el método de GAR desarrollado. En el apartado 5 se muestran los resultados y la evaluación del sistema. Finalmente, en el apartado 6 se analizan las conclusiones extraídas y se subrayan algunas posibles líneas de trabajo futuro.

2. Trabajo Previo

Según Sparck-Jones (Sparck-Jones, 1999), un resumen consiste en la transformación de un texto a través de la reducción de su contenido, bien por selección o por generalización de lo que es importante. La información presentada en el resumen dependerá de las necesidades del usuario. Mientras que los *resúmenes adaptativos* seleccionan los contenidos que son de interés para el lector, los *resúmenes genéricos* tratan de preservar el punto de vista del autor y la organización original del texto. Por otra parte, en función del número de documentos que intervienen en la elaboración del resumen, cabe hablar de resúmenes *mono-documento* y resúmenes *multi-documento*. A pesar de que los trabajos más recientes centran su atención en estos últimos, lo cierto es que los resultados en generación mono-documento aún presentan notables deficiencias en cuanto a contenido y coherencia gramatical se refiere.

Una clasificación de alto nivel de los sistemas de GAR es la que distingue entre aquellos que utilizan *técnicas de extracción*; es decir, generan resúmenes compuestos íntegramente por material del documento original, y aquellos que utilizan *técnicas de abstracción*; es decir, generan resúmenes que incluyen contenidos que no están presentes, al menos explícitamente, en la entrada. Aunque típicamente los humanos realizan resúmenes mediante abstracción, la mayor parte de la investigación hoy día sigue siendo en extracción.

Los sistemas basados en extracción de oraciones realizan un análisis superficial del texto, y no van más allá del nivel sintáctico. Los primeros trabajos se limitaban a localizar segmentos clave en el original, utilizando *criterios estadísticos*, como la frecuencia de las palabras en el documento(Luhn, 1958; Edmundson, 1969); *criterios posicionales*, teniendo en cuenta la posición que ocupa cada oración (Brandow, Mitze, y Rau, 1995); y *criterios lingüísticos*, que evalúan la presencia de ciertas expresiones o palabras indicativas (Edmundson, 1969). Muchos trabajos (Edmundson, 1969) combinan algunos o todos los criterios anteriores, mientras que los enfoques más sofisticados utilizan técnicas de aprendizaje automático para determinar el conjunto de atributos que mejor se comportan en la extracción (Kupiec, Pedersen, y Chen, 1995; Lin, 1999).

En los últimos años, han cobrado relevancia los enfoques que, al igual que este trabajo, utilizan algoritmos basados en grafos para representar la estructura de los documentos y elaborar el resumen (Yoo, Hu, y Song, 2007). Un trabajo muy representativo de este tipo de aproximaciones se presenta en (Erkan y Radev, 2004), donde se aborda el problema de la GAR multi-documento. Los autores proponen la construcción de un grafo para el conjunto de los textos, en el que existe un vértice por cada oración, representada por sus vectores de frecuencias (tf*idf), y calculan la similitud entre ellas utilizando la

métrica del coseno. No obstante, el enfoque anterior presenta algunos problemas importantes, derivados de la no consideración de la estructura semántica del documento y de las relaciones entre los términos que lo componen (sinonimia, hiperonimia, homonimia, coocurrencias o asociaciones semánticas). Para ilustrar alguno de estos problemas, consideremos las siguientes oraciones extraídas de (Yoo, Hu, y Song, 2007).

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

Puesto que ambas secuencias contienen términos diferentes, con una aproximación basada en las frecuencias de los términos resulta imposible determinar que las dos oraciones presentan una semántica común.

El método que se propone trata de solventar este problema. Para ello, se ha adoptado un enfoque basado en la representación del documento en forma de grafo, utilizando los conceptos de UMLS asociados a sus términos, extendidos con sus correspondientes hiperónimos y relaciones asociativas. A diferencia de los trabajos de (Yoo, Hu, y Song, 2007; Erkan y Radev, 2004), que se centran en la construcción de clusters de oraciones para determinar los temas comunes en múltiples documentos, y en la identificación de las oraciones centrales de cada cluster, en este trabajo el algoritmo de clustering es aplicado a la identificación de grupos de conceptos estrechamente relacionados, que delimitan los distintos subtemas que se tratan dentro de un texto, y cuya presencia en las oraciones determinará su grado de relevancia. El enfoque presenta la ventaja adicional de ser fácilmente extensible a la GAR multidocumento.

3. Ontologías y Recursos Lingüísticos para Biomedicina: UMLS

Las ontologías biomédicas proveen un marco organizativo de los conceptos involucrados en entidades y procesos biológicos, en un sistema de relaciones jerárquicas y asociativas que permite razonar sobre el conocimiento del dominio. Sin duda alguna, las más utilizadas en recuperación de información, son SNOMED[1], UMLS[2] y MeSH[3].

En este trabajo se ha utilizado UMLS (*Unified Medical Language System*), un sistema desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos, compuesto por tres aplicaciones: el *meta-tesauro*, que es una base de datos multilingüe con información sobre conceptos biomédicos y sus relaciones; la *red semántica*, que proporciona una clasificación de los conceptos representados en el meta-tesauro; y el *lexicón especializado*, que incluye términos biomédicos junto con información sintáctica, morfológica y ortográfica sobre los mismos.

UMLS presenta algunas ventajas frente a las otras ontologías mencionadas. En primer lugar, proporciona una mayor riqueza semántica, al recopilar el vocabulario y la organización de distintas ontologías, incluyendo MESH y SNOMED. En segundo lugar, permite restringir las fuentes de conocimiento que se desea consultar, posibilitando la comparación entre distintas terminologías. En tercer lugar, contempla vocabularios en distintos idiomas, lo que resulta muy interesante de cara a futuros trabajos en acceso a información multilingüe. Finalmente, se encuadra en un proyecto muy activo, y que cuenta con el respaldo de un considerable número de aplicaciones que lo utilizan (*PubMed*, *Indexing Initiative* de NLM o *Enterprise Vocabulary Services* del NCI).

4. Generación Automática de Resúmenes

En este apartado se presenta el método propuesto para resolver la tarea, a través de las distintas etapas que conducen a la elaboración del resumen. Los documentos utilizados en la experimentación proceden del corpus desarrollado por la editorial *BioMed Central*[4], especialmente concebido para la investigación en minería de texto. Está compuesto por más de 23900 artículos completos publicados, incluyendo una versión etiquetada y

[1] SNOMED International. URL: http://www.snomed.org/snomedct

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

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

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

estructurada de los mismos en XML, que permite identificar los distintos elementos que conforman el artículo (título, abstract, autores, secciones, palabras clave, etc.)

Como paso previo, el artículo es sometido a una etapa de preprocesamiento en la que el texto se divide en tokens, se realiza su etiquetado morfosintáctico y se divide el texto en en oraciones. Para ello, se han utilizando los módulos *Tokenizer*, *Part of Speech Tagger* y *Sentence Splitter* de la librería GATE[5]. Finalmente, se eliminan las palabras genéricas utilizando una lista de parada[6], así como los términos que presentan una alta frecuencia en el documento, puesto que no van a ser de utilidad a la hora de discriminar entre contenidos importantes e irrelevantes.

Con el objetivo de clarificar el funcionamiento del algoritmo, a lo largo de la exposición se hará referencia a un ejemplo concreto de GAR desarrollado para uno de los documentos del corpus. A continuación se muestra un extracto del documento utilizado. El texto completo presenta un total de 60 oraciones, y puede encontrarse en el sitio web de BioMed Central[7].

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 [¡abbr bid="B1»1¡/abbr¿]. 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)...

4.1. Construcción del Grafo del Documento

El objetivo de esta etapa es construir una representación del documento en forma de grafo, en la que los vértices representan los conceptos en UMLS asociados a cada término, y las aristas representan las relaciones *isa* existentes entre los conceptos.

Para ello, los términos de las oraciones del documento se traducen a conceptos de la ontología, utilizando *Metamap*[8](Aronson, 2001). MetaMap es una herramienta desarrollada por la National Library of Medicine (NLM), inicialmente pensada para su uso en indexación y recuperación de artículos en MEDLINE, y hoy en día utilizada en todo tipo de tareas de PLN en biomedicina. El uso de MetaMap en este proyecto presenta dos atractivos fundamentales. En primer lugar, MetaMap utiliza el lexicón especializado de UMLS para normalizar los términos del documento, considerando todas sus posibles variantes morfológicas antes de obtener el concepto asociado en el meta-tesauro. En segundo lugar, y puesto que ante un mismo término el meta-tesauro de UMLS puede recuperar distintos conceptos, realiza la desambiguación necesaria para determinar cuál es el concepto correcto de acuerdo con el contexto de la oración. Además, la indexación se realiza seleccionando n-gramas en lugar de términos individuales. De este modo, no sólo se consigue una mayor precisión en la interpretación semántica, sino que se reduce considerablemente el tamaño del grafo.

A continuación, los conceptos extraídos se expanden con sus hiperónimos a través de las relaciones *isa* de la ontología, y se construye la jerarquía que representa a la oración. La Figura 1 muestra el árbol correspondiente 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."*

Seguidamente, a cada arista que une un concepto con su padre se le asigna un peso que es proporcional a la profundidad del concepto en la jerarquía; es decir, será tanto mayor cuanto mas específicos sean los conceptos que conecte. El cálculo de los pesos se realiza utilizando una medida de la similitud entre conjuntos (Rada et al., 1989), de acuerdo con la expresión (1), donde α representa el conjunto de todos los ancestros de un concepto determinado, incluido el propio concepto, y β representa el conjunto de todos los ancestros del concepto del nivel inmediatamente superior(Figura 2).

$$\frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} = \frac{|\beta|}{|\alpha|} \quad (1)$$

Finalmente, los grafos de las distintas oraciones se fusionan en un único grafo, que se completa con las relaciones *associated-with* entre los grupos semánticos de UMLS a los que pertenecen los conceptos. El peso de estos enlaces se calculará siguiendo el criterio adoptado para las relaciones *isa*.

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

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

[7]BioMed Central: www.biomedcentral.com/content/download/xml/cvm-2-6-254.xml

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

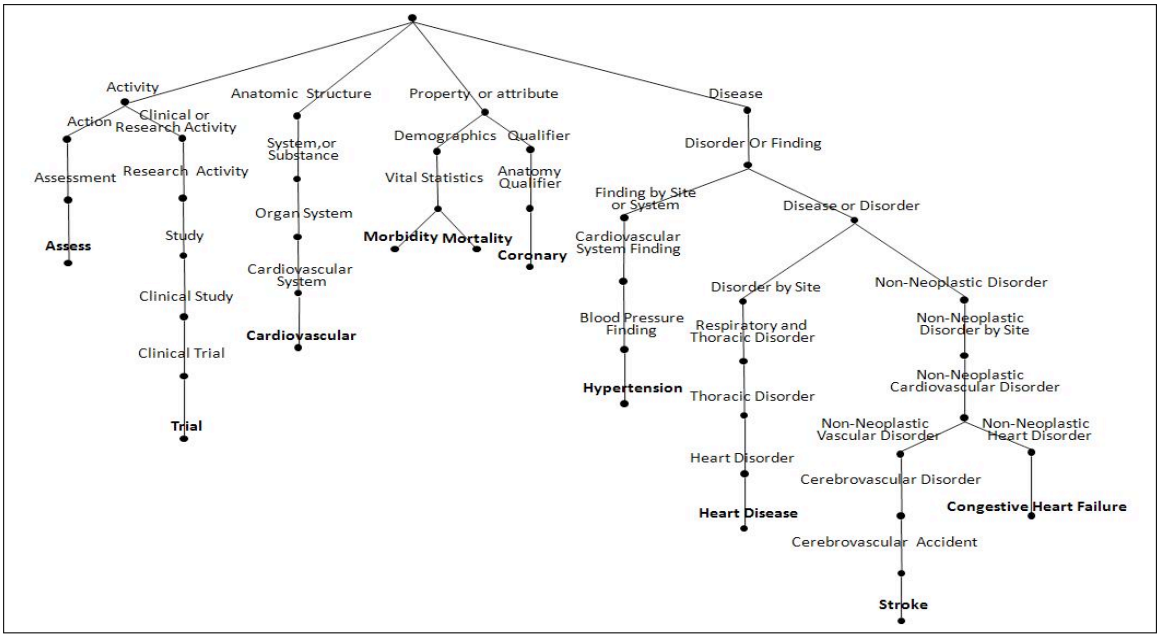

Figura 1: Grafo de una oración

Por ejemplo, los conceptos *trial* y *hypertension* están asociados, ya que sus respectivos tipos semánticos(*Research Activity* y *Disease or Syndrome*) presentan una relación *associated_with* en UMLS (Figura 2).

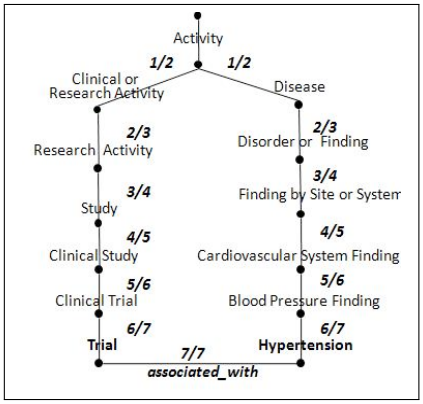

Figura 2: Asignación de pesos y relaciones *associated_with*

4.2. Clustering de Conceptos. Identificación de Subtemas

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

Se parte de la hipótesis de que el grafo obtenido constituye un ejemplo de *red libre de escala* (Barabasi y Albert, 1999). Una red libre de escala es un tipo específico de red compleja en la que algunos nodos están altamente conectados (nodos *hub*); es decir, poseen un gran número de enlaces a otros nodos, aunque el grado de conexión de casi todos los nodos es bastante bajo.

Siguiendo a (Yoo, Hu, y Song, 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 expresión (2).

$$salience(v_i) = \sum_{e_j | \exists v_k \wedge e_j conecta(v_i, v_k)} weight(e_j) \quad (2)$$

A continuación, se seleccionan los n vértices de mayor *salience*, y se agrupan formando *Hub Vertex Sets* (HVS), que constituirán los centroides de los clusters a construir. El resto de vértices se asignan al cluster para el que presenten la mayor conectividad con alguno de sus vértices, reajustando los HVS y los vértices asignados en un proceso iterativo. Para el ejemplo que nos ocupa, se han generado 4 clusters.

Finalmente, se asigna cada oración del documento a uno de los clusters anteriores. Para

ello, es preciso definir una medida de la similitud entre el cluster y el grafo de la oración. Es importante aclarar que, puesto que ambas representaciones son muy distintas en cuanto a tamaño se refiere, las métricas clásicas de similitud entre grafos (i.e. la distancia de edición) no resultan adecuadas. En su lugar, se utiliza un mecanismo de votos (Yoo, Hu, y Song, 2007). Cada vértice (v_k) de una oración (O_j) asigna a cada cluster (C_i) en el que se encuentra presente una puntuación ($p_{i,j}$) distinta dependiendo de si pertenece o no al HVS de dicho cluster (3).

$$similitud(C_i, O_j) = \sum_{v_k | v_k \in O_j} w_{k,j} \quad (3)$$

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 se han establecido a 1,0 y 0,5 respectivamente, lo que significa que se atribuye el doble de importancia a los conceptos que pertenecen a los HVS que a los restantes.

4.3. Selección de Oraciones Relevantes

El último paso del algoritmo consiste en extraer oraciones completas del texto original en función de su distancia semántica a los distintos clusters. El número total de oraciones a seleccionar (N) dependerá de la tasa de compresión utilizada. En esta etapa, se han investigado tres heurísticas.

- **Heurística 1:** Todos los clusters contribuyen a la construcción del resumen con un número de oraciones (n_i) proporcional a su tamaño. Por lo tanto, para cada uno de los clusters, se seleccionan las n_i oraciones con las que presenta mayor similitud.
- **Heurística 2:** El cluster de mayor tamaño (esto es, el que representa el *theme* principal en el documento), es el único que debería tenerse en consideración para la generación del resumen. Por lo tanto, se seleccionan las N oraciones con las que presenta mayor similitud.
- **Heurística 3:** Para cada oración, se calcula el total de sus votaciones a todos los clusters, ponderadas por el tamaño de estos últimos, según la expresión (4). Se seleccionan las N oraciones con mayor puntuación total.

$$score(O_j) = \sum_{C_i} \frac{similitud(C_i, O_j)}{|C_i|} \quad (4)$$

El problema de la ordenación de las oraciones es trivial al tratarse de un resumen mono-documento, y se resuelve tomándolas en el mismo orden en el que aparecen en el documento original.

5. Resultados y Evaluación

En este apartado se analiza un ejemplo del extracto generado con las distintas heurísticas, para el documento presentado al inicio del apartado 4, utilizando una tasa de compresión del 20 %.

La tabla 1 recoge las oraciones seleccionadas por cada heurística, junto con su puntuación. Si bien los resultados obtenidos no son estadísticamente significativos, su análisis muestra algunos aspectos en los que el algoritmo se comporta satisfactoriamente. Llama la atención que las heurísticas 1 y 3 presentan a la oración 0 como la más relevante, con una puntuación muy superior al resto de oraciones. Esto concuerda con el criterio posicional adoptado en muchos trabajos de seleccionar la primera oración del documento para el resumen, por ser la que generalmente contiene información más significativa. El motivo por el que la heurística 2 no selecciona esta oración es que el cluster de mayor tamaño (es decir, aquel del que dicha heurística extrae todas las oraciones) es el número 2, mientras que la oración 0 pertenece al cluster 0. La número 58 presenta un claro ejemplo de oración que, estando posicionada al final del documento, recoge conclusiones sobre la exposición, y por lo tanto, tiene un alto contenido informativo. Por su parte, la número 19 ejemplifica la sobrevaloración de oraciones de gran longitud. En esta oración, el mapeo al metatesauro de UMLS da como resultado un total de 23 conceptos, cuando el resto de oraciones presentan en torno a 10-12 conceptos. Por ello, y a pesar de que la mayoría de los conceptos que contienen no son centrales (es decir, no pertenecen a los HVS), recibe una puntuación elevada(20.0). Otro aspecto a destacar es que las heurísticas 1 y 3 comparten un gran número de oraciones;

en concreto, 9 de las 12 seleccionadas, mientras que la heurística 2, por su parte, se aleja bastante de las otras dos. De hecho, un análisis detallado del resultado de la segunda heurística demuestra que esta estrategia ignora algunos tópicos importantes del texto original. Por último, la oración 28 pone de manifiesto los problemas de inconsistencia típicos de los métodos de extracción, al tratarse de una oración que no es autocontenida, y que no tiene sentido incluir en el resumen si no se incluye también la oración que la precede.

Dado que los artículos del corpus se presentan acompañados del resumen elaborado por su autor, resulta interesante realizar una comparación entre éste y los resultados de las distintas heurísticas. A pesar de que las longitudes de los resúmenes varían significativamente (de las 3 oraciones del *abstract* a las 12 oraciones del resumen automático), se observa que las heurísticas 1 y 3 cubren la totalidad de los temas tratados en el *abstract*. En primer lugar, las oraciones 0 y 4 presentan el punto de partida del estudio, mientras que las oraciones 15, 17, 19, 20 y 25 presentan resultados de distintas investigaciones en ALLHAT, y de tratamientos con doxazosin. Ambos grupos de oraciones abarcan el contenido de la primera oración del *abstract*. Por su parte, las oraciones 43 y 58 advierten de la poca efectividad de las terapias con doxazosin contra las enfermedades cardiovasculares (oraciones 2 y 3 del *abstract*). La heurística 2, por su parte, no cubre satisfactoriamente el contenido del *abstract*.

6. Conclusiones y Trabajo Futuro

En este artículo se ha presentado un método para la GAR de textos biomédicos, basado en la representación del documento como un grafo extendido de conceptos y relaciones de UMLS, y en el cálculo de la relevancia de las oraciones a extraer en relación al prestigio o *salience* de los conceptos que las componen en este grafo. De este modo, se construye una representación más rica en conocimiento que la que se tendría utilizando un modelo del espacio vectorial, y se consiguen solventar los problemas identificados en el apartado 2.

En el apartado 5, se han evaluado distintas heurísticas para la extracción de las oraciones del resumen. Como resultado, se ha comprobado que la heurística número 2 no cubre todos los contenidos importantes, a la vez que selecciona oraciones de poca relevancia relativa. Por lo tanto, se concluye que no es válida para la resolución del problema. En cuanto a las heurísticas 1 y 3, y a falta de evaluarlas formalmente, se observa que presentan resultados muy similares y que cubren todos los tópicos importantes.

Por otra parte, el trabajo realizado hasta el momento ha puesto de relieve la gran complejidad de la tarea, y subrayado algunas deficiencias y posibles mejoras. En primer lugar, el método extrae oraciones completas, lo que implica que las de mayor longitud, al contener un mayor número de conceptos, tienen mayor posibilidad de ser seleccionadas. Una solución a considerar sería dividir la puntuación de las oraciones entre el número de conceptos que las componen.

En segundo lugar, los conceptos con un significado muy general no aportan información a la hora de identificar los tópicos y de discriminar entre oraciones relevantes e irrelevantes. Por lo tanto, pueden ser eliminados de los grafos, consiguiendo representaciones más compactas, y en consecuencia, mejorar el rendimiento de la aplicación. Los tipos semánticos de UMLS se pueden utilizar para identificar los términos asociados a conceptos muy generales. Por ejemplo, se podrían ignorar los términos correspondientes a los tipos *functional concept*, *temporal concept*, *entity*, *idea or concept* y *language*.

Otro problema a resolver sería el de la extracción de oraciones cuyo contenido esté relacionado con el de otras oraciones que no hayan sido seleccionadas. Para solucionar este tipo de inconsistencias, podría implementarse una estrategia basada en detectar palabras o grupos de palabras que actúan como conectores de oraciones, y no seleccionar las oraciones que las contengan a menos que también se seleccione la oración inmediatamente anterior.

Asimismo, se está estudiando una posible modificación del algoritmo para generar resúmenes adaptados al usuario. Para ello, se necesitaría disponer de un modelo del usuario, entendido como una representación de sus intereses y preferencias (Díaz y Gervás, 2004).

Otra línea de trabajo futura será la extensión del método poder realizar resúmenes a partir de múltiples documentos sobre un mismo tema. Como ya se mencionó en el apartado 2, las modificaciones no supondrán cam-

Heurística 1												
Oraciones	0	4	19	58	7	28	25	20	21	8	43	15
Puntuación	99.0	20.0	19.0	18.5	17.0	16.5	16.0	15.5	15.5	13.5	13.5	12.0
Heurística 2												
Oraciones	19	28	20	12	35	5	41	51	38	52	57	59
Puntuación	19.0	16.5	15.5	12.5	12.0	10.5	9.0	9.0	7.5	7.0	7.0	7.0
Heurística 3												
Oraciones	0	4	19	58	7	28	21	33	25	17	20	43
Puntuación	98.8	18.7	17.9	16.3	15.3	14.5	13.4	13.0	13.0	12.7	12.7	12.2

Tabla 1: Resultados

bios sustanciales, aunque deberán resolverse algunos problemas adicionales derivados de la consideración de varias fuentes (evitar la redundancia, ordenar las oraciones, etc.)

Finalmente, para garantizar la adecuación del método, se está realizando una evaluación a gran escala sobre los documentos del corpus de BioMed, basada en el cálculo de las medidas ROUGE-1, ROUGE-L y ROUGE-W (Lin, 2004), utilizadas en las conferencias DUC (Document Understanding Conferences)[9]. Esta evaluación servirá para el ajuste de parámetros, así como para evaluar objetivamente la bondad de cada una de las heurísticas definidas.

Bibliografía

Aronson, A. R. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. En *Proceedings of American Medical Informatics Association.*

Barabasi, A.L. y R. Albert. 1999. Emergence of scaling in random networks. *Science*, páginas 286–509.

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.

Díaz, A. y P. Gervás. 2004. User-Model Based Personalized Summarization. *Information Processing and Management*, 43(6):1715–1734.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 2(16):264–285.

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

Erkan, G. y D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–479.

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.

Lin, C-Y. 1999. Training a Selection Function for Extraction. En *Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM)*, páginas 55–62, Kansas City.

Lin, C-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. En *In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain.*

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Rada, R., H. Mili, E. Bicknell, y M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, páginas 17–30.

Sparck-Jones, K. 1999. Automatic Summarizing: Factors and Directions. En *I. Mani y M.T. Maybury, Advances in Automatic Text Summarization.* The MIT Press.

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).