Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 131-139 recibido 1-05-2009; aceptado 5-06-2009

Propuesta y evaluación de un método extractivo de generación de resúmenes en el ámbito biomédico basado en conceptos

A proposal and evaluation of an extractive method for summarization in the biomedical domain based on concepts

Manuel de la Villa y Manuel J. Maña
Departamento Tecnologías de la Información
Universidad de Huelva.
Campus La Rabida. Edif. Torreumbría,
21618, Palos de la Frontera, Huelva, España
{manuel.villa, manuel.mana}@dti.uhu.es

Resumen: Los métodos de generación de resúmenes basados en técnicas extractivas han demostrado ser muy útiles por su adaptabilidad y eficiencia en tiempo de respuesta en cualquier tipo de dominios. En el ámbito biomédico son numerosos los estudios que hablan de la sobrecarga de información y recogen la necesidad de aplicación de técnicas eficientes de recuperación y generación de resúmenes para una correcta aplicación de la medicina basada en la evidencia. En este contexto vamos a presentar una propuesta metodológica de generación automática de resúmenes basada en ontologías y grafos, aplicando técnicas de similitud y la frecuencia de aparición de los conceptos para obtener las frases más relevantes. Se realiza una evaluación de la propuesta frente a otras metodologías con la herramienta ROUGE y se analizan los resultados. Aunque la extensión del conjunto de evaluación no permite extraer conclusiones significativas, los resultados son suficientemente prometedores como para confiar en la efectividad de la propuesta presentada.
Palabras clave: resumen automático, método extractivo, conceptos biomédicos, UMLS
Abstract: The methods for automatic summarization generation based in extractive techniques have widely shown its utility for his adaptability and efficiency in the manner of response time at any kind of application domain. In Biomedical field are numerous the research results about the overload information and the need of application of efficient recovery and summarization methods for the proper use of evidence based medicine. In this context we are going to present a proposal of methodology for automatic summarization based on structured knowledge and graph's use, applying similarity methods between phrases and considering concepts appearance frequency. Finally, a methodology's evaluation is made to compare with other methods using the ROUGE tool and analyzing their results.
Although the size of the evaluation set doesn't allow extracting noteworthy conclusions, the results collected are enough promising to trust in the effectiveness of the proposal handed in.
Keywords: automatic summarization, extractive method, biomedical concepts, UMLS

1 Introducción

La generación de resúmenes de texto es un proceso de reducción de la información, que permite a un usuario tomar idea o conocer el contenido de un texto completo, sin tener que leer todas sus frases. Esta reducción de la cantidad de información a leer produce una mayor rapidez en la búsqueda de información relevante y una mayor asimilación de conceptos con menor esfuerzo.

Numerosos artículos certifican la sobrecarga de información tan común hoy día en nuestra sociedad, y en especial en el ámbito biomédico, donde la información está disponible desde una variedad de fuentes, incluyendo artículos científicos, bases de datos de resúmenes, bases de datos estructuradas o semiestructuradas, servicios web, webs de documentos o historia clínica de pacientes (Afantenos et al., 2005).

ISSN: 1135-5948

Si a ello unimos el hecho de que gran parte de los resultados de la investigación biomédica se encuentran en forma de literatura escrita en formato libre que se acumulan en grandes bases de datos en línea, podemos concluir que el proceso de reducción de información que producen los resúmenes automáticos es especialmente útil en el ámbito biomédico.

Por otro lado, el rápido crecimiento de los resultados de la investigación del dominio biomédico está produciendo un importante cuello de botella. MEDLINE[1] (*Medical Literature Analysis and Retrieval System Online*), la principal base de datos bibliográfica de EE.UU (de la *National Library of Medicine*), contiene más de 16 millones de referencias a artículos de revistas, centrados principalmente en biomedicina. Entre 2000 y 4000 referencias completas se añaden cada día, más de 670000 fueron añadidas en 2007.

Es evidente que en este dominio, los profesionales en general necesitan herramientas orientadas a proporcionar medios para acceder y visualizar la información adecuada para sus necesidades.

En este trabajo vamos a presentar el modelo de generación de resúmenes de carácter extractivo apoyado en conceptos del dominio biomédico así como una evaluación realizada con un mini-corpus, con el que podemos obtener unas primeras conclusiones. Para ello estructuramos el documento de la siguiente manera: en primer lugar se comentan trabajos de interés que son específicos del dominio. También presentamos UMLS y el conjunto de herramientas de procesamiento de lenguaje natural orientadas al ámbito biomédico que incorpora. En la sección tres introducimos el modelo de generación de resúmenes en que estamos trabajando, dividido en cuatro fases. En la sección 4 se recoge el procedimiento seguido para evaluar la efectividad de nuestro sistema, enfrentándolo a sistemas presentes en el mercado, ya sean resultados de investigación o herramientas comerciales. Para ello se explica el uso de la herramienta automática de evaluación ROUGE, se explica el corpus de documentos a evaluar y una breve reseña de cada sistema contra el que nos comparamos. En la sección 5, se presentan los resultados de la evaluación y se interpretan. En la sección 6 se establecen los resultados y las conclusiones del presente trabajo.

2 Trabajos relacionados en el ámbito biomédico y recursos UMLS.

Una primera propuesta de nuestro trabajo, (de la Villa y Maña, 2009) recoge en detalle el proceso de generación de resúmenes, los principales trabajos de referencia en el ámbito, así como una descripción detallada de los recursos UMLS que usa.

En el ámbito biomédico destacaremos los métodos de generación de resúmenes extractivos como *BioChain*, (basado en cadenas de conceptos o relaciones semánticas entre conceptos vecinos en texto), *FreqDist* (centrado en el uso de las distribuciones de frecuencia, construyendo un resumen con similar distribución que el original) y *Chainfreq* (híbrido de los dos anteriores), que usan conceptos específicos del dominio biomédico para identificar las sentencias destacables del texto completo (Reeve et al., 2007).

Los trabajos específicos de un ámbito usan conceptos en vez de términos, para lo que necesitan herramientas que den soporte a la identificación de los conceptos en una estructura de conocimiento del dominio y capaces de determinar relaciones semánticas entre estos conceptos.

Para el procesado semántico, consistente en el análisis e identificación de los conceptos y relaciones subyacentes en un texto, se requiere una estructura de conocimiento, como la que en el ámbito biomédico proporciona el proyecto Unified Medical Language System (UMLS) (Humphreys et al., 1998). El objetivo de este proyecto es el desarrollo de herramientas que ayuden a investigadores en la representación del conocimiento, recuperación e integración de información biomédica.

UMLS consiste en tres componentes, el SPECIALIST Lexicon, el Metathesaurus y la UMLS Semantic Network (Rindflesh et. al., 2005). SemRep es una herramienta de procesado semántico que integra los tres anteriores componentes de UMLS para analizar de manera automática textos con lenguaje

[1] www.nlm.nih.gov/pubs/factsheets/medline.html

médico identificando los conceptos y relaciones que representan el contenido del documento.

Usaremos el Metathesaurus y la herramienta Metamap Transfer (MMTx) para la identificación de los conceptos biomédicos de cada frase, base para el cálculo del solape entre frases. En cuanto a SemRep, añadiremos esta lista de relaciones al grafo dirigido para posteriores trabajos.

3 Propuesta de generación del resumen

Los métodos de generación de resúmenes basados en técnicas extractivas han demostrado ser muy útiles por su adaptabilidad y eficiencia en tiempo de respuesta en cualquier tipo de dominios. Por contra, los métodos abstractivos, por la necesidad de recursos léxicos, sintácticos y semánticos han proporcionado unos mejores resultados en cuanto a comprensibilidad a costa de un mayor esfuerzo computacional y por tanto, de tiempos de respuesta, aparte de la especificidad del ámbito de uso de la herramienta.

Nuestro objetivo es intentar combinar la capacidad y rapidez de los métodos extractivos con la efectividad y concreción de los métodos abstractivos. Para ello vamos a presentar la propuesta en que venimos trabajando de una metodología de generación automática de resúmenes apoyada en conocimiento estructurado y grafos de ranking.

Nuestra propuesta, basada en (Milhacea y Tarau, 2006) es eminentemente extractiva, de modo que el proceso podría resumirse en identificar las sentencias en el texto de origen, seleccionar aquellas que sean relevantes para el usuario a la vez que disminuimos la redundancia de la información. Para ellos asignamos una puntuación a cada frase de acuerdo a un conjunto de características. Las *n*-primeras frases en cuanto a puntuación se extraen y se presentan al usuario en su orden de aparición en el texto original.

3.1 Fase 1. Generación del grafo.

Independientemente del tamaño del texto, sea un texto completo o un *abstract*, la primera tarea debe consistir en la identificación de cada una de las sentencias del texto de origen, así como en la creación de un grafo que incluya un vértice en el grafo por cada sentencia. De manera simultánea, se identifican con la ayuda de Metamap Transfer (integrada en SemRep), los conceptos biomédicos incluidos en la frase y se incluyen en el nodo, así como las relaciones semánticas.

3.2 Fase 2. Aplicación de algoritmo de similitud.

Para la extracción de sentencias en resúmenes, un concepto importante es la 'similaridad' o grado de solapamiento entre sentencias, cuánto del contenido de una sentencia se encuentra incluido en otra. Es como si consideráramos el solape como una "recomendación" de una frase de dirigirse a otras que tratan y abundan los mismos conceptos. Una función de similaridad, que tome en cuenta el grado de repetición de *tokens* entre sentencias de manera normalizada proporcionará una medida de este concepto. Aplicamos una versión modificada (con conceptos en vez de términos) de la formula de similaridad de Milhacea y Tarau (2006):

$$Similitud(V_i, V_j) = \frac{\left|\{T_k \mid T_k \in V_i \wedge T_k \in V_j\}\right|}{\log(|V_i|) + \log(|V_j|)} \quad (1)$$

que podríamos expresar como sigue:

$$Similitud_p(V_i, V_j) = \frac{\left(\{df(C_k) \mid C_k \in V_i \wedge C_k \in V_j\}\right)}{\log(|V_i|) + \log(|V_j|)} \quad (2)$$

La recomendación que hace cada concepto sobre sus apariciones en otras frases no es equitativa sino que está ponderada. Su influencia será mayor cuanto mayor sea su frecuencia de apariciones en el documento (*df, document frequency*).

3.3 Fase 3. Aplicación de algoritmo de ranking

Los algoritmos de ranking basados en grafos, a partir de la asignación arbitraria de valores a cada nodo, realizan cálculos para obtener la puntuación S(Vi) de cada nodo de manera iterativa, hasta que se produce convergencia bajo un determinado umbral. Las referencias entre nodos y/o conceptos son tratadas como 'votos' para decidir el elemento más importante. La puntuación de cada vértice se obtiene aplicando *Pagerank* (Brin y Page, 1998):

Tras la ejecución del algoritmo, los nodos se ordenan atendiendo al peso o puntuación asociada, que define la notoriedad (*saliency*) de cada vértice en un grafo dirigido y ponderado.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in I_n(V_i)} \frac{W_{ji}}{\sum_{V_k \in O_{ut}(V_j)} W_{jk}} WS(V_j) \quad (3)$$

3.4 Fase 4. Creación del resumen

Los nodos de mayor puntuación definirán las frases a incluir en el resumen. El número de frases puede ser fijo o basado en un umbral. En nuestro prototipo es el usuario el que decide el porcentaje de frases del documento original que formarán parte del resumen. Para facilitar la legibilidad del resumen, la secuencialidad de presentación de las frases seleccionadas se hace atendiendo a su ordenamiento original.

4 Evaluación

Aunque la mayoría de trabajos de generación automática de resúmenes de texto tienen una componente teórica importante, suelen establecer hipótesis o proponer técnicas y algoritmos que necesitan ser validados, evaluados y comparados de una manera formal y rigurosa. Tradicionalmente, la evaluación de resúmenes ha requerido el juicio humano de diferentes métricas de calidad, como p.ej., coherencia, concisión, gramaticalidad, legibilidad y contenido (Mani, 2001).

No obstante, incluso la evaluación manual de resúmenes a gran escala sobre unas cuantas cuestiones de calidad lingüística y cobertura de contenido como las realizadas en la conferencia DUC (Document Understanding Conference) hubieran requerido sobre 3000 horas de esfuerzos humanos (Over y Yen, 2003). Una tarea como ésta es muy cara y difícil de llevar a cabo de manera frecuente, por lo que la evaluación automática de resúmenes es un desafío de investigación en el que la comunidad de generación de resúmenes ha puesto su objetivo en los últimos años.

Ante la escasez de propuestas y la dificultad de la tarea, ROUGE (Hovy, Lin 2003) se ha convertido en el paquete de evaluación de resúmenes automatizados más frecuentemente usado. Tras la exitosa aplicación en la evaluación de traductores automáticos (*MT Machine Translations*) de herramientas como BLEU (Papineni et al. 2001), Lin y Hovy presentaron ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). En dicho trabajo mostraban que los resúmenes producidos por jueces humanos no eran fiables como 'resúmenes ideales' (*gold standard*), a causa de las fuertes discrepancias que presentan. Como consecuencia, un resumen de consenso obtenido por la aplicación de métricas basadas en el contenido, como el solape de N-gramas, parecía mucho más fiable como resumen ideal para usarlos como referencia en el contraste de nuevos modelos.

ROUGE, desarrollado por el *Information Science Institute* en la *University of Southern California* es una herramienta automática que compara un resumen generado por un sistema automatizado con uno o más resúmenes ideales, los llamados 'modelos'. ROUGE usa N-gramas para determinar el solapamiento entre el resumen generado y los modelos. ROUGE ha sido usado desde 2004 en las Conferencias DUC (*Document Understanding Conference*) como herramienta de evaluación en las competiciones y es un estándar 'de facto' asumido por la comunidad internacional del ámbito que usaremos nosotros para evaluar la corrección de nuestro trabajo.

Según (Lin, 2004) ROUGE-2, ROUGE-L, ROUGE-W y ROUGE-S funcionan bien en tareas de resumen monodocumento. Del mismo modo, afirma que ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, y ROUGE-SU9 dan grandes resultados en la evaluación de resúmenes muy cortos.

4.1 Descripción del corpus de evaluación

Para comparar los resúmenes generados automáticamente y probar nuestro sistema, hemos usado una pequeña colección de cuatro documentos médicos, cedidos por Plaza, Díaz y Gervas (2008), que también trabajan en la generación de resúmenes usando grafos y ontologías. Estos documentos obtenidos de la colección *Biomed Central* han sido resumidos manualmente por expertos médicos.

Indicar que sobre los documentos se realizó previamente un trabajo de preprocesado, eliminando de cada documento el título, el *abstract* así como las cabeceras de sección (elementos que de ser tratados y tenidos en

cuenta supondrían un gran valor añadido, por la relevancia del título y del abstract), tablas e imágenes.

Para comparar los resúmenes generados automáticamente por los diferentes sistemas, vamos a usar 5 modelos o resúmenes ideales de cada documento de la colección, que representan supuestamente diferentes versiones de un resumen ideal. Dos resúmenes han sido elaborados por dos expertos para un ratio de compresión del 20%, otros dos resúmenes más, elaborados por el primero de los expertos con ratios del 30% y el 50%. Como quinto modelo ideal usaremos el *abstract* de cada artículo.

4.2 Generadores de resúmenes usados para evaluación.

Para la evaluación de nuestra propuesta vamos a presentar dos modelos candidatos, el primero de ellos (UHU2) es un baseline con un proceso inicial de tratamiento de los textos médicos, que no tiene en cuenta la frecuencia de aparición de conceptos. El segundo de ellos (UHU1) incorpora en la fórmula de cálculo de la similitud entre frases, la frecuencia de aparición de los conceptos que se solapan.

Para tener un mayor grado de conocimiento de la eficiencia genérica del proceso, se incluye en la comparativa los resultados obtenidos sobre el mismo corpus aplicando diferentes herramientas, plenamente admitidas y utilizadas como referencia por la comunidad internacional, tanto resultado de proyectos de investigación como herramientas comerciales.

Cada herramienta generó un resumen de un tamaño igual al 20% del tamaño del texto original. El tamaño del resumen no ha sido una decisión arbitraria. En el dominio de los resúmenes de noticia típicamente se selecciona un tamaño de cómo máximo 5 líneas, que representa sobre el 20% del tamaño de una noticia típica (Goldstein et al., 1999). Ha sido generalmente aceptado que un resumen no debería ser más corto del 15% ni más largo del 35% del tamaño del texto fuente (Hovy, 2005).

Hagamos una breve descripción de los generadores usados:

- *UHU1*. En esta primera versión se ha aplicado la metodología tal cual ha sido presentada en la sección anterior.
- *UHU2*. Es una versión anterior que no tiene en cuenta la ponderación de los conceptos, considera todos los conceptos iguales y aplica directamente la fórmula de similitud de Milhacea y Tarau (2006) cambiando términos por conceptos.
- *LEAD*. Uno de los que suele llamarse generadores de línea base, ya que su objetivo es dar alguna idea del nivel de rendimiento de una implementación muy simple. LEAD (o Lead) de manera secuencial recupera las primeras sentencias del texto, hasta completar el 20% del tamaño. Indicar que en textos periodísticos y científicos, las primeras líneas del documento suelen tener un alto grado de significatividad.
- *MS-Word (AutoSummarize)*. Esta función se encuentra incluida en el procesador de textos Microsoft Word v.2007 (concretamente, ensamblado en MS.Office.Tools.Word.v9.0.dll). Aunque los detalles concretos del algoritmo no son públicos, en la ayuda online del producto se afirma que las sentencias que usan palabras frecuentemente usadas tienen una mayor puntuación.
- *Copernic Summarizer*[2]. Es una herramienta multilingüe comercial de generación de resúmenes a partir de textos o páginas web con el objetivo de disminuir el tiempo de acceso del usuario a la información importante. Obtiene los conceptos clave y frases clave a partir de un ratio de compresión dado. Se integra fácilmente en procesadores de texto, navegadores y clientes de correo. Los algoritmos y técnicas usadas no son públicos, sólo se revela que usa 'sofisticados' algoritmos estadísticos y lingüísticos, eliminando automáticamente contenido y texto irrelevante.
- *Pertinence Summarizer*[3] es una herramienta comercial de generación de resúmenes que se basa en técnicas extractivas, mediante el procesamiento de la relevancia (pertinencia la denominan ellos) de cada sentencia, tomando en cuenta posibles palabras clave, diccionarios de términos y marcadores lingüísticos genéricos. Es multilingüe y se ha usado la versión online para la evaluación.

[2] http://www.copernic.com/en/products/summarizer/
[3] http://www.pertinence.net/index_en.html

- *Swesum*. Es un generador de resúmenes multilingüe (Hassel, 2007), inicialmente para sueco e inglés. Utiliza múltiples aspectos para valorar las sentencias, como su posición o valor numérico en un esquema, de modo que las sentencias iniciales tienen un peso adicional, así como las numeradas. Para la evaluación se ha usado la versión online[4], con las opciones por defecto.
- *Summ-It*. Es un módulo para generación de resúmenes integrado en la plataforma System Quirk[5], un banco de trabajo para el aprendizaje e investigación de técnicas de procesamiento del lenguaje natural.
- *Mead.* Es generador de resúmenes mono- y multidocumento (Radev et al., 2004), que usa múltiples criterios a la hora de puntuar las sentencias, como la posición de la sentencia en el texto, el solape de cada sentencia con la primera sentencia, la longitud de la sentencia y un método basado en el centroide de un clúster de documentos. Para la evaluación se ha usado la demo online[6] (MEAD, 2008), que los autores avisan de que es más limitada que la versión para descarga.
- *LexRank.* (Erkan y Radev, 2004) es un método multidocumento de generación de resúmenes extractivo orientado a la obtención de la relevancia de una frase en base al concepto de centralidad del vector propio (*eigen vector*) en una representación de sentencias en un grafo. Una matriz de conectividad basada en la similitud entre sentencias (método del coseno) es usada como matriz de adyacencia del grafo de sentencias. Este método quedó primero en la tarea de resúmenes del DUC 2004. Para la evaluación se ha usado la versión online[7] (LexRank, 2008), más limitada.

5 Resultados y discusión

Los resultados de la evaluación usando ROUGE se muestran en las siguientes tablas. Cada tabla se ha ordenado descendentemente atendiendo a su puntuación (medida F). Por lo tanto, el método de generación de resúmenes más eficiente aparece en la primera entrada de cada tabla mientras que el menos eficiente aparece en la última fila.

ROUGE es una herramienta que permite evaluaciones parametrizadas en función de ciertos valores, que orientan la tarea al tipo especial de documento original y resumen a generar. Vamos a diferenciar y separar entre dos evaluaciones, una con los parámetros que se han usado en la tarea de resumen del DUC y otra con los valores por defecto de ROUGE, en la que se realizarán toda la batería de pruebas posible:

5.1 Evaluación con los parámetros del DUC.

Desde 2004 hasta 2007, ROUGE ha sido la herramienta fundamental de evaluación en las Conferencias DUC. Entre las tareas principales y desafíos propuestos se encontraban principalmente tareas de resumen multidocumento, *question-answering* y detección de novedades. Para estas tareas, el método de evaluación usado en la tarea es ROUGE-2 y ROUGE-SU4, con *stemming* (corte de palabras a su raíz) y manteniendo *stopwords* (listas de palabras a ignorar). ROUGE-1.5.5 será procesado con los siguientes parámetros:

```
ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95
-r 1000 -f A -p 0.5 -t 0 -d
```

Método	Medida F
UHU1	**0.49456**
UHU2	**0.47399**
Word	0.46203
Copernic Summ.	0.46183
Swesum	0.45606
Mead	0.45552
Lead	0.45331
LexRank	0.44932
Pertinence Summ.	0.41740
Summ-It	0.40678

Tabla 1. Evaluación según ROUGE-1 con parámetros DUC 2005

-n 2 procesa ROUGE-1 y ROUGE-2
-x no calcula ROUGE-L
-m aplica algoritmo de Porter.
-2 4 procesa *Skip Bigram* (ROUGE-S4).
-u incluye los uni-gramas en *Skip Bigram* (ROUGE-SU)
-c 95 usa intervalo de confianza del 95%

[4] http://swesum.nada.kth.se/index-eng-adv.html
[5]http://www.computing.surrey.ac.uk/SystemQ/
[6] http://tangra.si.umich.edu/clair/md/demo.cgi
[7]http://tangra.si.umich.edu/clair/lexrank/

-f A puntuaciones promediadas sobre los múltiples modelos
-p 0.5 calcula la medida-F con *alpha* = 0.5

La mejor puntuación ha sido obtenida por nuestro algoritmo, en sus dos últimas versiones, con una mejora sobre el siguiente del 7,04%.

Método	Medida F
Copernic Summ.	0.35388
UHU1	**0.33964**
Swesum	0.33409
Lead	0.33263
Word	0.32381
Mead	0.31947
UHU2	**0.31794**
LexRank	0.30062
Pertinence Summ.	0.29155
Summ-It	0.26765

Tabla 2. Evaluación según ROUGE-2 con parámetros DUC 2005

Para bi-gramas, tamaño de palabra 2, nuestra última versión del algoritmo consigue el segundo mejor resultado, sólo superado por el generador comercial de la casa *Copernic*.

Método	Medida F
Copernic	0.36260
UHU1	**0.35972**
Lead	0.34231
Swesum	0.34213
UHU2	**0.33622**
Word	0.33173
Mead	0.32568
LexRank	0.31275
Pertinence Summ.	0.30078
Summ-It	0.28237

Tabla 3. Evaluación según ROUGE-SU4 con parámetros DUC 2005

Para ROUGE-SU4 nuestro algoritmo sigue superando claramente a la mayoría, salvo al de *Copernic*.

5.2 Resultados con los parámetros del DUC.

Podemos concluir después de esta evaluación que nuestro método es claramente uno de los mejores, en cuanto da el mejor resultado en ROUGE-1 y queda muy cerca del mejor en ROUGE-2 y ROUGE-SU4, a pesar de que estas métricas han sido seleccionadas por su conveniencia a la hora de medir la evaluación en sistemas multidocumento.

5.3 Evaluación con los parámetros ROUGE por defecto.

Dado que DUC se centra sólo en las métricas más adecuadas para sus tareas específicas, hemos decidido realizar una evaluación genérica que recoja y presente todas las métricas de que es capaz ROUGE.

```
ROUGE-1.5.5.pl -c 95 -2 -1 -U -r 1000 -n 4
-w 1.2
```

-2 -1 indica que *max-gap-length* no tiene límite
-U procesa los unigramas, incluso los regulares
-r 1000 remuestrea *bootstrap* 1000 veces para estimar el intervalo de confianza del 95%
-n 4 -n 2 procesa ROUGE-1, ROUGE-2, ROUGE-3 y ROUGE-4
-w 1.2 con un factor de peso de 1.2 para WLCS

5.4 Resultados con los parámetros ROUGE por defecto.

Analicemos los resultados de la Tabla 4. Aparecen en las filas los distintos métodos de generación de resúmenes, mientras en las columnas aparecen cada uno de los diversos métodos de evaluación que proporciona ROUGE. En cada columna se ha destacado en negrita el mejor resultado. Además, en la celda correspondiente a nuestro método, se indica entre paréntesis el lugar en el ranking que ocupa nuestro sistema para esa métrica ROUGE.

En este caso, los resultados de nuestros métodos han sido los más eficaces para los evaluadores ROUGE-1, ROUGE-L, ROUGE-S* y ROUGE-SU*, ocupando siempre las primeras cuatro posiciones.

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-S*	ROUGE-SU*	ROUGE-W-1.2
Copernic	0.4545	**0.3526**	**0.3273**	**0.3165**	0.4472	0.1868	0.1893	**0.1961**
Lead	0.4437	0.3307	0.3028	0.2909	0.4322	0.1620	0.1646	0.1936
Mead	0.4433	0.3181	0.2825	0.2642	0.4171	0.1877	0.1902	0.1853
Pertinence	0.4113	0.2902	0.2545	0.2398	0.3940	0.1493	0.1515	0.1432
Summ-It	0.3949	0.2648	0.2324	0.2197	0.3777	0.1383	0.1406	0.1482
Swesum	0.4493	0.3323	0.3009	0.2871	0.4322	0.1701	0.1727	0.1769
UHU1	**0.4834**	0.3372 (2)	0.2978 (4)	0.2843 (4)	**0.4657**	**0.2060**	**0.2085**	0.1881 (4)
UHU2	0.4635	0.3151	0.2772	0.2638	0.4449	0.1931	0.1956	0.1722
Word	0.4527	0.3223	0.2782	0.2588	0.4287	0.1723	0.1749	0.1675

Tabla 4. Tabla comparativa de evaluación con ROUGE, parámetros por defecto

También se ha de destacar que a medida que los N-gramas aumentan de tamaño, los resultados empeoran, si bien no podemos aún concretar la causa.

Si bien los resultados han sido muy buenos, hemos de ser conscientes de ciertos problemas intrínsecos a la tarea y a la solución aportada. Esta estrategia centrada en el concepto médico puede dar lugar a resúmenes más inconexos, como parece que indican los resultados para ROUGE mayores que 1. Sin duda, la mejora de la legibilidad del resumen habrá de plantearse como objetivo.

Las versiones de ROUGE L, W y S intentan arreglar ciertos problemas derivados de la traducción (aumento del espaciado entre palabras, cambios en el orden, etc...). Los resultados obtenidos en estas versiones han sido muy buenos, lo que podría derivarse del hecho de que otros métodos hagan un especial hincapié en la búsqueda de grupos de términos significativos o relevantes, algo no prioritario para nosotros, por lo que a medida que el n-grama a comparar aumenta de tamaño y se permiten saltos, nuestro método es favorecido por los resultados.

En definitiva, creemos que como primera evaluación y a pesar de usar un sistema de evaluación basado en términos, nuestros resultados son muy buenos, lo que parece indicar que el camino tomado y la estrategia de resolución del problema es la adecuada. Pero se ha de ser muy prudente en la evaluación de los resultados obtenidos, primero por la escasa representatividad de un corpus tan pequeño y segundo, por los malos resultados obtenidos por generadores de resúmenes genéricos de prestigio, como Mead o LexRank.

6 Conclusión

En este trabajo se ha presentado una metodología propia para la generación automática de resúmenes de texto. El método está basado en técnicas extractivas y en la representación del texto usando un grafo de frases y conceptos. El sistema hace uso de un analizador semántico que etiqueta el texto, identificando los conceptos y relaciona semánticamente los mismos, utilizando para ello un metathesauro médico todos los conceptos. Por ahora, sólo trabaja con textos en inglés.

Para conocer la calidad de los resúmenes generados hemos realizado un proceso de evaluación, valiéndonos para ello de la herramienta ROUGE. Esta herramienta evalúa automáticamente la calidad de resúmenes candidatos frente a un conjunto de resúmenes modelo, generados por jueces humanos. Mediante diversas métricas, principalmente de comparación de diversas variantes de N-gramas obtenemos valores de Cobertura, Precisión y Medida-F de cada generador automático de resúmenes.

Los resúmenes se han generado a partir de un pequeño corpus de documentos médicos del repositorio BIOMED Central, de los que dos expertos han realizado una serie de resúmenes manuales. Para obtener una más clara referencia de nuestra propuesta, se han generado resúmenes candidatos para la evaluación del mismo corpus usando una serie de generadores de resúmenes reconocidos por la comunidad investigadora.

Con la prudencia que merece el hecho de haber usado un corpus tan pequeño, parece que

los resultados confirman que el uso de conceptos del ámbito biomédico dentro de un proceso de generación extractiva de resúmenes produce mejores resultados que los basados en términos y que la propuesta presentada establece un buen *baseline* a partir del cual seguir mejorando.

Bibliografía

Afantenos, S. D., Karkaletsis, V. y Stamatopoulos P. "Summarization from Medical Documents: A Survey" en Artificial Intelligence in Medicine, 33(2):157-177. 2005.

Brin, S. y Page, L. "The anatomy of a large-scale hypertextual web search engine" en Computer Networks and ISDN Systems, 30 (1-7). 1998.

de la Villa, M., Maña, M. "Estableciendo una línea base para un generador de resúmenes extractivo basado en conceptos en el ámbito biomédico". Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, num.42 (Abril 2009)

Erkan, G. y Radev D. (2004) "LexRank: Graph-based Centrality as Salience in Text Summarization". Journal of Artificial Intelligence Research 22.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. "Summarizing text documents: sentence selection and evaluation metrics" SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States. 121-128. (1999)

Hassel, M. (2007) "Resource Lean and Portable Automatic Text Summarization", PhD-Thesis, School of Computer Science and Communication, KTH, ISBN-978-917178-704-0

Hovy, E. y Lin, C. Y. "Automatic evaluation of summaries using N-gram co-ocurrence statistics" en Proceedings of 2003 language technology conference (HLT-NAACL 2003) (Vol. 1(1), pag. 71-78). Edmonton, Canada.

Hovy, E.. Automated text summarization. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics, pp. 583-598. Oxford: Oxford University Press. (2005)

Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., y Barnett, G.O. "The Unified Medical Language System: An Informatics Research Collaboration", Journal of the American Medical Informatics Association, 5(1), 1-11. 1998.

Lin, C-Y. (2004) "ROUGE: a Package for Automatic Evaluation of Summaries" en Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain.

Mani, I. "Automatic Summarization" John Benjamins, Amsterdam / Philadelphia. (2001)

Milhacea R. and Tarau P: "TextRank: Bringing Order into Texts". In Proceedings of Empirical Methods in Natural Language Processing. ACL, 404-411, 2006.

Over P., Yen J. (2003) "Intrinsic Evaluation of Generic News Text Summarization Systems" DUC 2003. Workshop on Text Summarization. May 31-June 1, 2003. Edmonton, Canada

Papineni, K., S. Roukos, T. Ward, and W-J. Zhu. "BLEU: A method for automatic evaluation of machine translation". Research Report RC22176, IBM. (2001)

Plaza L., Díaz A. and Gervás P.:"Concept-graph based Biomedical Automatic Summarization using Ontologies" In Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing (COLING 2008).

Radev D.R., Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhang Zhu (2004) "MEAD - a platform for multidocument multilingual text summarization". En Proceedings of LREC 2004, Lisbon, Portugal.

Reeve, L.H., Han, H., Brooks, A.D. "The use of domain-specific concepts in biomedical text summarization" en Information Processing and Management 43, 1765-1776. 2007.

Rindflesh, T.C., Fiszman, M., Libbus, B. "Semantic interpretation for the biomedical research literature". Capítulo 14 del libro "Medical Informatics. Knowledge Management and Data Mining in Biomedicine " (Springer's Integrated Series in Information Systems), editores Chen, H., Fuller, S.S., Friedman C., Hersh, W. 2005