Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 107-115 recibido 21-04-2011 aceptado 24-05-2011

COMPENDIUM: Una herramienta de generación de resúmenes modular*

COMPENDIUM: A modular Text Summarization Tool

Elena Lloret y **Manuel Palomar**
Dept. Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de correos, 99 , 03080 Alicante, España
{elloret, mpalomar}@dlsi.ua.es

Resumen: En este artículo presentamos COMPENDIUM, una herramienta de generación de resúmenes de textos modular. Esta herramienta se compone de un módulo central con cinco etapas bien diferenciadas: i) análisis lingüístico; ii) detección de redundancia; iii) identificación del tópico; iv) detección de relevancia; y v) generación del resumen, y una serie de módulos adicionales que permiten incrementar las funcionalidades de la herramienta permitiendo la generación de distintos tipos de resúmenes, como por ejemplo orientados a un tema concreto. Realizamos una evaluación exhaustiva en dos dominios distintos (noticias de prensa y documentos sobre lugares turísticos) y analizamos diferentes tipos de resúmenes generados con COMPENDIUM (mono-documento, multi-documento, genéricos y orientados a un tema). Además, comparamos nuestro sistema con otros sistemas de generación de resúmenes actuales. Los resultados que se obtienen demuestran que la herramienta COMPENDIUM es capaz de generar resúmenes competitivos para los distintos tipos de resúmenes propuestos.
Palabras clave: Procesamiento del Lenguaje Natural, Generación de resúmenes de textos

Abstract: This paper presents COMPENDIUM, a modular text summarization tool. On the one hand, it consists of a core module, which comprises five distinct stages: i) linguistic analysis; ii) redundancy detection; iii) topic identification; iv) relevance detection; and v) summary generation. On the other hand, it integrates additional modules, with the purpose of increasing the capabilities of the text summarization tool, thus allowing the generation of different types of summaries, such as query-focused summaries. An exhaustive evaluation has been carried out in two domains (newswire and tourist places) in order to analyze the summaries generated with COMPENDIUM (single-document, multi-document, generic, and query-focused). Moreover, a comparison between our tool and other summarizers is also performed. The results obtained show that COMPENDIUM is able to generate competitive summaries for the different types of summaries proposed.
Keywords: Natural Language Processing, Text Summarization

1. Introducción

La generación de resúmenes no es una tarea nueva, ya que los primeros intentos de producir resúmenes automáticos se llevaron a cabo a finales de los años 50 (Luhn, 1958). Sin embargo, ha experimentado una gran evolución en la última década, sobre todo desde el rápido crecimiento de Internet. La gran cantidad de información disponible en formato electrónico crece de manera exponencial, dando lugar a millones de documentos cuya magnitud dificulta en gran medida su manejo. Debido a esto, la generación de resúmenes es de gran utilidad en el desarrollo de herramientas de Procesamiento del Lenguaje Natural (PLN) puesto que permite procesar grandes volúmenes de información y presentarla de forma resumida y sencilla, de modo que ofrezca al usuario, o a otras tareas de PLN, la posibilidad de gestionar la información requerida más eficientemente.

Según la Real Academia Española (RAE), "resumir" es *"reducir a términos breves y precisos, o considerar tan solo y repetir abreviadamente lo esencial de un asunto o materia"* (DRAE, 22ª edición). De esta definición se deduce que un resumen, si es elaborado correctamente, puede servir como sustituto del

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (beca FPI BES-2007-16268 y proyectos TIN2006-15265-C06-01 y TIN2009-13391-C04-01) y por la Conselleria d'Educació de la Generalitat Valenciana (proyectos PROMETEO/2009/119 y ACOMP/2011/001).

documento completo y ahorrar así, el trabajo de leerlo en su totalidad. La realización de un resumen requiere la lectura del documento en cuestión, saber extraer los conceptos e información más relevante y finalmente, reescribir toda esa información de manera que se obtenga un texto de menor tamaño que el original. Esto no es un proceso inmediato, sino que requiere tiempo y esfuerzo por parte de las personas que efectúen el resumen. En cambio, la obtención de dichos resúmenes de forma automática implicaría que apenas bastarían pocos segundos para resumir grandes cantidades de documentos. Ante los millones de documentos existentes en la web, supondría una gran ventaja disponer de este tipo herramientas automáticas.

En este artículo se presenta COMPENDIUM como una herramienta de generación de resúmenes modular, que es capaz de generar resúmenes extractivos (selección de las frases más importantes) de uno (mono-documento) o varios documentos (multi-documento), así como también resúmenes genéricos u orientados a un determinado tema. El sistema presentado se compone de diferentes etapas, que se explicarán más adelante, y gracias a su arquitectura flexible, se pueden añadir nuevos módulos para incrementar las funcionalidades del mismo. Los resultados obtenidos demuestran que los resúmenes generados con COMPENDIUM son competitivos con respecto a otros sistemas de generación de resúmenes existentes en la actualidad.

El resto del artículo está estructurado de la siguiente manera: la sección 2 presenta un breve estado de la cuestión. La sección 3 describe COMPENDIUM, junto con los diferentes módulos que lo componen. A continuación, la sección 4 muestra la evaluación del sistema y los resultados obtenidos. Finalmente, la sección 5 presenta las conclusiones y trabajos futuros.

2. Estado de la cuestión

En la actualidad, se pueden adoptar diversos enfoques para generar resúmenes de forma automática. Estos enfoques incluyen desde técnicas estadísticas, como la frecuencia de las palabras (Lloret y Palomar, 2009), hasta técnicas basadas en el análisis del discurso (Louis, Joshi, y Nenkova, 2010), pasando también por algoritmos basados en grafos (Plaza, Díaz, y Gervás, 2008) o de aprendizaje automático (Schilder y Kondadadi, 2008).

En lo que respecta a sistemas de generación de resúmenes, uno de los más conocidos es el sistema MEAD (Radev, Blair-Goldensohn, y Zhang, 2001), que es capaz de producir resúmenes extractivos tanto mono- como multi-documento, basándose en las siguientes fuentes de conocimiento: posición de la frase, solapamiento de una frase con respecto a la primera, y medidas para calcular la similitud de una oración respecto a la oración que constituye el centro de un clúster (*centroide*). Las frases que forman parte del resumen final se seleccionan en base al resultado de combinar linealmente las características previamente expuestas. Otro sistema que también se basa en la extracción y combinación de características estadísticas (tf-idf), posicionales y de similitud, como por ejemplo la similitud de las frases con respecto al tema que queramos extraer, es SUMMA (Saggion, 2008). Ambos sistemas tratan en problema de la redundancia con medidas basadas en el cálculo de la similitud del coseno. El sistema propuesto por (Steinberger et al., 2007), diseñado especialmente para mono-documento, propone el uso de *Latent Semantic Analysis* para determinar los temas más importantes de un documento combinado con información léxica y resolución de la anáfora. SummGraph (Plaza, Díaz, y Gervás, 2008) es un método basado en grafos semánticos que es capaz de identificar las frases más importantes de un documento, y extraerlas, analizando previamente los conceptos que componen las frases del documento y las relaciones entre ellos.

Nuestro sistema, COMPENDIUM, se diferencia principalmente de los sistemas anteriormente citados en dos aspectos: i) integra un módulo específico para tratar el problema de la redundancia, basado en la implicación textual, que va más allá del cómputo de la similitud léxica entre oraciones; y ii) además de emplear características estadísticas, utiliza fuentes de información basadas en teorías lingüísticas y cognitivas. En la siguiente sección se explica de forma detallada el sistema y cada una de las etapas que lo componen.

3. Arquitectura de COMPENDIUM

COMPENDIUM es una herramienta modular, capaz de producir distintos tipos de resúmenes de textos. Como entrada puede recibir uno o más documentos (mono- o multi-documento) y como salida, puede producir

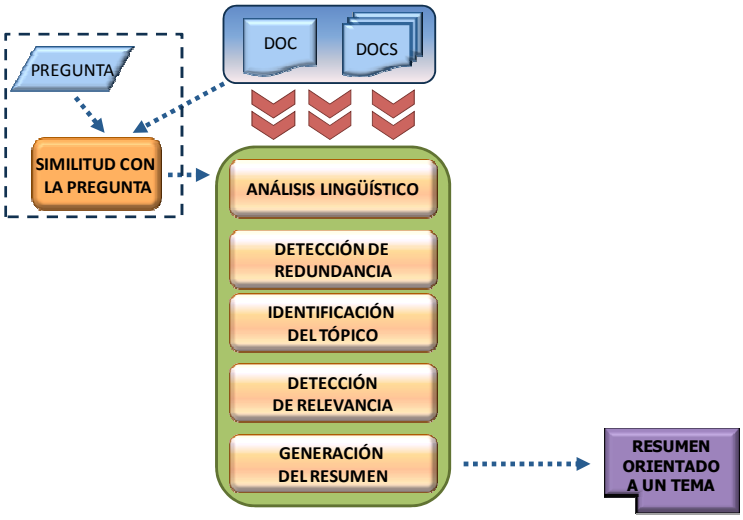

Figura 1: Arquitectura general de COMPENDIUM.

resúmenes genéricos u orientados a un tema en concreto. El objetivo de los resúmenes generados es que proporcionen al usuario las ideas más importantes de los correspondientes documentos fuente, y por lo tanto se pretende que estos resúmenes sean informativos. Finalmente, es importante destacar el hecho de que, por el momento, COMPENDIUM ha sido desarrollado y evaluado sólo para el inglés, lo cual no quita que no se pueda adaptar y probar para otros idiomas, como el castellano. Esto queda pendiente para investigaciones futuras.

En el proceso de generación de resúmenes que proponemos se pueden distinguir dos tipos de módulos. Por un lado, el núcleo de la herramienta (módulo central) está formado por cinco etapas, a partir de las cuales se producen extractos genéricos. Por otro lado, los módulos adicionales se proponen con el objetivo de incrementar las funcionalidades de COMPENDIUM, para producir tipos de resúmenes concretos, como por ejemplo resúmenes orientados a un tema concreto. Cada uno de estos módulos a su vez contiene una serie de etapas, dependiendo del objetivo que se persiga.

La figura 1 muestra la arquitectura general de COMPENDIUM, dónde la parte central representa el núcleo de la herramienta, mientras que los procesos enmarcados en rectángulos con bordes discontinuos se refieren a los módulos adicionales. Como se observa en la figura, por el momento solamente se ha desarrollado el módulo de *similitud con la pregunta*.

A continuación se hace una distinción entre las etapas que integran el núcleo de la herramienta y las que se consideran adicionales, y se explica con detenimiento cada una de ellas.

3.1. Núcleo de COMPENDIUM

Las etapas que forman el núcleo central de COMPENDIUM son: i) análisis lingüístico; ii) detección de redundancia; iii) identificación del tópico; iv) detección de relevancia; y v) generación del resumen.

3.1.1. Análisis lingüístico

En esta etapa se realiza un preprocesado básico del texto o de los textos de entrada, que consiste en:

- **Segmentación de oraciones.** El texto de entrada se segmenta en oraciones[1], ya que esta es la unidad que vamos a considerar para generar el resumen.
- **Segmentar en *tokens*.** Además de segmentar el texto de entrada en frases, también debemos segmentarlo en *tokens*[2] para poder luego calcular la frecuencia de aparición de cada uno, o bien

[1] http://duc.nist.gov/duc2004/software/
[2] http://cogcomp.cs.illinois.edu/page/tools_view/8

distinguir si se trata de una *stop word* o no.

- **Realizar *stemming*.** Este proceso nos permitirá obtener la raíz de una palabra[3].
- **Identificación de *stop words*.** Este proceso nos permitirá descartar palabras que carecen de información semántica, como artículos, preposiciones, etc., que no son necesarias para determinar la relevancia de una oración en el texto. Para poder identificarlas correctamente, nos basamos en una lista predefinida de *stop words* para el inglés[4].

3.1.2. Detección de redundancia

El objetivo de esta fase es detectar y eliminar la información redundante de un documento, para evitar así que el resumen contenga información repetida. Para lograr este objetivo, nos basamos en un módulo de reconocimiento de la implicación textual (TE) (Ferrández-Escámez, 2009), que nos indicará, dadas dos oraciones si una se puede deducir de la otra. Este sistema se basa en el cómputo de un conjunto de medidas léxicas (como por ejemplo, distancia de *Leveshtein*, *Smith-Waterman*, similitud del coseno), sintácticas (árboles de dependencia) y semánticas basadas en *WordNet 3.0*[5], aplicando un clasificador SVM con el objetivo de tomar la decisión final.

La idea principal realtiva al uso de la implicación textual en tareas automáticas de resúmenes siguiendo el enfoque propuesto, reside en conseguir un conjunto preliminar de oraciones formado por aquéllas que no tienen relación de implicación con ninguna otra frase del documento. La identificación de dichas relaciones de implicación ayudan a que el resumen final contenga la menor redundancia posible.

El siguiente ejemplo ilustra cómo aplicamos esta técnica para evitar que el resumen contenga información repetida. Partiendo del siguiente conjunto de frases:

$$S_1\ S_2\ S_3\ S_4\ S_5\ S_6$$

para obtener aquellas que no son redundantes, calculamos si existe una relación de implicación entre ellas, de la siguiente manera:

$$FrasesNoRedundantes = \{S_1\}$$
$$FrasesNoRedundantes \longrightarrow implica \longrightarrow S_2 \Rightarrow NO$$
$$FrasesNoRedundantes = \{S_1, S_2\}$$
$$FrasesNoRedundantes \longrightarrow implica \longrightarrow S_3 \Rightarrow NO$$
$$FrasesNoRedundantes = \{S_1, S_2, S_3\}$$
$$FrasesNoRedundantes \longrightarrow implica \longrightarrow S_4 \Rightarrow SI$$
$$FrasesNoRedundantes = \{S_1, S_2, S_3\}$$
$$FrasesNoRedundantes \longrightarrow implica \longrightarrow S_5 \Rightarrow SI$$
$$FrasesNoRedundantes = \{S_1, S_2, S_3\}$$
$$FrasesNoRedundantes \longrightarrow implica \longrightarrow S_6 \Rightarrow NO$$
$$FrasesNoRedundantes = \{S_1, S_2, S_3, S_6\}$$

Finalmente, obtenemos un conjunto de frases (S_1, S_2, S_3, S_6) no redundantes formado por las frases que no son implicadas entre sí.

3.1.3. Identificación del tópico

A pesar de que la frecuencia de las palabras fue de las primeras técnicas en utilizarse para generar resúmenes de forma automática (Luhn, 1958), todavía se sigue utilizando, ya que es una técnica sencilla de aplicar que obtiene muy buenos resultados.

Por lo tanto, usamos la frecuencia de las palabras para obtener el tema o temas principales de un documento, de tal manera que las palabras con mayor frecuencia en un documento (sin tener en cuenta las *stop words*, que han sido previamente eliminadas) indican cuáles son los tópicos más importantes del mismo.

3.1.4. Detección de relevancia

Esta etapa se encarga de asignar un peso a cada oración del documento, en función de su relevancia. Este peso se calcula combinando la frecuencia de cada término obtenido en la etapa anterior con el Principio de la Cantidad de Codificación (Givón, 1990).

Este principio es de origen lingüístico-cognitivo y establece que mientras más importante es la información, más prominente, más evidente y larga será el medio de codificación que la represente. Esto significa que un elemento encargado de presentar una determinada información en un texto, recibirá una codificación que será más o menos larga, en función del grado de relevancia que tenga dicha información en el texto. En cambio, si se trata de información menos importante, ésta se codificará con menor peso léxico. En (Ji, 2007), se ha demostrado que este principio se

[3] http://tartarus.org/ martin/PorterStemmer/

[4] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

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

cumple en textos escritos y además, está directamente relacionado con otro principio de carácter cognitivo, que es el Principio de la cantidad, la atención y la memoria (Givón, 1990), cuyas premisas son: (1) la codificación más prominente y distinta atraerá más la atención del receptor, y (2) la información que atrae más la atención se memoriza, almacena y recupera de forma más eficiente.

Como unidad de codificación, decidimos seleccionar los sintagmas nominales, puesto que son capaces de contener más o menos información según lo que se quiera transmitir, al poder incorporar modificadores (determinantes, adjetivos, nombres, o incluso cláusulas de relativo) que permiten aclarar y dar más información sobre un determinado sustantivo.

Nuestra hipótesis de trabajo al utilizar este principio es que las oraciones que contengan sintagmas nominales más largos formados por las palabras que constituyen el tema principal del documento, harán que dichas frases reciban mayor peso para ser seleccionadas y formar parte del resumen final.

Para identificar los sintagmas nominales de una frase, utilizamos la herramienta *BaseNP Chunker*[6]. Una vez que ya tenemos los sintagmas nominales identificados en las frases de los documentos, cada una de las frase se clasificará en función de un ránking obtenido mediante la fórmula 1.

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \qquad (1)$$

donde:

r_{s_i} = representa la relevancia de la oración i,
$\#NPi$ = número de sintagmas nominales que la oración i contiene,
tf_w = frecuencia de la palabra w que pertenece al sintagma nominal *NP*.

3.1.5. Generación del resumen

El objetivo de esta fase es generar un resumen de una determinada longitud en número de palabras. En esta fase, podemos distinguir entre dos tipos de resúmenes: resúmenes genéricos (COMPENDIUM$_E$), que contienen la información más relevante de un documento y resúmenes orientados a un tema concreto (COMPENDIUM$_{QE}$), que deben contener solamente la información más importante relacionada con ese tema.

[6] ftp://ftp.cis.upenn.edu/pub/chunker/

- **Resúmenes genéricos (COMPENDIUM$_E$).** Una vez calculada la relevancia de cada frase de los documentos, las frases cuyo valor de relevancia sea mayor serán seleccionadas para pertenecer al resumen final hasta alcanzar la longitud de resumen deseada.

- **Resúmenes orientados a un tópico (COMPENDIUM$_{QE}$).** Para generar este tipo de resúmenes, tenemos que tener en cuenta, además de la etapa de detección de redundancia, el módulo adidional de *similitud con la pregunta* (Sección 3.2.1). Una vez obtenidos estos dos valores, se combinarán mediante la fórmula 2, donde β puede obtener valores que oscilen entre 0 y 1, dependiendo de si se quiere dar más importancia a la relevancia de la frase o a la similitud con la pregunta. Las frases que obtengan mayor puntuación serán las que seleccionen para formar el resumen final.

$$Sc_{s_i} = (1+\beta^2)\frac{r_{s_i} * qSim_{s_i}}{\beta^2 * r_{s_i} + qSim_{s_i}} \qquad (2)$$

donde:

Sc_{s_i} = es la puntuación final para la frase i,
r_{s_i} = es la relevancia de la frase i,
$qSim_{s_i}$ = es la similitud entre la pregunta (o tema) y la frase i.

3.2. Etapas adicionales

Para poder generar resúmenes orientados a un tema en concreto integramos una etapa adicional (*similitud con la pregunta*) que nos permite identificar qué frases del documento tratan sobre dicho tema.

3.2.1. Similitud con la pregunta

Previamente a la ejecución de las etapas que componen el núcleo central de COMPENDIUM y que nos permiten eliminar la información redundante, identificar los temas principales del documento y seleccionar la información más relevante, establecemos esta etapa para determinar qué frases están relacionadas con una pregunta o tema inicial (por ejemplo, "lugares de interés para visitar en Alicante") que se proporciona junto con los documentos a resumir.

La similitud de cada una de las frases del documento con respecto a la pregunta se calcula usando la medida del coseno que está implementada en el herramienta *Text Similarity*[7].

Con el objetivo de dar mayor cobertura a la hora de encontrar frases similares a la pregunta proporcionada, resolvemos la anáfora utilizando la herramienta JavaRap[8]. Esto puede ayudar a que la medida de similitud utilizada (en nuestro caso, el coseno) sea más precisa a la hora de identificar frases similares con la pregunta. Por ejemplo, si la pregunta es *"Euston Railway Station"* y una de las frases es *"It is located in London"*, si no resolvemos la anáfora para esta frase, la fórmula del coseno no detectará ningún tipo de similitud entre estas dos frases.

Por tanto, el peso que se corresponde con la similitud entre las frases y la pregunta se calcula de acuerdo a la fórmula 3:

$$qSim_{s_i} = SimCoseno(S_i, P) \quad (3)$$

donde:

$SimdCoseno(S_i, P)$ = es la similitud calculada en base al coseno entre la frase S_i y la pregunta P.

4. Evaluación de COMPENDIUM

Para evaluar los resúmenes generados con COMPENDIUM utilizamos la herramienta ROUGE[9] (Lin, 2004). Esta herramienta, muy utilizada para la evaluación automática de resúmenes, permite obtener los valores de precisión, cobertura y F-medida, para diferentes niveles de solapamiento entre distintos resúmenes, siempre y cuando dispongamos de resúmenes de referencia. En este trabajo sólamente mostraremos los valores obtenidos para la cobertura que nos permitirán compararnos con otros sistemas. Las medidas ROUGE que usaremos para nuestra evaluación comprenden ROUGE-1 y ROUGE-2, que determinan la cobertura basada en unigramas y bigramas, respectivamente, entre el resumen candidato y el resumen (o resúmenes) modelo; ROUGE-L, que se basa en obtener la subsecuencia común más larga, consecutiva o no, entre dos textos y ROUGE-SU4, que permite calcular los bigramas comunes entre un resumen automático y uno de referencia, permitiendo un número de palabras entre ellos no superior a cuatro.

Concretamente, COMPENDIUM se ha evaluado sobre un conjunto de documentos, que pertenecen a dos dominios distintos: noticias de prensa y documentos sobre lugares de interés turístico. Para primer dominio se han utilizado los datos de las competiciones DUC[10] de las ediciones del 2002, 2003, y 2004. Por otro lado, para los documentos sobre lugares turísiticos se ha usado el corpus creado por Aker y Gaizauskas (2010).

Corpus	**No. clústers**	**No. docs.**	**Longitud media**
DUC 2002	59	567	630
DUC 2003	30	298	669
DUC 2004	50	500	601
Lugares turísticos	308	3.080	690

Cuadro 1: Propiedades de los corpus utilizados.

El cuadro 1 muestra a modo de resumen el número de clústers y documentos para cada conjunto de datos, así como también la longitud media de los documentos en número de palabras.

En los siguientes apartados se muestran los resultados obtenidos junto con una discusión de los mismos.

4.1. Noticias de prensa

Como ya se ha comentado anteriormente, para evaluar COMPENDIUM sobre este tipo de documentos, usamos los datos proporcionados en las conferencias DUC. Estos documentos nos permitirán evaluar COMPENDIUM$_E$ para la generación de resúmenes genéricos mono- y multi-documento. Además, estos corpus contienen resúmenes modelo elaborados por expertos humanos, que serán los que utilicemos para evaluar nuestro sistema. Siguiendo las mismas guías que en dicha competición, generamos resúmenes genéricos mono-documento y multi-documento de 100 palabras cada uno.

El cuadro 2 muestra el valor de la cobertura para distintas métricas de ROUGE obtenidas para COMPENDIUM$_E$ evaluado sobre el corpus de resúmenes mono-documento de las conferencias DUC 2002. En el mismo cuadro se observa los resultados obtenidos para

[7] http://www.d.umn.edu/tpederse/text-similarity.html

[8] http://aye.comp.nus.edu.sg/qiu/NLPTools/JavaRAP.html

[9] http://berouge.com/default.aspx

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

Mono-documento (DUC 2002)				
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-SU4**
COMPENDIUM$_E$	**0,46008**	0,20431	**0,41744**	**0,22399**
Mejor stma. DUC	0,42776	**0,21769**	0,38645	0,17315
Lead	0,41132	0,21075	0,37535	0,16604
Random	0,29963	0,11095	0,09004	0,27951

Cuadro 2: Resultados de COMPENDIUM$_E$ para resúmenes mono-documento.

COMPENDIUM$_E$ en comparación con otros sistemas de resúmenes (el que mejores resultados obtuvo en dicha edición del DUC y dos *baselines*: *Lead*, seleccionando las n primeras frases del documento, y *Random*, seleccionando frases aleatoriamente).

De manera análoga, el cuadro 3 muestra los resultados obtenidos para la tarea multi-documento de las ediciones DUC 2002, DUC 2003 y DUC 2004.

Multi-documento (DUC 2002, 2003 y 2004)			
ROUGE-1	**2002**	**2003**	**2004**
COMPENDIUM$_E$	0,30341	0,29355	0,31362
Mejor stma. DUC	0,35151	0,37980	0,38232
Lead	0,22771	0,20967	0,31293
Random	0,27131	0,27559	0,29680

Cuadro 3: Resultados de COMPENDIUM$_E$ para resúmenes multi-documento.

Como se observa, los resúmenes mono-documento obtienen mejores resultados (46 % para ROUGE-1) que los multi-documento (30 % de media). Para mono-documento, COMPENDIUM$_E$ mejora en todas las medidas ROUGE evaluadas (excepto para ROUGE-2) sobre el resto de resúmenes comparados. El incremento de mejora obtenido es de un 10 % de media con respecto al mejor sistema que participó en DUC 2002 y de un 14 % de media sobre *Lead*. Por otra parte, para los resúmenes multi-documento, COMPENDIUM$_E$ mejora los resultados obtenidos por los baselines *Lead* y *Random* en un 26 % y 8 %, respectivamente, promediando los resultados para las tres ediciones del DUC. Sin embargo, los resúmenes multi-documento generados no superan los resultados de los mejores sistemas en ninguna de estas ediciones. La tarea de generación de resúmenes multi-documento es más difícil de abordar que la tarea mono-documento. Para el caso de resúmenes multi-documento, COMPENDIUM no emplea ninguna técnica específica, sino que considera todos los documentos de entrada como uno solo. Esta limitación la solventaremos en trabajos futuros, a partir del análisis de otras estrategias más apropiadas para abordar la tarea de generación de resúmenes multi-documento.

4.2. Documentos sobre lugares turísticos

Para evaluar COMPENDIUM sobre este dominio, usamos el corpus creado por (Aker y Gaizauskas, 2010) que contiene 308 imágenes que representan a diferentes lugares o momumentos del mundo (por ejemplo, el Big Ben). Cada lugar está asociado con 10 documentos recuperados de Internet. Además, al igual que el dominio periodístico, disponemos de 932 resúmenes manuales en total (hasta un máximo de cuatro para cada lugar) de unas 200 palabras aproximadamente cada uno. Utilizando este corpus generamos resúmenes multi-documento genéricos y orientados a un tema[11] con el objetivo de comparar primero las versiones genéricas y orientadas a un tema concreto de COMPENDIUM para este escenario de evaluación concreto (COMPENDIUM$_E$ y COMPENDIUM$_{QE}$, respectivamente). Posteriormente, se compararán los mejores resultados con otros sistemas.

El cuadro 4 muestra el valor de la cobertura obtenida para COMPENDIUM para las métricas ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) y ROUGE-SU4 (R-SU4). De estos resultados se puede concluir que los resúmenes orientados a un tema (COMPENDIUM$_{QE}$) son más apropiados en este caso que los resúmenes genéricos, ya que los resultados para COMPENDIUM$_{QE}$ mejoran en un 3 % y además esta mejora es estadísticamente significativa (test de Wilcoxon).

Los resultados obtenidos son lógicos, puesto que los resúmenes orientados a un tema contendrán información relacionada con el lu-

[11] Como tema consideramos el nombre del lugar de interés.

	R-1	R-2	R-L	R-SU4
COMPENDIUM$_E$	0,35875	0,08551	0,33083	0,13371
COMPENDIUM$_{QE}$	**0,36298***	**0,08864***	**0,33551***	**0,13892***

Cuadro 4: Evaluación de COMPENDIUM$_E$ y COMPENDIUM$_{QE}$ en el dominio de lugares turísticos.

gar turístico, que es lo que nos interesa. Ahora vamos a comparar los resultados obtenidos con COMPENDIUM$_{QE}$ con otros enfoques para la generación de resúmenes. Concretamente, los sistemas que utilizamos para esta comparación son: SUMMA (Saggion, 2008) que combina características estadísticas, posicionales y de similitud, y un enfoque basado en modelos de lenguaje propuesto por (Aker y Gaizauskas, 2009)[12]. Adicionalmente, generamos un *baseline*, que extrae como resumen las primeras 200 palabras del artículo de la Wikipedia correspondiente a un lugar.

	R-1	R-2	R-L	R-SU4
Wikipedia	0,357	0,096	0,329	0,142
COMPENDIUM$_{QE}$	**0,363**	0,089	**0,336**	0,139
Modelos lenguaje	-	0,071	-	0,119
SUMMA	0,298	0,066	0,273	0,110

Cuadro 5: Comparación de los resultados obtenidos por COMPENDIUM$_{QE}$ con otros sistemas.

Los resultados obtenidos demuestran que COMPENDIUM$_{QE}$ se comporta mejor que el enfoque basado en modelos de lenguaje y SUMMA, incrementando los resultados en un 21 % y 27 % para estos dos sistemas respectivamente. Por el contrario, ninguno de los sistemas de generación de resúmenes consigue superar los resultados obtenidos utilizando las 200 primeras palabras de la Wikipedia para las métricas R-2 y R-SU4. Esto se debe a que estos resúmenes considerados como *baseline* son difíciles de superar, debido principalmente a que estos artículos han sido creados por humanos, y además en las primeras líneas, la información que contienen está directamente relacionada con el tópico. Aun así, es importante remarcar que COMPENDIUM$_{QE}$ supera a la Wikipedia en las métricas R-1 y R-L. Para el resto de métricas evaluadas, los resultados obtenidos con COMPENDIUM$_{QE}$ son muy prometedores, sobre todo para la métrica R-SU4, que es bastante similar al resultado obtenido por Wikipedia. Al igual que ocurría con las noticias de prensa, para generar resúmenes multidocumento no utilizamos ninguna técnica sofisticada, con lo que mejorando este aspecto de COMPENDIUM, se espera que los resultados mejoren y logren superar los resultados de la Wikipedia para todas las métricas de ROUGE.

[12]Para este enfoque sólo disponemos de los resultados para ROUGE-2 y ROUGE-SU4

5. Conclusión y trabajo futuro

En este artículo se ha presentado una herramienta para la generación de resúmenes, COMPENDIUM. Se trata de un sistema modular, que permite la incorporación de etapas y módulos específicos para generar distintos tipos de resúmenes, manteniendo el mismo núcleo central del sistema. Además, otra de las novedades es la incorporación de una etapa específica para identificar y eliminar la información redundante basada en la resolución de la implicacioń textual, así como también técnicas basadas en principios lingüísticos y cognitivos para detectar oraciones relevantes en los documentos. La evaluación del sistema en dos dominios distintos y para diferentes tipos de resúmenes demuestra que COMPENDIUM obtiene resultados competitivos, si bien es necesario mejorar algunos aspectos. Por lo tanto, como trabajo futuro a corto plazo nos planteamos investigar en técnicas que permitan la generación de mejores resúmenes multi-documento. A medio y largo plazo, nuestro objetivo será enriquecer COMPENDIUM para que sea posible generar otros tipos de resúmenes, como por ejemplo resúmenes subjetivos o incluso resúmenes abstractivos, y para otros idiomas como el castellano.

Bibliografía

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*.

Aker, A. y R. Gaizauskas. 2010. Model summaries for location-related images. En *Proceedings of the 7th Language Resources and Evaluation Conference*.

DRAE. 22ª edición. Diccionario de la lengua española. http://rae.es.

Ferrández-Escámez, Óscar. 2009. *Textual Entailment Recognition and its Applicability in NLP Task.* Ph.D. tesis, Universidad de Alicante.

Givón, Talmy, 1990. *Syntax: A functional-typological introduction, II.* John Benjamins.

Ji, Shaojun. 2007. A textual perspective on givon's quantity principle. *Journal of Pragmatics*, 39(2):292–304.

Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. En *Proceedings of ACL Text Summarization Workshop*, páginas 74–81.

Lloret, Elena y Manuel Palomar. 2009. A gradual combination of features for building automatic summarisation systems. En *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, páginas 16–23.

Louis, Annie, Aravind Joshi, y Ani Nenkova. 2010. Discourse indicators for content selection in summarization. En *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, páginas 147–156.

Luhn, H. P. 1958. The automatic creation of literature abstracts. En *Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization, páginas 15–22. MIT Press.

Plaza, Laura, Alberto Díaz, y Pablo Gervás. 2008. Uso de grafos de conceptos para la generación automática de resúmenes en biomedicina. *Sociedad Española para el Procesamiento del Lenguaje Natural*, (41):191–198.

Radev, Dragomir R., Sasha Blair-Goldensohn, y Zhu Zhang. 2001. Experiments in Single and Multi-Document Summarization using MEAD. En *Proceedings of the 1st Document Understanding Conference*, páginas 1–7.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Languages*, 49:103–125.

Schilder, Frank y Ravikumar Kondadadi. 2008. Fastsum: Fast and accurate query-based multi-document summarization. En *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, páginas 205–208.

Steinberger, Josef, Massimo Poesio, Mijail A. Kabadjov, y Kerel Ježek. 2007. Two Uses of Anaphora Resolution in Summarization. *Information Processing & Management*, 43(6):1663–1680.