Generación de resúmenes de textos basados en Tecnologías del Lenguaje Humano[1]

Elena Lloret Pastor

Director: Manuel Palomar Sanz

Memoria de suficiencia investigadora
Programa de doctorado: Aplicaciones de la Informática (5099)

Departamento de Lenguajes y Sistemas Informáticos

Universidad de Alicante

Julio 2009

A todas las personas que creen en mí cada día.
Gracias por vuestra comprensión, vuestros
ánimos y sobre todo, por vuestra paciencia.

Prefacio

Resumen

Este trabajo de investigación es fruto del trabajo desarrollado en el programa de doctorado *Aplicaciones de la informática (5099)* durante los cursos 2007-08 y 2008-09, en el Departamento de Lenguajes y Sistemas Informáticos (DLSI), de la Universidad de Alicante, y más concretamente, dentro del Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI), gracias a una beca de Formación de Personal Investigador (FPI) concedida por el Ministerio de Ciencia e Innovación (BES-2007-16268), que se enmarca dentro del proyecto TEXT-MESS (TIN2006-15265-C06).

El Procesamiento del Lenguaje Natural (PLN) es un campo de la Inteligencia Artificial cuyo objetivo es investigar mecanismos eficaces computacionalmente para facilitar y mejorar la comunicación hombre-máquina. La comprensión del lenguaje es una tarea todavía no resuelta, puesto que implica leer un texto, entenderlo, poder contestar preguntas, extraer la información deseada, resumirlo, etc. tal y como lo haría un humano. Cada una de estas tareas presenta un reto en sí, puesto que el lenguaje es difícil de procesar computacionalmente por la complejidad inherente al mismo y por los aspectos cognitivos en los que está implicado (cuando las personas utilizan el lenguaje como medio de comunicación, poseen un conocimiento del mundo previo que influye en cómo se utilizará el lenguaje y con qué finalidad). Esta investigación se centra solamente en una pequeña parcela de todo el amplio abanico de posibilidades que el PLN ofrece, concretamente nos dedicaremos a investigar en detalle la tarea de generación automática de resúmenes.

Como primer paso en la citada investigación, este trabajo presenta por un lado, un estado de la cuestión en la tarea de generación de resúmenes, donde se describen diferentes sistemas y las técnicas que se utilizan. Además, se realiza un estudio de las principales herramientas de evaluación existentes y los problemas asociados. Por otro lado, se presenta una propuesta para un sistema de generación de resúmenes de texto, combinando diferentes características y técnicas, así como su evaluación y su aplicación en el ámbito de la minería de opiniones. Por último, se plantea un marco de evaluación cualitativo para determinar la calidad de un resumen en función de unos criterios lingüísticos establecidos.

Estructura

Esta memoria de investigación se divide principalmente en dos partes:

- En la primera parte se realiza una breve descripción de las asignaturas cursadas durante el periodo de docencia en el marco del programa de doctorado *Aplicaciones de la informática*.
- La segunda parte corresponde al trabajo desarrollado durante el periodo de investigación tutelado.

La memoria finaliza con una apartado de publicaciones relacionadas con este trabajo y una bibliografía que ha servido de base durante toda la investigación desarrollada.

Índice general

I Memoria del periodo de docencia (2007-2008) 13

1. Asignaturas cursadas 15
1.1. Tecnologías del lenguaje humano 15
1.2. Extracción de información textual 17
1.3. Uso y diseño de ontologías 19
1.4. Búsquedas inteligentes de información en la web 21
1.5. Cómo se escriben y se publican trabajos científicos 23

II Memoria del trabajo de investigación tutelado 25

1. Trabajo de investigación tutelado 27
1.1. Obtención de resúmenes de documentos 27

2. Introducción 29
2.1. Motivación 29
2.2. Objetivos 30
2.3. Organización del trabajo de investigación 31

3. Estado de la cuestión en la generación de resúmenes 33
3.1. Resúmenes genéricos 38
3.1.1. Enfoques basados en conocimiento 38
3.1.1.1. Enfoques basados en grafos 40
3.1.1.2. Enfoques lingüísticos 41
3.1.2. Enfoques basados en corpus 42
3.2. Resúmenes orientados a tareas 43

4. Estado de la cuestión en la evaluación de resúmenes 49
4.1. Evaluación cuantitativa 50
4.1.1. Problemas de la evaluación cuantitativa 53
4.2. Evaluación cualitativa 56

5. Sistema de generación de resúmenes basado en conocimiento 59
5.1. Frecuencia de palabras 60
5.2. Implicación textual 61
5.3. Principio de la cantidad de codificación 63
5.4. Combinación de las técnicas 64
5.5. Evaluación de la propuesta 67
5.5.1. Análisis de los resultados 69
5.6. Investigación en progreso 70
5.6.1. Resúmenes multi-documento 70
5.6.2. Extensión a otros dominios 72

6. Generación de resúmenes de opiniones 75
6.1. Participación en TAC 2008 75
6.1.1. Resultados 80
6.2. Mejora del sistema después del TAC 2008 82
6.2.1. Problemas en la evaluación 83
6.2.2. Mejora de los resultados del TAC 84
6.2.3. Sistema de resúmenes genérico 86

7. Marco de evaluación cualitativa de resúmenes 89

8. Conclusiones y trabajo en progreso 93
8.1. Conclusiones 93
8.2. Trabajo futuro 96

Bibliografía 99

A. Aportaciones científicas a través de publicaciones 111

Índice de cuadros

3.1. Características más destacadas de los sistemas más representativos 47

4.1. Características de los métodos de evaluación existentes 53

5.1. Evaluación del sistema de resúmenes propuesto 68
5.2. Comparación de resultados con los sistemas del DUC 2002 69
5.3. Resultados para multi-documento 71
5.4. Resultados para los cuentos infantiles 73

6.1. Resultados de la participación en TAC 2008 82
6.2. Mejora del sistema incorporando TE (medida-F) 85
6.3. Análisis de la corrección gramatical (*Grammaticality*) 85
6.4. Resultados del sistema genérico 87

Índice de figuras

3.1. Taxonomía propuesta por Karen Spärck-Jones 34
3.2. Taxonomía propuesta por Hovy y Lin 35
3.3. Taxonomía propuesta por Mani y Maybury 36

4.1. Precisión, cobertura y medida-F 51
4.2. Documento AP880911-0016 55
4.3. Ejemplo de resúmenes humanos (A, B) y automático (C) 56

5.1. Ejemplo documento *AP880911-0016* 60
5.2. Ejemplos de implicación textual 61
5.3. Cálculo de la relevancia aplicando el CQP. 65
5.4. Arquitectura del sistema 66
5.5. Resumen generado por nuestra propuesta 68

6.1. Ejemplo de tema, preguntas y snippets 76
6.2. Sistema de resúmenes de opiniones 77
6.3. Ejemplo de “Procesamiento de la pregunta” 78
6.4. Ejemplos de resúmenes generados 81

7.1. Propuesta de evaluación cualitativa 90

Parte I

Memoria del periodo de docencia (2007-2008)

Capítulo 1

Asignaturas cursadas

1.1. Tecnologías del lenguaje humano

Código	62571
Profesorado	Dr. Antonio Ferrández Rodríguez Dr. Antonio Molina Marco
Créditos	4
Tipo	Fundamental
Calificación obtenida	Sobresaliente

El Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial. Todo sistema de PLN intenta emular el comportamiento lingüístico humano y para lograrlo, investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre-máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales.

El objetivo de esta asignatura es introducir al alumno en el área del PLN, impartiendo una serie de fundamentos teóricos relacionados con este campo, para que el alumno adquiera la capacidad suficiente para estudiar y desarrollar la tecnología apropiada para cada aplicación.

Para lograr dichos objetivos, el contenido de la asignatura se estructura en los siguientes módulos:

- Preliminares del Procesamiento del Lenguaje Natural
- Análisis léxico
- Análisis sintáctico
- Interpretación semántica
- Interpretación contextual

- Disertaciones bibliográficas

Además de estudiar los módulos que componen cualquier aplicación de PLN, se estudian también algunas aplicaciones que necesitan de estos módulos, tales como traducción automática, acceso y extracción de información de Bases de Datos, recuperación o búsqueda de información, etc.

Esta asignatura constituye la base para el resto de asignaturas, combinando las clases magistrales (teóricas) con sesiones prácticas y seminarios participativos para fomentar la interacción y mejorar el aprendizaje.

Para obtener la calificación final de la asignatura, se realizaron dos trabajos prácticos. El primero de ellos consistió en evaluar y comparar las prestaciones de distintos etiquetadores morfosintácticos para el español, utilizando el paquete *ACOPOST*. A partir del corpus *LexEsp* y siguiendo una metodología de validación cruzada sobre diez particiones, se entrenó y se evaluó el corpus atendiendo a diferentes características: tamaño del corpus de entrenamiento, juego de categorías morfosintácticas utilizadas, o el método de suavizado para las palabras desconocidas, entre otras. Por el contrario, el segundo trabajo práctico de la asignatura consistió en realizar un conjunto de ejercicios sobre análisis sintáctico que incluía el desarrollo de una gramática en Prolog, con el objetivo de comprobar cómo dicha gramática era capaz de reconocer ciertas oraciones. Además, también se realizó una serie de experimentos con el programa SUPAR. Por una parte, esto sirvió para familiarizarnos con el programa, mientras que por otra, se utilizó SUPAR para detectar posibles errores gramaticales, a partir del análisis sintáctico de varias oraciones de un corpus dado.

En cuanto a la bibliografía recomendada para esta asignatura, destacamos como bibliografía más relevante las siguientes referencias:

- Introducción al Procesamiento del Lenguaje Natural. L. Moreno, M. Palomar, A. Molina, A.Ferrández. Servicio de publicaciones de la Universidad de Alicante. 1999
- Tecnologías del Lenguaje. Martí, M.A. Et al. Editorial UOC. 2003
- The Oxford Handbook of Computational Linguistics. Oxford University Press. 2003

1.2. Extracción de información textual

Código	62572
Profesorado	Dr. Rafael Muñoz Guillena Dr. Patricio Manuel Martínez Barco Dr. Horacio Rodríguez Hontoria Dr. Estela Saquete Boró
Créditos	4
Tipo	Fundamental
Calificación obtenida	Sobresaliente

La tarea Extracción de Información (EI) consiste en seleccionar automáticamente información estructurada o semiestructurada a partir de documentos escritos en lenguaje natural.

El objetivo fundamental de esta asignatura es realizar una introducción a la tarea de EI. Para ello, se sitúa la tarea de EI dentro del PLN, definiendo y explicando las capacidades que aportan los sistemas de EI, así como también sus necesidades. Se proponen también una serie de técnicas para transformar la información no estructurada en estructurada. Además, otra de las metas de esta asignatura es dar a conocer recursos y sistemas de EI, como por ejemplo LaSIE, EXIT o PROTEUS.

Los contenidos que se tratan en esta asignatura se dividen en:

- Introducción a la Extracción de Información
- El Procesamiento del Lenguaje Natural en la Extracción de Información
- Extracción de información
  - Aplicaciones de EI
  - Bibliotecas digitales: Una fuente de información que requiere extracción de información automática
- Tareas de los sistemas de EI
- Reconocimiento de Entidades
- Resolución de la correferencia
- Definición y relleno de plantillas
- Aplicaciones de la Extracción de Información

Además de las clases teóricas, se proponen también una serie de sesiones prácticas para conocer algunas herramientas y recursos, con el fin de construir un

sistema de extracción de información. Para la consecución de la asignatura, se propuso el diseño de un sistema de extracción de información muy básico mediante el desarrollo de una serie de patrones, utilizando la herramienta de extracción de información Jet, que es una herramienta diseñada para el análisis del lenguaje natural, y muy especialmente para la extracción de información. Los documentos con los que se trabajó fueron documentos en inglés que proporcionaban información sobre las víctimas americanas de operaciones militares durante el año 2007. Los tipos de información a extraer para cada víctima incluían fechas, lugares, nombres, números, etc.

Finalmente, entre todas las referencias bibliográficas existentes para esta tarea, destacan:

- Information Extraction: Technique and Challenges. R. Grishman. In Maria Teresa Pazienzia (Ed.). LNCS, 1299:10-27. Springer-Verlag 1997
- Information Extraction. J. Cowie, Y. Wilks. In R. Dale, H. Moisl and H. Somers (eds.) The Handbook of Natural Language Processing. 2000.

1.3. Uso y diseño de ontologías en procesamiento del lenguaje natural y la web semántica

Código	62573
Profesorado	Dr. Andrés Montoyo Guijarro Dr. José Luis Vicedo González Dr. German Rigau Claramunt Dr. Armando Suárez Cueto
Créditos	4
Tipo	Fundamental
Calificación obtenida	Sobresaliente

La introducción al uso de ontologías para el PLN es de gran ayuda para el alumno, ya que lo que se pretende con esta asignatura es el estudio y comprensión de los conceptos básicos sobre ontologías y sus componentes. Además, se presentan los recursos ontológicos más relevantes, y se explica con profundidad el proceso de diseño de ontologías para su aplicación en tareas de PLN y la web semántica.

La asignatura se desarrolla en torno a clases teórico-prácticas, cuyos contenidos comprenden:

- Conceptos y componentes de una ontología aplicados a PLN y web semántica
- Recursos ontológicos
- Uso de ontologías
  - Las ontologías y el Procesamiento del Lenguaje Natural
    - Las ontologías en el WSD
    - Las ontologías en la Búsqueda de Respuestas
  - Las ontologías y la web semántica
- Diseño de ontologías
- Disertaciones bibliográficas

Para la evaluación de la asignatura se realizaron varios cuestionarios on-line, así como también una práctica sobre el diseño de ontologías y etiquetado semántico. A partir de una descripción y un esquema de un sitio web ficticio, el objetivo que se perseguía era etiquetar semánticamente todas las páginas web del sitio dado, empleando una ontología de datos y una ontología de navegación, previamente diseñadas mediante el editor de ontologías *Protégé*.

Finalmente, algunas referencias bibliográficas relevantes son:

- Eduard Hovy. Using an Ontology to simplify data access. Communications of the ACM 46(1). 2003.
- Vossen, P. 2000. EuroWordNet: a Multilingual Database with WordNets in 8 languages. The ELRA Newsletter, 5(1):9-10. 2000.

1.4. Búsquedas inteligentes de información en la web

Código	62338
Profesorado	Dr. Antonio Ferrández Rodríguez Dr. Jesús Peral Cortés Dr. Fernando Llopis Pascual Dr. Luis Alfonso Ureña López
Créditos	4
Tipo	Fundamental
Calificación obtenida	Sobresaliente

La creciente expansión de Internet y el rápido desarrollo tecnológico posibilita la existencia de enormes cantidades de información en la web. Sin embargo, el problema subyacente radica en la dificultad, cada vez mayor, para encontrar la información realmente relevante para un usuario acerca de un tema concreto. Los sistemas de Recuperación de Información (RI) y de Búsqueda de Respuestas (BR) proporcionan mecanismos para poder acceder más rápidamente a la información deseada y han facilitado, en gran medida, esta tarea.

Esta asignatura se centra en conocer el funcionamiento de los buscadores de información y mejorar su funcionamiento, analizando la calidad y la precisión de los resultados devueltos por estos sistemas, y aplicando técnicas de PLN a estos buscadores. Además, en la segunda parte de la asignatura se explica el funcionamiento de los sistemas de búsqueda de respuestas (o sistemas de Question Answering), cuya misión es contestar preguntas formuladas por los usuarios, devolviendo la respuesta a la pregunta concreta que el usuario ha realizado, y no el documento completo que la contiene. Finalmente, se introducen también técnicas que permiten añadir capacidad multilingüe a dichos buscadores.

Por tanto, el temario de esta asignatura abarca las partes que a continuación se enumeran:

- Introducción
- Sistemas de búsqueda de información (Information Retrieval). Estado del arte.
- Incorporación de técnicas de lenguaje natural a los buscadores de información.
- Sistemas de búsqueda de respuestas (Question Answering).
- Recuperación de información multilingüe (Cross Language Information Retrieval).

Este curso, de manera similar a los mencionados anteriormente, también combina clases teóricas y prácticas, proporcionando en primer lugar una base teórica para ponerla luego en práctica sobre diferentes sistemas de recuperación. La evaluación de la asignatura consistió en un trabajo práctico en el cual se comparaban varios sistemas de recuperación de información (utilizando la fórmula de Kashkiel, Deviation from Randomness -DFR- y coseno pivotado). El objetivo era analizar los resultados obtenidos por cada uno de estos sistemas y compararlos con respecto a los otros. Como trabajo adicional, se propuso la lectura de un artículo del CLEF[1] de la edición del 2007, para realizar un resumen de los aspectos más importantes del artículo: qué se planteaba, cómo se abordaba el problema, qué resultados se obtenían y finalmente, que conclusiones se derivaban de los resultados obtenidos.

Como referencias bibliográficas más importantes, se deben incluir:

- Modern Information Retrieval. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Addison Wesley. 1999.
- INFORMATION RETRIEVAL BOOK: C. J. van RIJSBERGEN.

[1] http://www.clef-campaign.org/

1.5. Cómo se escriben y se publican trabajos científicos

Código	61345
Profesorado	Dr. Mikel L. Forcada Zubizarreta Dr. Juan Antonio Pérez Ortiz
Créditos	4
Tipo	Metodológica
Calificación obtenida	Sobresaliente

Escribir y redactar de forma adecuada, ya sea en inglés o en español, es una tarea esencial para el desarrollo de la ciencia, y por tanto constituye una tarea imprescindible para cualquier investigador de cualquier materia. Esta asignatura proporciona una formación a los investigadores en este ámbito, ofreciendo una serie de guías y pasos a seguir para que el alumno pueda realizar una correcta difusión de su investigación.

Los objetivos que se persiguen con la realización de este curso son, entre otros, que el alumno comprenda la necesidad de métodos especializados de documentación y comunicación para el trabajo científico y tecnológico, así como conocer los mecanismos de publicación de trabajos científicos, para aprender a preparar y realizar presentaciones orales y defender pósters en congresos. Como objetivos adicionales, se pretende dar a conocer los mecanismos básicos de obtención de documentos científicos e información en torno a ellos, con énfasis en los recursos actuales de la Universidad de Alicante, y los aspectos básicos relacionados con los derechos de autor y de explotación de los trabajos publicados.

Por lo tanto, el programa de la asignatura queda estructurado de la siguiente forma:

- Introducción
- Ciencia y comunicación
- La preparación de una comunicación científica escrita
- La preparación de una comunicación científica oral
- La búsqueda en bases de datos de publicaciones
- Preparación de documentos y presentaciones con LaTeX
- Derechos de autor y de explotación

La metodología propuesta para impartir este curso de doctorado comprende la realización de actividades en el aula, tanto de forma individual como conjunta, para introducir conceptos y discutirlos. Estas sesiones se complementan con clases

prácticas o sesiones técnicas en las que se profundiza más en determinados aspectos de la asignatura.

Como trabajo final de la asignatura, se realizaron varias presentaciones (tanto en forma de comunicación oral como póster), así como también un trabajo de investigación más extenso para la obtención de la calificación final.

Algunas referencias bibliográficas interesantes son:

- Day, R.A. (1994) ”How to Write and Publish a Scientific Paper”, Oryx.
- Kopka, Helmut; Daly, Patrick W. (2003) ”Guide to LaTeX”. Pearson Education.

Parte II

Memoria del trabajo de investigación tutelado

Capítulo 1

Trabajo de investigación tutelado

1.1. Obtención de resúmenes de documentos basada en cadenas de referencia

Código	60462
Profesorado	Dr. Manuel Palomar Sanz
Créditos	12
Tipo	Trabajo de investigación
Calificación obtenida	Sobresaliente

En este trabajo de investigación se pretende desarrollar técnicas de obtención de resúmenes de textos que toman como entrada documentos completos. Entre ellas se encontraría la resolución de la correferencia lingüística, para identificar las cadenas de correferencia existentes en el texto, y en base a estas cadenas, construir un método capaz de obtener resúmenes.

Los objetivos pedagógicos de este trabajo de investigación incluyen estudiar técnicas de Procesamiento de Lenguaje Natural aplicadas a resúmenes de documentos, construir sistemas de resúmenes de documentos, y finalmente, evaluar y comparar resultados frente a otros sistemas similares.

Para poder llevar a cabo todo este trabajo, algunas referencias bibliográficas básicas que se recomiendan son las siguientes:

- Introducción al Procesamiento del Lenguaje Natural. L. Moreno, M. Palomar, A. Molina, A.Ferrández. Servicio de publicaciones de la Universidad de Alicante. 1999
- The Oxford Handbook of Computational Linguistics. Oxford University Press. 2003

Capítulo 2

Introducción

El Procesamiento del Lenguaje Natural (PLN) es una subdisciplina de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos para facilitar la interrelación hombre-máquina, permitiendo una comunicación mucho más fluída y menos rígida que los lenguajes formales (Moreno Boronat et al., 1999).

A pesar de su sencilla definición, su puesta en práctica resulta sumamente compleja. Su dificultad radica por la naturaleza misma del lenguaje natural y por el contexto socio-cultural en el que se enmarca. Por un lado, se debe tener constancia de las estructuras propias de una lengua concreta y de los fenónemos y propiedades que en dicho lenguaje se producen. La ambigüedad, propiedad inherente en todas las lenguas naturales, o mecanismos de economía lingüística como la elipsis, son dos ejemplos de ello, ampliamente tratados debido a su complejidad de procesamiento automático. Por otro lado, se debe disponer también de un conocimiento general acerca del mundo para comprender las ideas que se pretenden transmitir a través del lenguaje. Por esta razón, el PLN puede abarcar el tratamiento del lenguaje desde documentos completos, hasta las unidades que forman las palabras, por ejemplo los morfemas. Esto da lugar a un amplio abanico de subtareas, que comprende desde aplicaciones más generales, como la *recuperación de información*, *búsqueda de respuestas*, *extracción de información*, *generación de resúmenes*, *atribución de autoría*, etc., hasta aplicaciones intermedias, tales como analizadores morfológicos, analizadores sintácticos, desambiguadores del sentido de las palabras, reconocedores de entidades, etc. que constituyen los pilares básicos permitiendo el desarrollo de aplicaciones generales.

2.1. Motivación

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. 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 del PLN que permiten, de algún manera, procesar dicha información y presentarla de forma resumida y sencilla, de modo que ofrezca al usuario, o a otras tareas del 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 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. Lamentablemente, debido a todos los retos que presenta la tarea de generación automática de resúmenes, el conseguir resúmenes perfectos (igual que los haría un humano) es todavía una utopía, ya que la "inteligencia" que pueda tener un ordenador no es comparable a la humana. Sin embargo, gracias a los esfuerzos realizados por la comunidad científica, esta tarea se va perfeccionando poco a poco, consiguiendo cada vez, resúmenes mejores y de diversos tipos, según el cometido perseguido. Así, puesto que la tarea de generación de resúmenes conlleva una gran aplicación práctica y de gran utilidad, es necesario seguir investigando en técnicas que permitan identificar la información más relevante de uno o varios documentos, y presentarla en forma de resumen.

2.2. Objetivos

Este trabajo persigue los objetivos que siguen a continuación:

- Estudio del estado de la cuestión de los sistemas de generación automática de resúmenes, analizando también la difícil tarea de evaluación, a partir de los sistemas existentes actualmente.

- Análisis de la influencia que determinadas técnicas o tareas del PLN ejercen sobre la generación de resúmenes, examinando las ventajas e inconvenientes que cada una de ellas aporta.

- Propuesta de un sistema de generación automática de resúmenes modular, que aplique y permita combinar las distintas técnicas previamente estudiadas.
- Evaluación del sistema propuesto, analizando las fortalezas y debilidades de las técnicas empleadas.
- Desarrollo y evaluación de un sistema de resúmenes de opiniones, combinando técnicas de minería de opiniones y generación de resúmenes, a través de la participación en la competición TAC[1] 2008 organizada por el NIST[2].
- Propuesta inicial de un marco de evaluación cualitativo para evaluar resúmenes de forma automática en base a unos criterios de calidad preestablecidos.

2.3. Organización del trabajo de investigación

Conforme a los objetivos planteados anteriormente, el resto del trabajo de investigación se organiza en torno a los siguientes capítulos:

- **Estado de la cuestión en generación de resúmenes** (capítulo 3). En este apartado se hace un repaso a los distintos tipos de resúmenes que podemos obtener y, en base a esos tipos, se describirán sistemas y aproximaciones existentes, proporcionando información sobre las técnicas y métodos empleados para generar resúmenes de forma automática, no sólo de documentos textuales, sino también de documentos de otras tipologías, como páginas web o blogs.
- **Estado de la cuestión en la evaluación de resúmenes** (capítulo 4). El objetivo de este capítulo es explicar los mecanismos existentes para evaluar los resúmenes automáticos. Como consecuencia, se distingue entre dos tipos de evaluación - cuantitativa y cualitativa - y se describen las herramientas y métodos existentes en función cada tipo. Además, se plantean los problemas que tiene la evaluación cuantitativa en el marco de esta tarea.
- **Sistema de generación de resúmenes basado en conocimiento** (capítulo 5). Esta sección presenta una propuesta de un sistema de resúmenes de textos, que emplea tres técnicas para detectar la información relevante dentro de un documento (frecuencia de palabras, implicación textual y el principio de la cantidad de codificación), con la posibilidad de aplicarlas de forma individual o conjunta. Además, en este mismo capítulo se explica detalladamente la experimentación realizada y se muestran los resultados obtenidos, sobre los que se hace un breve análisis.

[1] Text Analysis Conference: www.nist.gov/tac/
[2] National Institute of Standards and Technology: www.nist.gov/

- **Resúmenes de opiniones** (capítulo 6). Lo que se pretende aquí es analizar la aplicación práctica y real que puede tener un sistema de generación de resúmenes desde la perspectiva de la tarea de minería de opiniones. Para ello, se propone un nuevo sistema que genera resúmenes a partir de opiniones contenidas en blogs. Además, se analiza el comportamiento y la aplicación del sistema de resúmenes propuesto en capítulo 5 para plantear su validez en esta tarea concreta.

- **Marco de evaluación cualitativa de resúmenes** (capítulo 7). En este apartado se propone un marco de evaluación para la tarea de resúmenes, en el que a diferencia de las herramientas automáticas existentes para evaluar resúmenes, se centra en aspectos relacionados con la presentación y la estructura del resumen final, y no tanto en el contenido del mismo. Se explicará la idea subyacente a esta propuesta y se expondrán ejemplos de criterios que podrían ser evaluados.

- **Conclusiones y trabajo en progreso** (capítulo 8). Finalmente, se presentan las conclusiones y se enumeran las posibles líneas de investigación futuras para mejorar la calidad de los resúmenes obtenidos, incorporando nuevas técnicas al sistema, o bien estudiando y examinando la influencia de otras tareas del PLN.

Capítulo 3

Estado de la cuestión en la generación de resúmenes

Según la definición dada en (Spärck Jones, 2007), el objetivo de la tarea de generación de resúmenes es obtener una versión reducida del documento o documentos fuente, reduciendo su contenido de tal forma que se seleccionen y queden presentes en el resumen los conceptos más importantes de dichos documentos. Por lo tanto, de la definición de esta tarea se deduce que un resumen debe contener la información más significativa de uno o varios documentos, teniendo un tamaño considerablemente inferior al del documento(s) fuente.

Los sistemas de generación de resúmenes se pueden clasificar en base a múltiples factores. Una de las taxonomías más conocidas en este campo es la propuesta en (Spärck Jones, 1999), en la que se distinguen tres tipos de factores que pueden influir en los resúmenes: factores relacionados con la entrada (*input factors*), con la salida (*output factors*), y con la finalidad del resumen (*purpose factors*). Los factores relacionados con la entrada tratan aspectos como el género, idioma o registro. La finalidad del resumen depende de a quién vaya dirigido y del uso que se le quiera dar, ya que no es lo mismo generar un resumen para informar acerca de las últimos datos de la bolsa, que sobre los hechos históricos más importantes acaecidos en un país, por poner un ejemplo. Finalmente, la salida el resumen en sí, viene determinada por la finalidad y el objetivo que se persiga con éste. La figura 3.1 muestra la taxonomía mencionada, con las tres clases de factores comentados.

Como se puede observar de la figura, esto permite que un resumen se pueda clasificar atendiendo a diferentes criterios, algunos de los cuales se explicarán a continuación de forma más detallada. Relacionado con los factores de entrada, podemos producir un resumen a partir de un solo documento (mono-documento) o a partir de un conjunto de ellos (multi-documento). Además, no siempre debe tratarse de documentos textuales, ya que podemos realizar resúmenes a partir de otros tipos de texto, como páginas web, blogs, imágenes, vídeos, reuniones, etc. En lo que respecta al resumen generado, éste se puede conseguir siguiendo dos

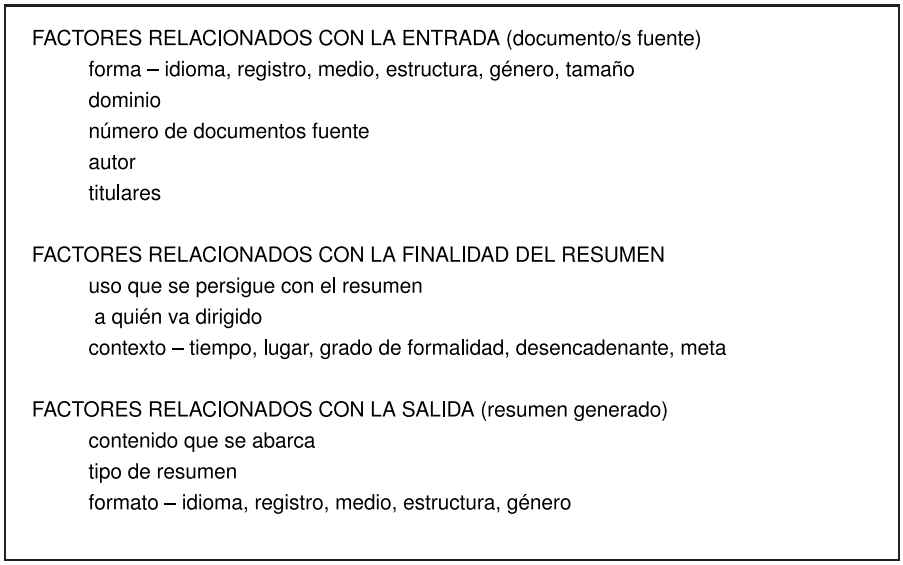

Figura 3.1: Taxonomía propuesta por Karen Spärck-Jones

posibles estrategias: extractiva o abstractiva. Si se sigue una estrategia extractiva, se seleccionarán y extraerán, literalmente, las frases más importantes sin realizar ninguna modificación sobre ellas. Sin embargo, si se opta por llevar a cabo una estrategia abstractiva para producir un resumen, será necesario realizar algún tipo de transformación en las frases seleccionadas o en los conceptos que se deseen que formen parte del resumen, de tal manera que la información que aparezca en el resumen estará expresada de una forma distinta a la que aparece en el documento original. Un método alternativo, propuesto por Horacio Saggion, que combina ambas estrategias sería el de extracción y generación, dónde se extraen informaciones de los documentos para producir representaciones abstractas, que luego se combinarán utilizando métodos de generación automática de textos. Otro aspecto a tener en cuenta en un resumen es si se trata de un resumen genérico o bien si debe estar centrado u orientado a algún tema o necesidad específica. Como veremos más adelante, hay sistemas que optan por generar resúmenes de carácter genérico, simplemente intentando captar la información más relevante, mientras que otros contienen solamente información acerca de determinados aspectos concretos del documento, por ejemplo ante una necesidad específica de un usuario. Otra distinción posible, referente a los tipos de resumen que los sistemas automáticos suelen abordar, es la distinción entre resúmenes indicativos o informativos. Dependiendo de la finalidad que se le quiera dar al resumen, optaremos por uno u otro tipo. Los resúmenes indicativos son aquellos que simplemente proporcionan unas pequeñas pinceladas acerca de un documento, de manera que podemos tener una idea de si vale la pena acudir o no al documento completo, para encontrar en él la información que buscamos. Por el contrario, si producimos un resumen informativo, lo que pretendemos es que éste contenga la

máxima información relevante posible, de tal manera que el resumen por sí solo nos proporcione la información requerida, sirviendo como sustituto del documento original. Otro factor interesante a tener en cuenta es el idioma de los documentos origen y del resumen, ya que esto también puede dar lugar a diferentes tipos de resumen. Diremos que un sistema de resúmenes es monolingüe si únicamente es capaz de producir resúmenes para un determinado idioma (por ejemplo, el español). En este caso, los documentos fuente y el resumen estarán en el mismo idioma. Si por el contrario, un sistema produce resúmenes para diferentes lenguas (inglés, español, italiano, etc.) nos encontramos ante un sistema multilingüe, pero tanto los documentos fuente como el resumen siguen estando en el mismo idioma. Ahora bien, si el sistema no sólo no es capaz de tratar varios idiomas, sino que también puede generar resúmenes en un idioma determinado (por ejemplo, español) aunque los documentos a partir de los que tiene que generar el resumen estén en un idioma distinto (por ejemplo, italiano), estaremos ante el caso de un sistema *cross-lingual*.

Hovy y Lin (Hovy y Lin, 1999) presentan una taxonomía similar a la anterior, a través de la cual se clasifican también los factores que caracterizan a los resúmenes. Análogamente a los factores relacionados con la entrada, la salida y la finalidad del resumen descritos en (Spärck Jones, 1999), esta taxonomía distingue entre características del documento fuente, del resumen generado y del uso que se persiga con el mismo. Las principales diferencias entre estas dos taxonomás es que en esta última se introducen conceptos relativos a la coherencia del resumen y al grado de subjetividad del mismo. La figura 3.2 ilustra la taxonomía completa.

CARACTERÍSTICAS DEL DOCUMENTO/S FUENTE (entrada)
- tamaño de la entrada: *mono-documento vs. multi-documento*
- dominio: *dominio concreto vs. general*
- género y longitud del documento
- idioma: *monolingüe vs. multilingüe*

CARACTERÍSTICAS DEL RESUMEN (salida)
- paradigma de generación: *extractivo vs. abstractivo*
- coherencia: *coherente vs. no coherente*
- carácter: *objetivo vs. subjetivo*
- flexibilidad: *fijo vs. flexible*

CARACTERÍSTICAS DEL USO DEL RESUMEN (finalidad)
- audiencia: *genérico vs. orientado a un tema concreto*
- uso: *indicativo vs. informativo*
- contenido: *novedoso vs. detallado*

Figura 3.2: Taxonomía propuesta por Hovy y Lin

Sin embargo, la clasificación presentada por Mani y Maybury (Mani y Maybury, 1999) difiere de las dos anteriores en que ahora lo que se clasifican son

enfoques y no características del resumen. Así, se realiza una distinción entre los tipos de enfoques que se pueden adoptar a la hora de generar un resumen de forma automática, abordando el problema desde tres perspectivas: superficial, entidad y discurso. Las aproximaciones que se basan en características superficiales tratan de captar la información más importante del documento por medio de la combinación de características simples a través de alguna función para calcular la relevancia de información. Este tipo de características abarcan, entre otras, la posición de la frase, la frecuencia de los términos o la presencia de términos del título. Los enfoques basados en entidades son aquellos que construyen una representación interna del documento o documentos para modelar las entidades y las relaciones que aparecen. Las relaciones entre las entidades pueden consistir en relaciones de similitud, de proximidad, de correferencia, o relaciones lógicas de implicación o causalidad, por citar algunos tipos. Finalmente, los enfoques que se basan en la estructura del discurso pretenden modelar la estructura del documento, incluyendo características presentes en el formato del documento, modelando el hilo del discurso a través de los temas tratados en el documento, o bien captando las estructuras de los diferentes tipos de textos, como por ejemplo, textos narrativos o argumentativos. En la figura 3.3 se muestran todas estas características según la clasificación propuesta por Mani y Maybury.

ENFOQUES BASADOS EN CARACTERÍSTICAS SUPERFICIALES
- características estadísticas: *tf, tf*idf*
- posición: *posición de una unidad dentro de un documento, párrafo, sección*
- origen de la información: *presencia de términos del título, de la pregunta*
- locuciones: *"en resumen", "como conclusión", "el objetivo"*

ENFOQUES BASADOS EN ENTIDADES Y SUS RELACIONES
- relaciones de similitud
- relaciones de proximidad
- relaciones de coocurrencia
- relaciones semánticas
- correferencia entre entidades
- relaciones lógicas
- relaciones sintácticas

ENFOQUES BASADOS EN LA ESTRUCTURA DEL DISCURSO
- formato del documento: *enlaces, etiquetas, palabras en negrita*
- desarrollo del tema y subtemas
- estructura retórica

Figura 3.3: Taxonomía propuesta por Mani y Maybury

Como vemos, las tres taxonomías presentadas se abordan desde dos perspectivas distintas. Las taxonomías propuestas por Spärck Jones y la propuesta por Hovy y Lin son de granularidad fina, definiendo distintos tipos de resúmenes según diferentes criterios, mientras que la propuesta por Mani y Maybury agrupa los sistemas

en función del tipo de enfoques que utilicen para generar el resumen.

El principal problema de tener una taxonomía con tanto nivel de detalle radica en que es muy difícil poder clasificar un sistema de resúmenes en base a esos criterios, ya que la mayoría de los sistemas se enmarcarán en varios grupos, dando lugar a una clasificación poco clara y difícil. A su vez, el problema que presenta la taxonomía propuesta por Mani y Maybury es que presentan una clasificación pura, y actualmente, la mayoría de los sistemas optan por enfoques híbridos (Mani y Maybury, 1999) adoptando, por ejemplo, un enfoque basado en la estructura del discurso, pero en el que se tienen en cuenta también características superficiales, como la posición de una oración en el documento, lo que conlleva también a que la clasificación de los sistemas de resúmenes actuales sea difícil de realizar. Por lo tanto, en nuestro caso se ha optado por seguir una clasificación similar a la propuesta por Mani y Maybury, en el sentido de que los sistemas se agrupan en función del tipo de enfoque que adopten, pero que difiere de la misma en que por un lado diferenciamos entre sistemas que producen resúmenes que no se aplican a ninguna tarea concreta, de los que sí que están orientados a aplicaciones reales. Adicionalmente, dentro de esta clasificación, los sistemas se agrupan según sigan un enfoque basado en conocimiento, o basado en corpus.

A continuación, vamos a describir algunos enfoques y técnicas que se han utilizado para afrontar esta tarea, centrándonos especialmente, en resúmenes de textos. Como ya se ha mencionado, en la clasificación que proponemos se distingue entre sistemas de resúmenes genéricos y sistemas orientados a tareas. Por un lado, la finalidad de los resúmenes englobados en el primer tipo es producir un resumen a partir de un documento o conjunto de documentos de un determinado dominio sin más, mientras que por otro, los resúmenes orientados a tareas persiguen unos objetivos concretos dependiendo del uso que se quiera dar al resumen y de la aplicación en la que éste se desee integrar. Para cada uno de estos dos grupos, dependiendo de cómo se empleen las técnicas para producir resúmenes de forma automática, podemos establecer una amplia distinción entre los enfoques basados en conocimiento y los enfoques basados en corpus[1]. El primer tipo, analiza las posibles técnicas existentes y las combina, estableciendo un peso para cada una de las técnicas empleadas. Dicho peso se asigna dependiendo de la influencia que se quiera dar a cada una de las fuentes de conocimiento utilizadas. Por el contrario, los enfoques basados en corpus, se basan en el uso de algoritmos de aprendizaje automático para entrenar un conjunto de resúmenes modelo y poder extraer características de forma automática, de tal manera que dichas características permitan detectar las frases más importantes de los documentos para componer el resumen final. La ventaja derivada de los enfoques basados en corpus es que permiten rápidamente la identificación de características relevantes para seleccionar oraciones, pero su principal desventaja es que necesitamos a priori, un conjunto de pares documentos-resúmenes modelos (escritos por humanos) lo

[1]Sin embargo, esta distinción no se realizará para los resúmenes orientados a tareas, puesto que de entre todos los sistemas descritos, los enfoques basados en conocimiento son los que predominan.

suficientemente grande para poder entrenar el sistema y extraer características de forma automática, y conseguir un corpus de este tipo a veces no es tan fácil (Mani, 2001a).

3.1. Resúmenes genéricos

En este grupo se distinguirán entre enfoques basados en conocimiento y enfoques basados en corpus. Además, dentro de los enfoques basados en conocimiento, dedicaremos secciones especiales a aquellos enfoques que utilizan algoritmos basados en grafos y a aquellos que tienen en cuenta aspectos lingüísticos con la finalidad de producir resúmenes mejor organizados y más coherentes, puesto que ambos enfoques están adquiriendo una gran importancia en los últimos años.

3.1.1. Enfoques basados en conocimiento

Las investigaciones en resúmenes de textos comenzaron a finales de los años cincuenta. Los primeros investigadores en estudiar posibles técnicas para poder generar resúmenes de forma automática fueron Luhn (Luhn, 1958) y Edmundson (Edmundson, 1969), que propusieron técnicas como la frecuencia de las palabras, la posición de las frases en un documento o la identificación de *cue words*. Esta última técnica consiste en calcular la relevancia de una frase en función de si tiene locuciones como “en conclusión”, “el objetivo de este artículo”, etc. que pueden ser indicadoras de que dicha frase pueda contener información significativa. Además de estas técnicas, en la literatura se pueden encontrar diversos enfoques basados en conocimiento para generar resúmenes tanto mono- como multi-documento. Por ejemplo, en (Qiu et al., 2007) se proponen diferentes estrategias para producir resúmenes multi-documento, dependiendo de la finalidad que se persiga (por ejemplo, centrado en personas, lugares, eventos, objetos u otros), combinando varias técnicas en cada una de ellas. La idea subyacente es encontrar la combinación de técnicas más adecuada para cada tipo de resumen, y esto se consigue analizando los documentos de entrada. Las características que se proponen incluyen la frecuencia de palabras, la longitud y posición de las frases, o la importancia de las entidades nombradas. Otro sistema clásico que también se basa en la combinación de características es MEAD (Radev, Blair-Goldensohn, y Zhang, 2001). Este sistema es capaz de producir resúmenes tanto mono- como multi-documento mediante un enfoque extractivo, 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 de la oración *centroide*. Las frases que formarán parte del resumen final serán seleccionadas en base al resultado de combinar linealmente las características previamente expuestas.

La estrategia que siguen otros sistemas es la de localizar el tema principal

en el documento, y ver cómo va cambiando a lo largo de éste, o analizar cómo se va descomponiendo en subtemas, de tal manera que en función de dichos cambios, se seleccionan las frases más adecuadas para el resumen. Existe una gran variedad de métodos mediante los cuáles se puede determinar el tema principal de un documento. Además, el poder usar técnicas que segmentan un documento en función de los cambios de tema también es muy útil para la tarea de resúmenes. En la literatura, encontramos ejemplos de sistemas que utilizan estas técnicas para generar resúmenes, como es el caso de (Neto et al., 2000), (Angheluta, Busser, y Moens, 2002), (Boguraev y Neff, 2000), (Harabagiu y Lacatusu, 2005). En este último enfoque, la estructura de temas se caracteriza en términos de lo que se denominan *topic themes*, que son representaciones de eventos o estados que se repiten a lo largo de un documento. En esta aproximación se analizan cinco maneras diferentes para representar el tema de un documento: (1) *topic signatures*. Las *topic signatures* (Lin y Hovy, 2000) se basan en representar un tema en función de un conjunto de términos relacionados con ese tema. (2) Representar el tema mediante la identificación de relaciones entre las *topic signatures*. (3) Identificar el tema de un documento segmentando el texto con el algoritmo TextTiling (Hearst, 1997) para ver en qué puntos se produce una ruptura temática. (4) Representar el tema del documento modelando el contenido del mismo, utilizando modelos ocultos de Markov. (5) Finalmente, el último mecanismo está basado en el uso de plantillas (*templates*), como si de una tarea de extracción de información se tratase.

Por otra parte, en (Teng et al., 2008) se propone un sistema para producir resúmenes mono-documento, identificando en una primera etapa, los posibles temas locales del documento. Esto se realiza agrupando las oraciones en función de su similitud, y una vez hecho esto, mediante la técnica de la frecuencia de las palabras, se extraen las frases más relevantes de cada uno de los grupos que se han creado, para representar los diferentes temas del documento.

Es importante tener en cuenta que la identificación del tema principal de un documento no es la única técnica posible. Por ejemplo, el sistema propuesto en (Kuo y Chen, 2008) genera resúmenes multi-documento en base a palabras que denotan eventos o que tienen un alto contenido de información. La idea subyacente en esta aproximación es que este tipo de palabras suelen indicar conceptos y relaciones importantes de los documentos y, por tanto, pueden utilizarse para detectar las frases más relevantes dentro de un conjunto de documentos. Además, en este sistema se lleva a cabo también la resolución de las expresiones temporales, para poder transformar todas las fechas al mismo formato y no tener problemas a la hora de ordenar los eventos y las frases seleccionadas.

Hasta ahora los sistemas que hemos comentado producen resúmenes genéricos, ya sean mono-documento o multi-documento. En cambio, la diferencia de estos sistemas con respecto a los que producen resúmenes centrados en un tema concreto, es que estos últimos normalmente están guiados por unos requerimientos iniciales por parte del usuario, generalmente, en forma de pregunta, y por tanto, el resumen deberá contener la respuesta a dichas necesidades. Existen muchos sistemas que se han centrado en generar resúmenes de este tipo aunque, en esta memoria,

solamente comentaremos algunos de ellos. Para producir resúmenes centrados en un determinado tema, lo que hacen la mayoría de los sistemas es calcular, mediante alguna técnica, la similitud entre la pregunta y cada oración del documento, para que de esta manera, se seleccionen y extraigan las frases que guarden relación con dicha pregunta. Algunos de los sistemas propuestos más recientes, como es el caso de (Fuentes, Rodríguez, y Ferrés, 2007), (Toutanova et al., 2007) o (Gotti et al., 2007), producen resúmenes a partir de un conjunto de documentos, con el objetivo de proporcionar respuestas a preguntas complejas. En el primero de ellos, *FEMsum* (Fuentes, Rodríguez, y Ferrés, 2007), los resúmenes se producen teniendo en cuenta aspectos sintácticos y semánticos de las frases. Este sistema contiene diferentes módulos, entre ellos, el que detecta información relevante, el que procesa la pregunta y el que compone el resumen. El segundo sistema mencionado, *PYTHY* (Toutanova et al., 2007), se basa en generar diferentes resúmenes alternativos para la misma pregunta de tal forma que, posteriormente, de entre los posibles resúmenes, se elige el mejor de ellos. El tercer sistema, *GOFAIsum* (Gotti et al., 2007) se basa principalmente en la técnica *tf*idf* para determinar qué términos son los importantes del documento, y en base a ellos, asignar un peso a las frases. Las frases con mayor peso son filtradas según la similitud con la pregunta, y finalmente, se lleva a cabo un postprocesamiento para dar al resumen la forma final.

3.1.1.1. Enfoques basados en grafos

El uso de los algoritmos basados en grafos aplicados a la tarea de generación automática de resúmenes ya sea para producir, bien resúmenes genéricos o bien resúmenes centrados en un tema concreto, también ha sido muy utilizado recientemente. Esto se debe a que son muy efectivos y obtienen buenos resultados. Básicamente, la idea es representar un documento en forma de grafo, de manera que los nodos del mismo representen elementos del texto (normalmente, palabras o frases), mientras las aristas son las relaciones entres dichos elementos (por ejemplo, relaciones de similitud entre frases, o de sinonimia, hiperonimia, etc. entre palabras). Una vez que tenemos el documento representado en forma de grafo, la idea es que dicha representación nos ayudará a determinar los elementos más importantes del documento, por ejemplo atendiendo al grado de salida de cada nodo del grafo. En (Mihalcea, 2004) se realiza un análisis de varios algoritmos basados en grafos para generar resúmenes. Además, en (Wan, Yang, y Xiao, 2007) se propone un enfoque basado en grafos de afinidad, para producir tanto resúmenes genéricos como orientados a una necesidad específica. La idea aquí consiste en extraer frases que contengan una gran riqueza informativa, a la vez que incorporen información nueva. Esto se consigue a través de la similitud entre cada par de frases, teniendo en cuenta también información relacionada con el tema del documento y, diferenciando también entre enlaces dentro de un mismo documento y entre documentos diferentes. Finalmente, se elimina la información redundante.

Otros enfoques basados en grafos se pueden encontrar en (Plaza, Díaz, y Gervás, 2008), cuya idea es construir grafos de conceptos para producir resúmenes

de documentos de carácter biomédico; y en (Giannakopoulos, Karkaletsis, y Vouros, 2008), en la que los documentos se representan utilizando grafos de n-gramas, para determinar qué información es relevante a partir de un conjunto de documentos.

3.1.1.2. Enfoques basados en la estructura del discurso y otros aspectos lingüísticos

Además de todas las técnicas expuestas anteriormente, la tarea de generación automática de resúmenes se ha tratado de abordar también desde un punto de vista lingüístico, basándose por ejemplo, en la estructura del discurso. La estructura retórica del discurso (Mann y Thompson, 1988) sirvió como punto de partida para el enfoque propuesto en (Marcu, 1999), dónde la relevancia de una frase venía dada por las relaciones discursivas entre las oraciones de un documento, ya que se distinguían diversos tipos de relaciones (elaboración, causalidad, etc) con diferente importancia. Este enfoque fue aplicado en (Khan, Khan, y Mahmood, 2005), donde se llevó a cabo la integración de esta teoría en un sistema de resúmenes que no tenía en cuenta aspectos lingüísticos, para ver si la combinación de ambos aspectos mejoraba la calidad final de los resúmenes obtenidos.

Por otro lado, criterios como la coherencia y cohesión están cobrando cada vez más importancia en los sistemas de resúmenes. Los investigadores de esta tarea son conscientes de que la estructura y la presentación de un resumen automático es casi tan importante como su contenido, porque si el resumen no es legible y no se entiende, tendrá poca aplicación práctica. Como consecuencia, los sistemas de generación automática de resúmenes incorporan módulos para tener en cuenta ciertos aspectos lingüísticos a la hora de producir los resúmenes. Una técnica bastante extendida para, de algún modo, intentar mantener la coherencia del resumen es la técnica de las cadenas léxicas propuesta originalmente en (Barzilay y Elhadad, 1999), pero que ha sido posteriormente adoptada en muchos sistemas. Utilizando esta técnica, se pueden identificar los principales temas de un documento y la información más relevante del mismo en función de los distintos niveles de importancia de dichas cadenas (según los elementos que contengan). Algunos ejemplos de otros sistemas que se basan en esta técnica son (Medelyan, 2007) y (Ercan y Cicekli, 2008). Otra idea similar es la identificación de las cadenas de correferencia, a nivel de documento o entre documentos distintos. En (Nunes Gonçalves, Rino, y Vieira, 2008), se identifican las cadenas de correferencia en un documento para tratar de minimizar problemas de cohesión en los resúmenes. Para ello, establecen una etapa de postprocesamiento (una vez generado el resumen) en la que se reescriben las expresiones que hacen referencia a otros objetos. Pero el uso de las cadenas de correferencia, para intentar mantener la coherencia del texto, y por tanto, del resumen cuando se aplica en tareas de resúmenes de texto, no es una técnica nueva, en el sentido de que, ya en (Baldwin y Morton, 1998) y (Azzam, Humphreys, y Gaizauskas, 1999) aparecían las primeras aproximaciones para sacar provecho de las cadenas de correferencia

previamente identificadas en un texto, y poder extraer información relevante del documento sin que la coherencia del resumen se viera muy perjudicada. Son varios los sistemas que han optado por seguir una aproximación similar a las anteriores basándose también en las cadenas de correferencia (Witte, Krestel, y Bergler, 2005), (Hendrickx y Bosma, 2008), puesto que se ha demostrado que, además de producir resúmenes más coherentes, disminuye también el fenómeno conocido como *dangling anaphora*[2] tan frecuente en los resúmenes automáticos. Para minimizar este problema, lo que algunos sistemas hacen es incorporar sistemas de resolución de la anáfora en los sistemas de resúmenes automáticos, por ejemplo, en (Orasan, 2004) o (Mitkov et al., 2007). En estos trabajos, primero se realiza un proceso de resolución de la anáfora sobre el conjunto de documentos, y una vez sustituidos los pronombres por sus correspondientes antecedentes, se lleva a cabo el proceso de generación de resúmenes, analizando hasta qué punto, la resolución de la anáfora beneficia a la tarea de resúmenes. En teoría, este enfoque es una idea muy buena y sencilla de implementar, pero en la práctica, no se obtienen resultados tan buenos, debido en parte al modesto rendimiento de los sistemas de resolución de anáfora (en torno al 60 %). Sin embargo, si tuviéramos disponible un sistema ideal de resolución de la anáfora, obtendríamos beneficios considerables en la tarea de generación de resúmenes, como se demuestra en (Orasan, 2007), dónde se realiza una serie de experimentos resolviendo la anáfora de forma manual. En (Steinberger et al., 2007), se expone que los beneficios que puede obtener un sistema de resúmenes incorporando un sistema de resolución de la anáfora, no dependen tanto del rendimiento de éste último, sino de cómo se utilice la relaciones anafóricas. En el sistema de resúmenes que proponen, utilizan la anáfora con dos finalidades diferentes: por un lado, para mejorar la calidad de los resúmenes generados, y por otro, para comprobar la coherencia del mismo, una vez ya producido, identificando las cadenas de correferencia en el documento original y en el resumen y comprobando los elementos de ambas.

3.1.2. Enfoques basados en corpus

Los sistemas que a continuación se presentan, realizan el proceso de generación de resúmenes basándose en algoritmos de aprendizaje automático. Las primeras técnicas de aprendizaje automático que se usaron en esta tarea incluyeron clasificadores binarios (Kupiec, Pedersen, y Chen, 1995), Modelos Ocultos de Markov (Conroy y O'leary, 2001), (Schlesinger et al., 2002) y redes bayesianas (Aone, Okurowski, y Gorlinsky, 1998). Pero no son los únicos algoritmos que se pueden utilizar para entrenar un conjunto de datos y extraer características. En los últimos años, sistemas como *NetSum* (Svore, Vanderwende, y Burges, 2007) apuestan por el uso de redes neuronales, basándose en el algoritmo de aprendizaje *RankNet* (Burges et al., 2005). Además de utilizar características como son las

[2]Este fenómeno consiste en que los resúmenes automáticos (realizados siguiendo un enfoque extractivo) suelen contener frases con pronombres, sin que esté claro a qué objetos hacen referencia, dando lugar a referencias incorrectas o a interpretaciones erróneas.

palabras clave y la posición de la frase, se propone también, un nuevo conjunto de características basadas en la Wikipedia[3], que da preferencia a las frases que contengan términos que estén contemplados en dicho recurso.

En *FastSum*, sistema propuesto por (Schilder y Kondadadi, 2008), se seleccionan las frases usando los algoritmos *Support Vector Regression* (SVR), y *Least Angle Regression* para seleccionar y extraer características. SVR ya se utilizó previamente para generar resúmenes en (Li et al., 2007), donde se usaban características a nivel de palabra y frase, así como también la posición de las frases y la presencia de entidades nombradas, para entrenar el sistema de forma automática y puntuar cada frases del documento. En (Wong, Wu, y Li, 2008), se realiza la generación de resúmenes utilizando métodos de aprendizaje automático supervisados (utilizando *Support Vector Machines (SVM)*) y semi supervisados (combinando SVM con clasificadores *Naïve Bayes*). Las características utilizadas se agrupan en distintos tipos e incluyen, entre otras, la longitud de la frase, la frecuencia de las palabras, o la similitud entre las frases.

3.2. Resúmenes orientados a tareas

En los últimos años, los resúmenes multi-documento y orientados a usuarios están cobrando especial importancia, puesto que Internet se está convirtiendo en una fuente de conocimiento muy utilizada por todos los usuarios, dónde acuden cuando necesitan encontrar algún tipo de información específica, generalmente utilizando algún motor de búsqueda existente. Por tanto, debido a la gran cantidad de información que podemos encontrar en la web, lo que algunos sistemas de resúmenes intentan es ayudar en cierta medida, a los usuarios, facilitándoles el trabajo de tener que navegar documento por documento hasta encontrar la información concreta que buscan, y proporcionando un breve resumen informativo, que permita al lector ahorrar tiempo en leer la gran cantidad de información recuperada, teniendo sólo que echar un vistazo a los resúmenes, en vez de a los documentos completos. Tal es el caso de los sistemas *SWEet* (Steinberger, Jezek, y Sloup, 2008) y el descrito en (Yang y Liu, 2008) que tienen su punto de partida en los documentos devueltos por un motor de búsqueda, y a partir de dicho conjunto de documentos, realizan los resúmenes sobre cada uno de ellos (resumen mono-documento) o sobre clusters (resumen multi-documento), utilizando diversas técnicas, como *Latent Semantic Analysis* en el primer caso, y grafos semánticos en el segundo, para identificar así, la relevancia de cada frase. Otras aproximaciones de sistemas basadas en la web, como la versión mejorada de *NetSum* (Svore, Vanderwende, y Burges, 2008), producen resúmenes muy cortos a partir de documentos de noticias periodísticas, con el objetivo de extraer la información más destacada del texto (*highlights*). Como ya se comentó en la sección 3.1.2, *NetSum* está basado en técnicas de aprendizaje automático y por tanto, sigue un enfoque basado en corpus.

[3] http://www.wikipedia.org

En lo que respecta a resúmenes orientados a usuario, en (Díaz y Gervás, 2007) se plantea un sistema de resúmenes para el dominio periodístico capaz de contener información relevante de acuerdo a un perfil de usuario determinado. Para ello, se necesita tener primero un modelo de usuario para un individuo con el fin de, posteriormente, poder calcular la similitud entre cada una de las frases del documento y dicho perfil de usuario. Podremos encontrar diferentes resúmenes personalizados para un mismo usuario, dependiendo de la parte del perfil del usuario que se tome como referencia. A veces, no queremos que los resúmenes estén centrados en los usuarios, sino en temas concretos. El sistema *QCS* (Dunlavy et al., 2007) es un sistema complejo que integra tareas de recuperación de información, agrupación y resúmenes de textos. Centrándonos en la tarea de resúmenes, que es la que nos interesa, el sistema presentado realiza primero un resumen de cada documento recuperado, y a partir del conjunto de resúmenes generados para cada grupo de documentos relacionados, se realiza el resumen de todo el cluster.

Por otro lado, la generación de resúmenes que contengan solamente información novedosa o más actual respecto a un tema (conocidos en inglés como *update summaries*) es también un reto dentro de la tarea de resúmenes. La idea que se persigue es generar resúmenes, pero teniendo en cuenta que los lectores de los mismos tienen un conocimiento previo sobre dicho tema, y por tanto sólo desean conocer los acontecimientos más recientes y no una visión general de dicho tema. Por ejemplo, ante un tema de actualidad como puede ser el brote de gripe porcina, si conocemos información acerca del tema, posiblemente no nos interese conocer las causas, ni las vías de contagio puesto que ya lo sabremos, sino que si queremos estar actualizados, posiblemente sea más útil tener un sistema de resúmenes que nos proporcione diariamente el número de nuevos casos afectados en cada país. Un ejemplo de sistema que se basa en esta idea lo podemos encontrar en (Sweeney, Crestani, y Losada, 2008), en el que las frases que contienen novedades se determinan calculando la similitud entre las frases que ya forman parte del resumen y las posibles frases candidatas, y escogiendo las que menos similitud tengan. En cambio, en (Witte, Krestel, y Bergler, 2007) y (Bellemare, Bergler, y Witte, 2008), este tipo de resúmenes se generan a partir de grafos de clusters que se basan en el contexto del conjunto de documentos de los que queremos obtener el resumen. Es decir, se propone un método para determinar el ránking de las oraciones de los documentos en base a la cantidad de vocabulario común entre las frases pertenecientes a los clusters y el contexto, de tal forma, que finalmente se seleccionan las frases dependiendo de su posición en dicho ránking. En (Li, Wan, y Wang, 2008), se propone el concepto de “historia”, que hace referencia a aquellos documentos que el lector ya conoce. Además, se introducen métricas para filtrar frases mediante medidas de similitud, tales como el coseno, que evitan incorporar al resumen aquellas frases que guardan cierta similitud con alguna frase contenida en el histórico de documentos.

Otra aplicación de la tarea de resúmenes es generar resúmenes biográficos a partir de un conjunto de documentos referentes a una persona (Zhou, Ticrea, y Hovy, 2004). La idea subyacente es producir pequeños resúmenes que contengan

aspectos relevantes sobre una determinada persona, por ejemplo respondiendo a preguntas como *"¿Quién es Barack Obama?"*. En el sistema mencionado, para lograr esta tarea se utilizaron diversas técnicas de aprendizaje automático (redes bayesianas, SVM y árboles de decisión) para clasificar las oraciones y determinar cuáles eran las más apropiadas para formar parte del resumen, eliminando la información redundante en una etapa posterior. Otro sistema muy parecido para generar resúmenes biográficos lo encontramos en (Biadsy, Hirschberg, y Filatova, 2008), que está basado, al igual que el anterior, en algoritmos de aprendizaje automático. La diferencia entre ambos radica en que en este caso se utiliza un clasificador binario para decidir si una frase es de tipo biográfica o no, y además se utiliza la Wikipedia como corpus.

Por otro lado, aunque el dominio que más se ha utilizado para generar resúmenes ha sido el dominio periodístico (Nenkova, Siddharthan, y McKeown, 2005), (Nenkova, 2005), el científico (Jaoua y Hamadou, 2003), (Teufel y Moens, 2002) o incluso podemos encontrar algunos trabajos relacionados con el dominio legal (Saravanan, Ravindran, y Raman, 2006), (Cesarano, Mazzeo, y Picariello, 2007), algunos sistemas están intentando realizar resúmenes sobre documentos que pertenecen al dominio literario, ya se trate de relatos breves (Kazantseva, 2006) o de libros. En (Mihalcea y Ceylan, 2007) se exponen los problemas derivados al intentar producir resúmenes de libros. Se proponen, además, varias técnicas que pueden ser útiles para resumir este tipo de documentos, tomando como referencia el sistema MEAD (Radev, Blair-Goldensohn, y Zhang, 2001), con algunos cambios para adaptarlo a la generación de resúmenes de documentos con mayor extensión.

Recientemente, una de las aplicaciones de la tarea de generación de resúmenes es combinar esta tarea con la de minería de opiniones (también conocida como *Opinion Mining*) para producir resúmenes que estén orientados a opiniones, y por tanto expresen sentimientos o argumentos a favor o en contra de un determinado producto, lugar, persona, etc. En lo que respecta a este tipo de resúmenes, primero se deben detectar qué frases expresan opiniones y determinar el carácter de la opinión, es decir, si dicha opinión está a favor o en contra de algo. Una vez que las opiniones han sido clasificadas y agrupadas, habrá que construir el resumen. Algunos de los sistemas que participaron en la tarea piloto del TAC 2008[4] (*Opinion Summarization Pilot task*) seguían estos pasos (Conroy y Schlesinger, 2008), (He et al., 2008), (Balahur et al., 2008) o (Bossard, Généreux, y Poibeau, 2008). Por otro lado, fuera del ámbito de la competición también podemos encontrar enfoques interesantes para abordar esta tarea. Por ejemplo, en (Beineke et al., 2004) se propone un sistema de este estilo, en el que una vez identificados los fragmentos de texto que expresan opiniones, se utilizan técnicas de aprendizaje automático para seleccionar las frases que pertenecerán al resumen. De manera similar, en (Zhuang, Jing, y Zhu, 2006) se identifican las palabras que expresan opinión y el tipo de las mismas (bien expresando una opinión positiva o negativa), para componer posteriormente el resumen, pero orientado solamente a reseñas de películas de cine.

[4]Text Analysis Conference: www.nist.gov/tac/

Generalmente, las opiniones que se encuentran en la web vienen acompañadas muchas veces de índices numéricos que representan el grado de satisfacción de un usuario ante un determinado producto o servicio. En (Titov y McDonald, 2008) se propone un modelo (*Multi-Aspect Sentiment Model*) para identificar el carácter de las opiniones expresadas en dicho formato y poder extraer los fragmentos de textos correspondientes para generar automáticamente un resumen.

El cuadro 3.1 muestra las características más destacadas de algunos de los sistemas presentados anteriormente. En este cuadro se pretende reflejar por una parte, las características propias de los resúmenes (recogiendo los factores más importantes según la taxonomía propuesta por Hovy y Lin (Hovy y Lin, 1999)) y por otra, el enfoque que se adopta de acuerdo a la nueva clasificación propuesta en este trabajo de investigación. Los distintos colores ayudan a que sea más fácil visualizar los diferentes tipos de factores y de taxonomías. El color amarillo (enfoque) hace referencia a la taxonomía propuesta, en la que sólo vamos a distinguir si se trata de un sistema basado en conocimiento o en corpus. Por otra parte, el color gris representa los factores más importantes relacionados con la entrada del sistema, tales como si es mono-documento, monolingüe y de dominio específico. El color azul muestra el paradigma utilizado por el sistema para generar el resumen (factor relacionado con la salida), ya sea extractivo o no. Finalmente, en rosa se presentan dos factores relativos a la finalidad del resumen: si el resumen producido es genérico o no, y si es de tipo informativo. El símbolo ✓ indicará que un cierto sistema cumple un requisito concreto, mientras que una casilla con el símbolo ✗ significará que no lo cumple.

A partir del cuadro se pueden extraer a grandes rasgos las tendencias actuales respecto a la tarea de resúmenes y los problemas que todavía no están resueltos. En primer lugar, destaca el hecho de que todos los sistemas produzcan resúmenes informativos siguiendo un paradigma extractivo. La dificultad asociada a generar resúmenes de tipo abstractivo hace que casi todos los sistemas opten por un enfoque extractivo, limitándose a seleccionar las frases más relevantes de los documentos. Además, la razón por la que la mayoría de los sistemas generan resúmenes de tipo informativo se debe en parte a la definición de las tareas en determinadas conferencias, que posibilitan un conjunto de datos para realizar experimentos. También se observa que los sistemas se diseñan para un determinado dominio, puesto que cada dominio presenta unas características propias y por tanto, es difícil poder desarrollar un sistema de resúmenes completamente general para cualquier dominio. En relación al idioma con el que trabajan los resúmenes, vemos que casi todos los sistemas comentados trabajan solamente para un idioma, siendo en su mayoría el inglés y el chino. Sin embargo, en cuanto al número de documentos fuente se refiere, hay una fuerte tendencia a que los sistemas sean capaces de tratar múltiples documentos, debido a la gran cantidad de información disponible actualmente. A pesar de ello, todavía existen sistemas que se centran en la producción de resúmenes mono-documento, tarea a la que todavía le queda

SISTEMA	ENFOQUE	MONO-DOCUMENTO?	DOMINIO CONCRETO?	MONOLINGÜE?	EXTRACTOS?	GENÉRICO?	INFORMATIVO?
Qiu et al. (Qiu et al., 2007)	conocimiento	✗	✓	✓	✓	✗	✓
Teng et al.(Teng et al., 2008)	conocimiento	✓	✓	✓	✓	✓	✓
Kuo & Chen (Kuo y Chen, 2008)	conocimiento	✗	✓	✓	✓	✓	✓
FEMsum (Fuentes, Rodríguez, y Ferrés, 2007)	conocimiento	✗	✓	✓	✓	✗	✓
PYTHY (Toutanova et al., 2007)	conocimiento	✗	✓	✓	✓	✗	✓
GOFAIsum (Gotti et al., 2007)	conocimiento	✗	✓	✓	✓	✗	✓
Wan et al. (Wan, Yang, y Xiao, 2007)	conocimiento	✗	✓	✓	✓	✓[5]	✓
Plaza et al. (Plaza, Díaz, y Gervás, 2008)	conocimiento	✗	✓	✓	✓	✓	✓
Medelyan (Medelyan, 2007)	conocimiento	✓	✗	✓	✓	✓	✓
Ercan & Cicekli (Ercan y Cicekli, 2008)	conocimiento	✓	✓	✓	✓	✓	✓
Nunes et al. (Nunes Gonçalves, Rino, y Vieira, 2008)	conocimiento	✓	✓	✓	✓	✓	✓
NetSum (Svore, Vanderwende, y Burges, 2007)	corpus	✓	✓	✓	✓	✓	✓
FastSum (Schilder y Kondadadi, 2008)	corpus	✗	✓	✓	✓	✗	✓
Wong et al. (Wong, Wu, y Li, 2008)	corpus	✗	✓	✓	✓	✓	✓
SWEet (Steinberger, Jezek, y Sloup, 2008)	conocimiento	✗	✗	✗	✓	✗	✓
QCS (Dunlavy et al., 2007)	conocimiento	✗	✓	✓	✓	✗	✓
Kazantseva (Kazantseva, 2006)	conocimiento[6]	✓	✓	✓	✓	✓	✓
Titov & McDonald (Titov y McDonald, 2008)	conocimiento	✗	✓	✓	✓	✗	✓

Cuadro 3.1: Características más destacadas de los sistemas más representativos

mucho por mejorar. Centrándonos en el tipo de enfoque que siguen los sistemas, la mayor parte de ellos se basan en conocimiento, aunque también encontramos algunos ejemplos de sistemas que están basados en corpus. Para poder desarrollar un sistema basado en corpus, se necesitará disponer de un conjunto de pares documentos-resúmenes de referencia lo suficientemente grande para poder extraer las características de forma automática. En la tarea de resúmenes, por lo general resulta muy complicado y costoso conseguir un conjunto de datos de entrenamiento (resúmenes humanos) y por lo tanto, este es uno de los motivos por lo que muchos sistemas están basados en conocimiento.

Como se puede observar de todo este repaso del estado de la cuestión, existen muchos y muy diversos enfoques que se pueden adoptar a la hora de producir un resumen de forma automática, predominando los enfoques basados en conocimiento frente a los basados en corpus. Algunas técnicas clásicas, como la frecuencia de las palabras o la posición de las frases, siguen siendo muy utilizadas puesto que se trata de técnicas sencillas que obtienen buenos resultados,

[5] Este sistema, además de producir resúmenes genéricos, produce también resúmenes sobre temas concretos.

[6] Aunque también proponen un enfoque basado en corpus.

sobre todo si trabajamos con el dominio periodístico. Sin embargo, es difícil establecer qué técnicas son las técnicas perfectas para producir resúmenes de forma automática, ya que los resúmenes pueden estar orientados a diferentes objetivos, tener diversos tipos de entrada y producir distintos tipos de salidas, como se vio al inicio de este capítulo. Por tanto, a pesar del gran progreso que ha habido durante las últimas décadas en este área de investigación, la tarea de generación de resúmenes es una tarea en continua adaptación a las nuevas tecnologías que van surgiendo (teniendo aplicabilidad directa en numerosas tareas), por lo que todavía queda mucho campo por explorar y mejorar.

Capítulo 4

Estado de la cuestión en la evaluación de resúmenes

Los métodos para evaluar tareas del PLN se pueden clasificar en dos grandes grupos: métodos intrínsecos y extrínsecos (Jones y Galliers, 1996). Concretamente, aplicados a la tarea de generación de resúmenes, los métodos intrínsecos evalúan el resumen en sí, atendiendo al contenido del mismo. En cambio, los extrínsecos se centran en evaluar cómo de buenos son los resúmenes generados, para poder cumplir el cometido de otra tarea externa, como por ejemplo tareas de recuperación de información. En lo que respecta a la evaluación intrínseca, existen varios métodos que pueden ser útiles a la hora de evaluar un resumen automático. Según (Mani, 2001a), dentro de los métodos intrínsecos, se puede hacer una subclasificación entre evaluar la calidad del resumen, y realizar, así, una evaluación cualitativa, o bien evaluar el grado de información que contiene (*informativeness*), y por lo tanto realizar una evaluación cuantitativa. Además, se puede evaluar también el grado de similitud respecto al documento o los documentos origen, para determinar si el resumen cubre los mismos conceptos relevantes que el documento fuente. El problema de este método es cómo determinar la importancia de los conceptos en los documentos fuentes. Sin embargo, a pesar de existir diferentes metodologías para llevar a cabo una evaluación intrínseca, la más común es evaluar la información del resumen, comparando el contenido del mismo frente a un conjunto de resúmenes elaborados por humanos que se toman como referencia. Para lograr esta tarea, se han propuesto varios métodos y se han desarrollado varias herramientas a lo largo de estos años, que se explicarán en profundidad en la sección 4.1. No obstante, debido a la subjetividad inherente que presenta la tarea de generación de resúmenes, no se puede establecer de forma objetiva un resumen de referencia que sirva como modelo (*gold-standard*), puesto que podemos encontrar para un mismo documento, diferentes resúmenes igual de válidos. Diferentes personas pueden tener opiniones distintas acerca de lo bueno que puede ser un resumen, y además un resumen depende mucho de la finalidad que se pretenda conseguir. Estos motivos, junto con algunos otros, hacen que la tarea de evaluación

de resúmenes sea especialmente difícil, y aunque existen métodos para evaluar de forma automática un resumen, dichos métodos presentan, todavía una serie de inconvenientes y problemas que se describirán en la sección 4.1.1.

En cuanto a los métodos extrínsecos se refiere, aparte de poder evaluar los resúmenes automáticos a través de tareas externas, podemos encontrar varios escenarios planteados como juegos (Hassel, 2007). Ejemplos de algunos de estos escenarios son *The Question Game* donde, a partir de un resumen, se le propone al usuario responder una serie de preguntas, o *The Keyword Association*, cuyo objetivo es, dada una lista de palabras clave, ver hasta qué punto el resumen contiene información acerca de dichas palabras. Otro tipo de evaluación extrínseca se basa en tareas de comprensión lectora (Mani, 2001a), mediante la realización de tests de opción múltiple, por ejemplo, donde el usuario sólo tiene accesible o bien el resumen, o bien los documentos originales, para analizar el porcentaje de preguntas correctas que es capaz de responder con el resumen, y estudiar las diferencias con respecto a tener accesible todo el documento.

Puesto que la mayoría de las métricas y herramientas existentes en la actualidad se basan en métodos intrínsecos, en este capítulo se presentarán dichas herramientas clasificadas dependiendo del tipo de evaluación que realicen. Por un lado, en la sección 4.1, se presentarán las herramientas que llevan a cabo una evaluación cuantitativa (analizando la cantidad de contenido reflejado en el resumen), junto con los problemas e inconvenientes que plantean (sección 4.1.1), mientras que por otro lado, en la sección 4.2, se describirán los métodos que apuestan por una evaluación cualitativa (examinando la calidad del resumen).

4.1. Evaluación cuantitativa

Como ya se ha comentado, existen varios métodos para evaluar de forma automática un resumen generado por un sistema. Las medidas de evaluación de precisión, cobertura (*recall*), y medida-F, tan usadas en el ámbito de la recuperación de información (Van Rijsbergen, C. J., 1981), también han sido utilizadas en la tarea de evaluación de resúmenes automáticos (Nenkova, 2006). La cobertura determina qué proporción de frases seleccionadas por un humano para formar parte de un resumen, también han sido seleccionadas por un sistema, mientras que la precisión es el porcentaje de las frases identificadas por un sistema automático que son correctas. La medida-F es una combinación de ambas métricas, cobertura y precisión, donde generalmente beta toma el valor 1. La figura 4.1 muestra de forma gráfica dichas métricas.

El problema que se deriva de usar estas métricas, es que dos resúmenes totalmente válidos pueden ser evaluados de forma muy distinta, según coincidan en mayor o menor medida con las frases seleccionadas por los humanos. Para paliar este efecto negativo, se propuso la medida *Relative Utility* (Radev y Tam, 2003), en la que varios jueces decidían el grado de idoneidad de las frases de un documento para pertenecer al resumen, asignándoles una puntuación entre 0 y 10. Las frases

$$\text{Cobertura} = \frac{|\text{Frases comunes seleccionadas por el sistema y los humanos}|}{|\text{Frases seleccionadas por humanos}|}$$

$$\text{Precisión} = \frac{|\text{Frases comunes seleccionadas por el sistema y los humanos}|}{|\text{Frases seleccionadas por el sistema}|}$$

$$\text{Medida-F } (\beta=1) = \frac{2*\text{precision}*\text{cobertura}}{\text{precision}+\text{cobertura}}$$

Figura 4.1: Precisión, cobertura y medida-F

con más puntuación serían más adecuadas de cara a pertenecer al resumen final. En la misma línea que esta métrica surgió otra conocida como *factoid scores* (Teufel y Halteren, 2004), que no eran más que unidades atómicas de información que representan el significado de una frase. Una vez identificados dichos *factoids* a partir de un conjunto de resúmenes de referencia, se comprobaba cuántos de ellos estaban presentes en los resúmenes automáticos. Con una filosofía similar, apareció el método de las pirámides (*The Pyramid method*) (Nenkova, Passonneau, y McKeown, 2007), cuyo objetivo es identificar información con el mismo significado entre todos los resúmenes modelos. Estos fragmentos de información, que no tienen por qué ser oraciones enteras, se les conoce como *Summary Content Units* (SCU). Cada SCU identificada recibe un peso, dependiendo del número de humanos que coinciden en si ese fragmento debe formar parte del resumen, diferenciando así, el contenido más importante del menos. En (Fuentes et al., 2005) se propuso un método para intentar automatizar el proceso para comparar las SCU de referencia con el contenido de los resúmenes automáticos. A pesar de ello, la principal desventaja del método de las pirámides es que la identificación de las SCU se tiene que hacer de forma manual y por tanto es una tarea muy costosa.

En (Lin, 2004) se desarrolló una herramienta totalmente automática para llevar a cabo la evaluación de resúmenes. Esta herramienta llamada ROUGE[1] se ha convertido en una herramienta muy usada para la evaluación de resúmenes automáticos, y obtiene los valores de precisión, cobertura y medida-F para un resumen generado automáticamente, siempre y cuando tengamos disponible al menos, un resumen de referencia (escrito por un humano). La idea subyacente es que los textos con un significado similar, deben contener palabras o frases comunes. ROUGE se basa en la comparación de n-gramas entre ambos resúmenes, y para ello se establecen diferentes tipos de n-gramas, como unigramas (ROUGE-1), bigramas (ROUGE-2), subsecuencia común más larga (ROUGE-L), etc., siendo ROUGE-1 y ROUGE-2 las más utilizadas.

[1] *Recall-Oriented Understudy for Gisting Evaluation*: http://haydn.isi.edu/ROUGE/

Bajo la hipótesis de que la mejor métrica de similitud es la que mejor discrimina entre un resumen manual y uno automático, en (Amigó et al., 2005) se propuso un nuevo marco de evaluación: *QARLA*. Teniendo un conjunto de resúmenes de referencia y uno de resúmenes automáticos, la evaluación se realiza en base a los siguientes criterios: un criterio automático (*QUEEN*) de evaluación de resúmenes sobre un conjunto de modelos y un conjunto de métricas de similitud para estimar la calidad de un resumen automático; un criterio (*KING*) para estimar la calidad de la métrica de similitud escogida; y por último, el criterio *JACK* estima la fiabilidad del conjunto de resúmenes automáticos. Entre el conjunto de métricas de similitud propuestas (59 métricas en total), se encuentran métricas basadas en ROUGE y en el tamaño de las oraciones.

La herramienta ROUGE, explicada anteriormente, tiene una serie de inconvenientes, como por ejemplo que la comparación entre el resumen automático y el resumen de referencia se hace en base a n-gramas de tamaño fijo, pudiendo perjudicar al resumen automático si éste no contiene palabras expresadas exactamente de la misma forma que el resumen manual. Para intentar disminuir este problema se desarrolló otra herramienta, *Basic Elements* (Hovy et al., 2006), cuya idea se basa en segmentar una frase en unidades de contenido mínimo, definiendo tripletas formadas por un elemento principal (*head*), un modificador (*modifier*) y la relación (*relation*) entre ambos elementos. El objetivo era permitir mayor flexibilidad a la hora de identificar expresiones comunes entre un resumen automático y su correspondiente conjunto de resúmenes de referencia. Además, el problema de la similitud semántica está presente en todos los métodos comentados hasta ahora, puesto que las expresiones equivalentes que expresan la misma idea, pero escritas de distinta manera, no se detectarán como buenas, infravalorando la calidad del contenido del resumen. Con la finalidad de intentar solventar este problema, se propuso *ParaEval* (Zhou et al., 2006), un sistema de identificación automática de paráfrasis basado en tres niveles. En un primer nivel, se identificaban los fragmentos de texto equivalentes entre el resumen manual y el automático. Para aquellos fragmentos de los que no se había podido encontrar su expresión equivalente, se intentaba identificar los sinónimos entre palabras, y en una tercera fase, si esto también fallaba, se llevaba a cabo un emparejamiento léxico.

AutoSummENG (Giannakopoulos et al., 2008) es otra herramienta desarrollada para evaluar resúmenes de forma automática. El método usado para esta herramienta difiere de los anteriores principalmente en tres aspectos: (1) el tipo de información estadística extraído; (2) la representación escogida para dicha información; y (3) la manera de calcular la similitud entre resúmenes. La comparación entre resúmenes manuales y automáticos se realiza a partir de grafos de n-gramas de caracteres. Una vez construidos estos grafos, se lleva a cabo la comparación entre ambas representaciones para poder establecer el grado de similitud entre ambos.

Por otra parte, también podemos encontrar diversas metodologías de evaluación para idiomas distintos del inglés, como por ejemplo para el chino. Tal es el

caso de *HowNet*[2], que es una base de conocimiento para el inglés y el chino, y se diferencia de otros recursos similares, como WordNet[3] en la manera que se establecen las relaciones entre las palabras que forman el recurso. Además, *HowNet* proporciona más información para cada uno de los conceptos, asociando a cada uno de ellos una definición y relaciones con otros conceptos de forma no ambigua. Este recurso ha sido muy usado para el chino en diferentes tareas de PLN. Por ejemplo, en (Wang, Long, y Li, 2008) se propone un método para evaluar resúmenes de forma automática haciendo uso de este recurso, donde además de calcular la similitud entre pares de resúmenes en base a la co-ocurrencia de n-gramas, se tienen en cuenta también las palabras que guardan alguna relación de sinonimia entre ellas.

El cuadro 4.1 presenta algunas de las características de los métodos de evaluación comentados en esta sección. Por un lado se distingue si un método es totalmente automático o no, mientras que por otro, se proporciona información de aquellos métodos que necesitan algún tipo de anotación realizada por humanos, como es el caso de las SCU en el método de las pirámides. Como se puede observar, todos los métodos existentes necesitan resúmenes humanos que sirvan como referencia para poder realizar la comparación de los resúmenes automáticos. Esta dependencia en resúmenes humanos puede dar lugar a algunos problemas, debido al gran número de posibles resúmenes diferentes que podemos tener para un mismo documento o conjunto de documentos. En la siguiente sección se explicarán los problemas más comunes que se han identificado en la difícil tarea de evaluación de resúmenes.

Método	**Automático**	**Resúmenes humanos**	**Anotación manual**
Relative Utility	✗	✓	✓
Factoid Score	✗	✓	✓
Pyramid Method	✗	✓	✓
ROUGE	✓	✓	
QARLA	✓	✓	
Basic Elements	✓	✓	
ParaEval	✓	✓	
AutoSummENG	✓	✓	
HowNet	✓	✓	

Cuadro 4.1: Características de los métodos de evaluación existentes

4.1.1. Problemas de la evaluación cuantitativa

Aunque los métodos descritos anteriormente son de gran ayuda para la evaluación de los resúmenes automáticos, tienen ciertos problemas e inconve-

[2]http://www.keenage.com
[3]http://wordnet.princeton.edu/

nientes que merece la pena comentar. Además, también es necesario analizar los problemas y dificultades relacionados con la tarea de evaluación de resúmenes.

La primera dificultad para la evaluación de un resumen automático es poder disponer de un conjunto de resúmenes de referencia (producidos por humanos) puesto que actualmente, para determinar la corrección de un resumen automático, los métodos existentes se basan, como se ha comentado anteriormente, en la comparación directa de los resúmenes automáticos con los correspondientes resúmenes de referencia. Conseguir este conjunto de resúmenes modelos es una tarea muy difícil y muy costosa, sobre todo cuando debemos tener resúmenes multi-documento. Además, cada humano escribirá un resumen de manera distinta, teniendo diversas variantes de resúmenes para un mismo documento o conjunto de documentos. Esto significa que, aunque dos resúmenes puedan diferir en el contenido, no quiere decir por ello, que uno de los dos esté mal, sino que puede ocurrir que ambos sean correctos. Sí que es verdad, que lo que puede suceder es que, dependiendo de para qué se necesite el resumen, uno pueda ser más adecuado que otro. Para ilustrar este fenómeno, la figura 4.3 muestra tres ejemplos diferentes de resúmenes realizados sobre el mismo documento (AP880911-0016), cuyo contenido completo se puede ver en la figura 4.2. Este documento ha sido extraído del corpus de datos del DUC 2002[4] y pertenece al dominio periodístico. En la figura 4.3, los resúmenes A y B fueron generados por personas, mientras que el resumen C fue generado de forma automática, mediante un sistema basado en la frecuencia de palabras[5] para seleccionar y extraer las frases más importantes del documento. De la figura se puede observar como, si comparamos los resúmenes con el documento original, los dos primeros tienen naturaleza abstractiva, mientras que el resumen generado automáticamente es de tipo extractivo. Además, por otra parte, vemos cómo ninguno de los resúmenes es idéntico a otro, aunque la información que contienen es similar. Por ejemplo, hay palabras que aparecen en todos los resúmenes como *"tropical storm Gilbert in the eastern Caribbean"*, pero también encontramos hechos que están expresados de forma diferente, como *"Puerto Rico issued a flood watch for Puerto Rico"*, *"flooding is expected in Puerto Rico"*, *"Gilbert brought coastal flooding [...] to Puerto Rico's south coast"*. Por consiguiente, es una cuestión muy delicada decidir cuál de los tres resúmenes es el mejor, y en base a qué tomar dicha decisión. Por tanto, la comparación de un resumen automático con resúmenes humanos puede que no sea la mejor solución para la tarea de evaluación de resúmenes. Además, la dificultad aumenta si lo que tenemos que comparar son resúmenes producidos siguiendo un enfoque extractivo con resúmenes que han sido generados de manera abstractiva. Otro de los problemas derivados de contar con varios humanos que decidan qué frases de un documento deberían pertenecer al resumen final, es el poco grado de acuerdo (*agreement*) que se produce entre ellos para determinar qué frases son las más adecuadas. En (Donaway, Drummey, y Mather, 2000) y

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

[5] Este sistema ese describirá de forma más detallada en el capítulo 5.

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. ``There is no need for alarm,'' Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather'' rotating around the center of the storm. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast. There were no reports of casualties. San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night. On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 80 mph winds and sheets of rain. Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

Figura 4.2: Documento AP880911-0016

(Mani, 2001b) se demostró este fenómeno, viendo cómo variaba el valor para la cobertura, dependiendo de qué resumen humano se utilizara para la comparación.

Además, la equivalencia semántica entre diferentes palabras, por ejemplo mediante relaciones de sinonimia, o utilizando diferentes expresiones para expresar la misma idea, es otra de las dificultades de la tarea de evaluación (Nenkova, 2006), porque la mayoría de los métodos presentados sólo realizan un análisis superficial, y no tienen en cuenta el significado semántico de las oraciones. En lo que respecta a los métodos automáticos, varias son las críticas que se han realizado sobre la herramienta ROUGE. En (Sjöbergh, 2007), se describe cómo se puede generar un resumen incorrecto pero que obtenga valores altos al evaluarlo con ROUGE. Bajo esta premisa, decidieron generar resúmenes incorrectos desde el punto de vista humano, pero que al evaluarlos con ROUGE, alcanzaban un 41 % para la cobertura en la medida ROUGE-1, valor bastante bueno en el estado de la cuestión de la tarea de generación de resúmenes. Por otro lado, en (Liu y Liu, 2008) se demostró que la correlación entre ROUGE y los humanos no era tan elevada como se decía cuando se intentaban evaluar resúmenes generados a partir de conversaciones transcritas. En cuanto a la herramienta *Basic Elements*, el principal inconveniente es que usa algunos módulos dependientes del lenguaje como analizadores sintácticos, por lo que no se puede extender fácilmente para la evaluación de resúmenes en otros idiomas distintos del inglés, sobre todo si estos idiomas carecen de tales recursos.

Tropical Storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15mph with a broad area of cloudiness and heavy weather with sustained winds of 75mph gusting to 92mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast.
San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

Figura 4.3: Ejemplo de resúmenes humanos (A, B) y automático (C), respectivamente

4.2. Evaluación cualitativa

En general, un inconveniente que tienen los métodos actuales de evaluación es que sólo se centran en el contenido de los resúmenes (realizando por tanto una evaluación cuantitativa), y no tienen en cuenta otros aspectos igualmente importantes, como son la coherencia del resumen o la cantidad de información redundante que tiene. Este tipo de evaluación de la calidad de un resumen en base a aspectos más lingüísticos (evaluación cualitativa) siempre ha estado presente en la mente de los investigadores de este campo del PLN. El objetivo del protocolo *FAN*, descrito en (luc Minel, Nugier, y Piat, 1997), era evaluar la calidad de un resumen independientemente del documento origen y del contenido del mismo. Para conseguir dicho objetivo se propusieron cuatro criterios: número de referentes anafóricos incorrectos, número de segmentos de textos mal relacionados, presencia de frases repetidas, y legibilidad del resumen. Todos estos criterios eran evaluados de forma manual por dos jueces. Siguiendo la misma filosofía, se desarrolló también el protocolo *MLUCE*, cuya idea era que los posibles usuarios de los resúmenes los evaluaran para ver si les eran útiles en función de sus necesidades. De nuevo, esta evaluación se hacía de forma manual. Otros intentos (Saggion y Lapalme, 2000) se han centrado en evaluar si un resumen es adecuado o no, examinando, por una parte si los temas o conceptos más relevantes del

documento están correctamente identificados y por otra, viendo en qué grado, las frases seleccionadas en el resumen hubieran sido seleccionadas también por expertos humanos.

Más recientemente, en (Conroy y Dang, 2008) se expone la necesidad de disponer de herramientas que no sólo evalúen el contenido del resumen, sino que también tengan en cuenta aspectos lingüísticos. En las conferencias DUC[6] y TAC[7], los resúmenes se evaluaban de forma manual, generalmente en base a cinco criterios lingüísticos que representaban, de alguna manera, la calidad de un resumen. Estos criterios consistían en comprobar: la corrección gramatical del resumen (*Grammaticality*), la cantidad de información repetida o redundante existente en el resumen (*Non-redundancy*), si el resumen se centra en un tema o trata muchos temas inconexos (*Focus*), si las referencias anáforicas, incluyendo los pronombres, están claramente identificadas (*Referential Clarity*), y por último, evaluar la estructura y coherencia del resumen (*Structure and Coherence*). Estos criterios, evaluados manualmente, no requieren comparación ni con resúmenes generados por humanos ni con los documentos fuentes. Simplemente, los jueces determinan si el resumen cumple los objetivos para cada criterio en base a unas pautas que reciben, dándoles una puntuación entre 1 y 5, dependiendo del grado en que el humano piensa que el resumen es correcto. En (Pitler y Nenkova, 2008), podemos encontrar un estudio para predecir la calidad de un texto, analizando diversas métricas relacionadas con la legibilidad de un texto (*Readability*). Otros enfoques, como las propuestas en (Barzilay y Lapata, 2005), (Lapata y Barzilay, 2005) o (Elsner y Charniak, 2008), se centran en cómo modelar la coherencia de texto, con el objetivo de poder medir y evaluar la coherencia de un documento, de manera que estos enfoques podrían ser aplicados para evaluar si un resumen es o no coherente.

[6] http://www-nlpir.nist.gov/projects/duc/
[7] www.nist.gov/tac/

Capítulo 5

Sistema de generación de resúmenes basado en conocimiento

Una vez conocido el estado de la cuestión en la tarea de generación de resúmenes, y después de ver las posibles técnicas y aplicaciones que se pueden emplear para detectar la información relevante en un documento, en este capítulo, describiremos una propuesta para la generación de resúmenes automáticos, describiendo en detalle las técnicas utilizadas y las características relativas al mismo. Uno de los objetivos de esta investigación a largo plazo es construir un sistema de resúmenes robusto, y por tanto, en esta memoria se presentarán los primeros pasos para la consecución de dicho objetivo.

Las técnicas que se explicarán en las secciones siguientes, y que son las técnicas principales que sustentan el sistema de resúmenes, son: la frecuencia de las palabras (sección 5.1), la técnica de reconocimiento de implicación textual (sección 5.2), y el principio de la cantidad de codificación (sección 5.3) descrito en (Givón, 1990). La hipótesis de trabajo consiste en analizar cada una las técnicas propuestas de forma individual y ver si influyen de forma positiva en la tarea de resúmenes, para posteriormente combinar aquellas técnicas que tengan una influencia positiva y analizar si dicha combinación es también apropiada para la generación de resúmenes de forma automática. Para ello, en la sección 5.5 se presentará la evaluación realizada describiendo, por una parte, los experimentos planteados y por otra, los resultados de los mismos. Los resultados obtenidos nos permitirán llevar a cabo un análisis de errores y extraer conclusiones (sección 5.5.1) acerca del sistema propuesto, pudiendo identificar fortalezas y debilidades del mismo, con la finalidad de mejorar el sistema en un futuro.

5.1. Frecuencia de palabras

A pesar de que esta técnica fue de las primeras en utilizarse para generar resúmenes de forma automática (Luhn, 1958), todavía se sigue utilizando (véase el capítulo 3), ya que es una técnica sencilla de aplicar que obtiene muy buenos resultados. Además, en (Nenkova, Vanderwende, y McKeown, 2006) se analizó el impacto que la frecuencia de palabras tiene en los resúmenes humanos y se comprobó como éstos están formados por palabras que, a su vez, presentan una alta frecuencia en los documentos originales. A partir de estas premisas, decidimos incorporar la frecuencia de las palabras (WF) como la primera característica de nuestra propuesta para el sistema de resúmenes. La hipótesis de trabajo es que las palabras con mayor frecuencia en un documento indican cuáles son los conceptos más importantes del mismo, y por tanto, las frases que contengan dichos conceptos contendrán información relevante, por lo que tendrán mayor probabilidad de pertenecer al resumen final. Para cada oración del documento se calculará su relevancia a través de la fórmula 5.1, de tal manera que, una vez procesado todo el documento, cada frase tendrá un valor de relevancia asignado. La fórmula calculará dicho valor de relevancia para cada oración, sumando la frecuencia de las palabras que contiene (*tfi*) y el resultado se dividirá entre la longitud de la frase (*n*). Tanto para el cálculo de la frecuencia como para el de la longitud de la frase, no se tienen en cuenta las *stopwords*.

$$Sc_s = \frac{\sum_{i=1}^{n} tf_i}{n} \ . \tag{5.1}$$

Si consideramos, por ejemplo, las dos frases mostradas en la Figura 5.1, extraídas del corpus del DUC del 2002[1], la puntuación para la frase *a* es 3.2 y para la *b* es 1. Entre paréntesis se indica la frecuencia de cada palabra respecto a todo el documento, y se puede observar como las *stopwords* no se han tenido en cuenta para el cálculo de la frecuencia ni para contar la longitud de la oración.

S$_a$: *Tropical*(2) *Storm*(6) *Gilbert*(7) *formed*(1) *in*(0) *the*(0) *eastern*(1) *Caribbean*(1) *and*(0) *strengthened*(1) *into*(0) *a*(0) *hurricane*(7) *Saturday*(4) *night*(2).
S$_b$: *There*(0) *were*(0) *no*(0) *reports*(1) *of*(0) *casualties*(1).

Figura 5.1: Frases de ejemplo del documento *AP880911-0016* (corpus DUC 2002)

Una vez realizado este proceso, se ordenarán las frases en función de dicho valor, y se extraerán aquellas frases que tengan valores altos (el número de frases a extraer dependerá de las restricciones que tengamos en cuanto a la longitud que

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

debe tener el resumen), componiendo el resumen final. Siguiendo con el ejemplo de la figura anterior, si tuviéramos que escoger entre una de las dos frases mostradas para formar el resumen, elegiríamos la primera oración ya que la puntuación que obtenemos (3.2) es mayor que la obtenida para la segunda (1). Finalmente, para intentar que el resumen sea lo más coherente posible, una vez que sabemos las frases que formarán el resumen, éstas se ordenarán manteniendo el mismo orden en el que aparecían en el documento original.

Por tanto, como ya se ha explicado, la frecuencia de las palabras será la primera característica de nuestro sistema de resúmenes, donde las frases más importantes serán las que contengan las palabras más frecuentes.

5.2. Implicación textual

Otra técnica que decidimos analizar y estudiar si era adecuada para la generación de resúmenes fue el reconocimiento de la implicación textual[2]. El reconocimiento de la implicación textual es, en sí, otra tarea del PLN, cuyo objetivo es determinar si el significado de un fragmento de texto (hipótesis) se puede deducir de otro fragmento de texto (texto) (Glickman, 2006). Ambos fragmentos de texto deben ser oraciones coherentes escritas en lenguaje natural. Los siguientes pares de oraciones (figura 5.2), extraídos del corpus proporcionado en *Third RTE Challenge*[3](Giampiccolo et al., 2007), ilustran ejemplos en los que se puede ver cuándo se produce implicación textual y cuando no.

Par id=109 implicación=SI
T: ASCAP is a membership association of more than 200,000 US composers, songwriters, lyricists and music publishers of every kind of music.
H: More than 200,000 US composers, songwriters, lyricists and music publishers are members of ASCAP.
Par id=194 implicación=NO
T: US President George W. Bush has indicated he will invite Abbas to the United States for talks, something he never did with Abbas's predecessor, the late Yasser Arafat.
H: Yasser Arafat succeeded Abbas.

Figura 5.2: Ejemplos de implicación textual

En la literatura podemos encontrar algunos trabajos en los que el reconocimien-

[2]En inglés, esta tarea es conocida con el nombre de *Textual Entailment*.

[3]Esta edición pertenece al marco de conferencias orientadas al reconocimiento de la implicación textual (RTE, del inglés *Recognising Textual Entailment*).

to de la implicación textual se ha combinado con la tarea de resúmenes. Concretamente, el reconocimiento de la implicación textual se ha utilizado para:

- Evaluar resúmenes (Harabagiu, Hickl, y Lacatusu, 2007).
  Suponiendo que el resumen es la hipótesis y el documento original es el texto, se utiliza la tarea de reconocimiento de implicación textual para decidir qué resumen (de un conjunto de posibles resúmenes candidatos) mejor se deduce de su correspondiente documento.

- Segmentar el documento (Tatar et al., 2008).
  Se utiliza el reconocimiento de la implicación textual para segmentar el documento original, es decir, formar grupos a partir de frases implicadas entre sí, y una vez establecidos dichos grupos, poder extraer una o varias frases de cada uno de ellos para formar el resumen final.

- Eliminar información redundante. Utilizando la implicación textual se puede eliminar aquellas oraciones que contienen información repetida de un documento, y evitar, así, que un resumen contenga redundancia.

Este último enfoque, descrito originalmente en (Lloret et al., 2008a) y (Lloret et al., 2008b), será el que adoptaremos para nuestra propuesta del sistema de resúmenes. Para ello, utilizaremos el sistema de reconocimiento de implicación textual descrito en (Ferrández et al., 2007) con algunas mejoras y entrenado con el corpus proporcionado en la tercera edición del RTE (Giampiccolo et al., 2007). Este sistema se basa en el cómputo de un conjunto de medidas léxicas (como por ejemplo, distancia de *Leveshtein*, *SmithWaterman*, similitud del coseno) y semánticas basadas en *WordNet 3.0*[4], aplicando un clasificador SVM con el objetivo de tomar la decisión final.

La idea principal que sostiene el uso de la implicación textual en tareas automáticas de resúmenes siguiendo el enfoque propuesto, reside en conseguir un resumen preliminar formado por las oraciones 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.

Como ejemplo, supongamos que un documento está formado por el siguiente conjunto de frases:

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

y el documento reducido por el cómputo de las relaciones de implicación textual se obtiene de la siguiente manera:

$RESUMEN = \{S_1\}$
$RESUMEN \longrightarrow implica \longrightarrow S_2 \Rightarrow NO$

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

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

Por lo tanto, el documento reducido obtenido por el procesamiento de las inferencias de implicación textual comprenderá aquellas frases que no son implicadas por el conjunto de oraciones que no han producido relación de implicación previamente ($S1, S2, S3, S6$ en el ejemplo anterior).

Debido al tiempo computacional que necesita el sistema de implicación textual para procesar todas las posibles relaciones entre pares de frases, decidimos considerar solamente el cálculo de las relaciones en un sentido, comenzando por las primeras frases hasta procesar todo el documento.

5.3. Principio de la cantidad de codificación

El principio de la cantidad de codificación (CQP[5]) (Givón, 1990) es un principio de origen lingüístico que establece que mientras menos predecible (o 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 otras palabras, si la información es más importante, entonces recibirá una cantidad de codificación mayor, con lo que se codificará con mayor peso léxico. 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 cumple en textos escritos y además, está directamente relacionado con otro principio de carácter cognoscitivo, 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.

Por todas estas razones, optamos por analizar la influencia que este principio tiene en la tarea de resúmenes. 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.

[5]En inglés se le conoce como *The Code Quantity Principle*.

Por ejemplo, si tenemos estos dos sintagmas nominales *"the Academy of Motion Pictures Arts and Sciences"* and *"the Academy"*, el primero de ellos sería más completo que el segundo y además, evita posibles ambigüedades. Por tanto, si tuviéramos que generar un resumen a partir de un documento que contuviera estas dos frases, y previamente no hubiéramos incluido ninguna información acerca de qué tipo de academia se trata, nos convendría incluir la frase que tuviera el sintagma nominal más completo (en nuestro ejemplo, el primero de ellos), ya que, de esta forma, en el resumen quedaría claro qué se trata de la academia de las artes y las ciencias. Otro motivo para considerar los sintagmas nominales lo encontramos en el análisis realizado en (Mittal et al., 1999), donde se demostró que, en términos medios, la longitud de los sintagmas nominales de las oraciones que forman parte de un resumen es más del doble que las oraciones que no son seleccionadas para formar parte del resumen final.

Nuestra hipótesis de trabajo al utilizar este principio es que las oraciones que contengan sintagmas nominales más largos harán que las frases contengan más información y de manera menos ambigua, y por consiguiente, dichas frases recibirán mayor peso para ser seleccionadas y formar parte del resumen final.

Para identificar los sintagmas nominales de una frase, utilizamos la herramienta *BaseNP Chunker* disponible en *ftp://ftp.cis.upenn.edu/pub/chunker/*. Esta herramienta, como cualquier otra herramienta de PLN, tiene errores, pero su rendimiento es bastante bueno, alcanzando valores de precisión y cobertura del 93 % para sintagmas nominales simples, y 88 % para los sintagmas nominales más complejos (Ramshaw y Marcus, 1995). Una vez que ya tenemos los sintagmas nominales identificados en las frases de los documentos, las frases serán clasificadas en función de un ránking obtenido mediante la fórmula 5.2.

$$Sc_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |w| \ . \tag{5.2}$$

Tal y como la fórmula indica, para cada frase, sumaremos todas las palabras que pertenezcan a un sintagma nominal (cada palabra que se incluya en un sintagma nominal se cuenta como una unidad, a excepción de las *stopwords*) y dividiremos entre el número total de sintagmas nominales que contenga dicha frase. La figura 5.3 muestra un ejemplo de cómo se aplicaría esta fórmula sobre un par de frases extraídas del documento AP880217-0100 del corpus del DUC 2002. Siguiendo la misma idea que se explicó cuando se utilizó la frecuencia de las palabras, finalmente se seleccionaría la frase (o frases) con mayor puntuación para formar parte del resumen.

5.4. Combinación de las técnicas

Después de presentar cada una de las diferentes técnicas por separado, lo que nos interesa de cara a construir un sistema de resúmenes automático, es analizar si dichas técnicas se pueden combinar para mejorar la calidad del sistema. Lo

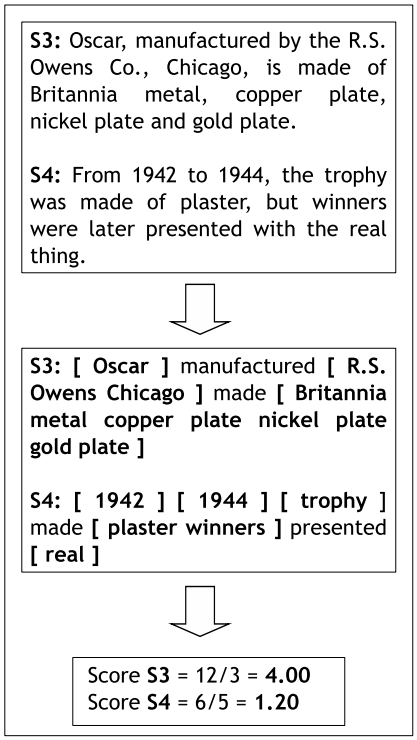

Figura 5.3: Cálculo de la relevancia aplicando el CQP.

que haremos en esta sección será describir la arquitectura del sistema propuesto, formado principalmente por tres módulos. Cada uno de ellos se corresponde con una de las técnicas descritas en las secciones anteriores. Posteriormente, en la sección 5.5 se presentará el entorno de evaluación, junto con los resultados obtenidos.

La figura 5.4 muestra un esquema general de la arquitectura del sistema. Se trata de una primera propuesta preliminar para generar automáticamente resúmenes, y nos permite añadir, combinar o eliminar módulos según queramos probar y analizar los resultados que diferentes técnicas pueden aportar a esta tarea del PLN.

Según el esquema presentado en la figura 5.4, el proceso de generación de resúmenes se puede dividir en tres etapas, siendo la segunda de ellas, la parte principal de todo el proceso, ya que en esta fase se identifican cuáles son las oraciones más relevantes de un documento. Las fases implicadas en el proceso de producción de resúmenes son: (1) una etapa de preprocesamiento del documento original; (2) la etapa de detección de relevancia, cuyo objetivo es determinar la información más importante dentro de un documento; y (3) una de postprocesamiento, cuya misión es agrupar y presentar las frases seleccionadas, para finalmente obtener el resumen. En la fase de preproceso, el documento se transforma a texto plano (si no lo está ya), eliminando además, todas las etiquetas

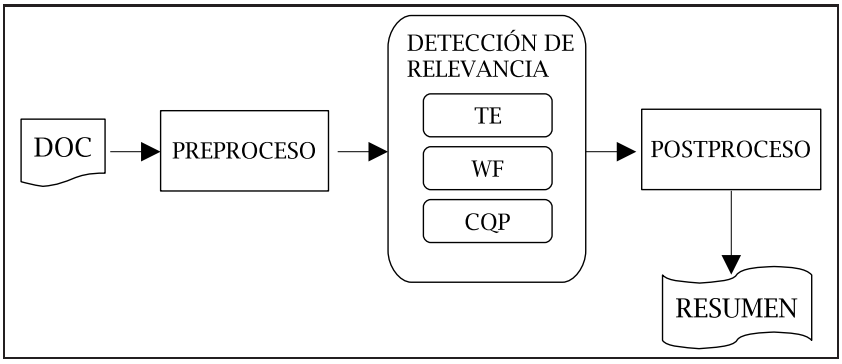

Figura 5.4: Arquitectura del sistema

innecesarias o que puedan introducir algún tipo de ruido en el resumen final. Para el texto restante, se calculará la frecuencia de cada palabra, sin considerar las *stopwords*, que nos servirá posteriormente para determinar qué frases son las más importantes. Una vez que la fase de preprocesamiento ha concluido, pasamos a la etapa de detección de relevancia, en la que encontramos diferentes aproximaciones según las características (descritas en las secciones anteriores) que nos interesen utilizar. Así, podemos elegir entre la frecuencia de las palabras (WF), la implicación textual (TE), el principio de la cantidad de codificación (CQP), o bien la combinación de ellas. Un aspecto a destacar sobre el módulo de reconocimiento de la implicación textual es que podría haberse incluido dentro de la etapa de preprocesamiento, ya que cuando se combina con otras técnicas siempre se ejecuta en primer lugar. Sin embargo, esta técnica ayuda a descartar frases que contienen información redundante y, por tanto, decidimos incluirla como una característica más. Lo que nos interesa especialmente de la arquitectura y las técnicas propuestas es analizar y comprobar cómo podemos integrar gradualmente distintas características y estudiar su influencia en la generación automática de resúmenes, para determinar qué combinación de técnicas proporciona los mejores resultados. Una vez que hemos seleccionado qué frases deben formar parte del resumen final dependiendo de la longitud que deba tener el resumen, pasamos a la etapa de postprocesamiento, en la que las frases seleccionadas se muestran en el mismo orden en el que aparecían en el documento original para mantener, en la medida de lo posible, la coherencia del resumen. Actualmente, la fase de postprocesamiento es muy simple, y es necesario tener en cuenta más aspectos para conseguir un resumen coherente, y no sólo el orden de las frases. De momento, la investigación se centra, básicamente, en analizar técnicas que ayuden a detectar información relevante dentro de un documento, dejando la etapa de postprocesamiento como un asunto pendiente para tratarlo como parte de la investigación futura.

5.5. Evaluación de la propuesta

En esta sección se describen los experimentos que se han realizado para probar el sistema, así como la herramienta de evaluación utilizada. Además, se realiza un análisis de los resultados obtenidos, para identificar los puntos débiles de la propuesta y poder mejorarla en el futuro.

Para realizar la evaluación del sistema, decidimos seguir las pautas descritas en la tarea mono-documento del DUC 2002. Esta tarea consistía en producir automáticamente resúmenes que tuvieran una extensión de 100 palabras, a partir de documentos en inglés, pertenecientes al dominio periodístico. Por tanto, decidimos también usar los documentos facilitados por la organización del congreso en esa edición, que eran en total 567 documentos. El motivo por el que se ha decidido trabajar con los datos del DUC 2002, es debido a que ese año tuvo lugar la última edición del DUC en el que se propuso una tarea de generación de resúmenes mono-documento. En las ediciones posteriores (hasta la última, en el 2007) todas las tareas estaban orientadas a resúmenes multi-documento, por lo que la mejor opción fue trabajar con los datos del 2002, ya que disponíamos, por un lado, de documentos con sus correspondientes resúmenes manuales y, por otro, de los resúmenes generados automáticamente por los sistemas participantes, para poder realizar una comparación y ver el rendimiento de nuestra propuesta.

Finalmente, debido a que algunos de los 567 documentos iniciales estaban repetidos, trabajamos con 533 documentos. Al aplicar el sistema de resúmenes, se ejecutaban las etapas previamente explicadas de manera secuencial: primero se realizaba un preprocesamiento de los documentos, después se identificaba la información relevante de cada documento, y por último se componía el resumen final con las frases seleccionadas. Los experimentos realizados se pueden agrupar en tres tipos. Primero, decidimos generar resúmenes utilizando solamente una de las técnicas propuestas, bien la frecuencia de las palabras o bien, el principio de la cantidad de codificación. La razón por la que el reconocimiento de la implicación textual no se probó de forma individual, fue porque el algoritmo para eliminar la redundancia se definió de tal forma que se daba preferencia a las frases que ocupaban las primeras posiciones del documento, dando lugar a resúmenes muy similares al *baseline* proporcionado en el DUC 2002[6], que utilizaremos para comparar los resultados. En segundo lugar, decidimos combinar las técnicas de dos en dos, obteniendo tres nuevas aproximaciones: TE+WF, TE+CQP, WF+CQP. Para poder combinar la frecuencia de las palabras con el principio de la cantidad de codificación, lo que hicimos fue asignar como peso de cada palabra perteneciente a un sintagma nominal, su frecuencia en el documento en vez de simplemente contar el número de palabras contenidas en cada sintagma nominal. Finalmente, combinamos todas las técnicas, aplicando el reconocimiento de la implicación textual para eliminar redundancia, en primer lugar, y calculando la importancia de

[6]El *baseline* propuesto en DUC consistía en formar el resumen con las 100 primeras palabras del documento.

cada una de las frases restantes en base al principio de la cantidad de codificación combinado con la frecuencia de las palabras. Para cada una de las aproximaciones, se produjeron resúmenes automáticamente para los 533 documentos. La figura 5.5 muestra un ejemplo de un resumen automático producido a partir de la combinación de TE+WF+CQP, para el documento AP880325-0239.

What do Charlie Chaplin, Greta Garbo, Cary Grant, Alfred Hitchcock and Steven Spielberg have in common?
They have never won Academy Awards for their individual achievements.
Oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the Academy of Motion Picture Arts and Sciences.
The honorary award has also proved useful to salve the Academy's conscience.
Douglas Fairbanks, Judy Garland, Noel Coward, Ernst Lubitsch, Fred Astaire, Gene Kelly, Harold Lloyd, Greta Garbo, Maurice Chevalier, Stan Laurel, Cary Grant, Lillian Gish, Edward G. Robinson, Groucho Marx, Howard Hawks and Jean Renoir are others who have received honorary awards.

Figura 5.5: Ejemplo de resumen automático generado por nuestra propuesta

En lo que respecta a la evaluación, decidimos utilizar, entre todas las herramientas de evaluación existentes (véase la sección 4.1 del capítulo 3), la herramienta ROUGE (Lin, 2004) básicamente porque esta herramienta tiene una alta correlación con los juicios humanos, y porque ha sido una herramienta muy utilizada para evaluar resúmenes de forma automática, al ser rápida y simple de utilizar. Además, al disponer de los resúmenes manuales en el corpus de datos del DUC 2002, no era necesario realizar ningún tratamiento adicional sobre dichos resúmenes, sino que era suficiente con compararlos con los resúmenes generados automáticamente. Las medidas ROUGE que utilizamos en nuestra evaluación fueron: ROUGE-1, ROUGE-2, ROUGE-L y ROUGE-SU4. ROUGE-1 y ROUGE-2 calculan la precisión y cobertura en función de los n-gramas de tamaño 1 y 2, respectivamente, comunes entre los resúmenes automáticos y los de referencia. ROUGE-L se basa en la subsecuencia común más larga entre ambos tipos de resúmenes, mientras que ROUGE-SU4 tiene en cuenta los bigramas comunes, pero sin necesidad de que sean consecutivos, permitiendo hasta un máximo de cuatro palabras entre ellos. El cuadro 5.1 muestra los resultados obtenidos para cada una de las aproximaciones comentadas para la medida-F, ya que ésta tiene también en cuenta la precisión y la cobertura.

Aproximación	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
WF	0,42951	0,16970	0,19618	0,38951
CQP	0,41751	0,17127	0,19261	0,38039
TE+WF	0,44491	0,18726	0,21056	0,40360
TE+CQP	0,42853	0,17885	0,19892	0,38722
WF+CQP	0,44153	0,18565	0,20795	0,39920
TE+WF+CQP	**0,45611**	**0,20252**	**0,22200**	**0,41382**

Cuadro 5.1: Evaluación del sistema de resúmenes propuesto

Por otra parte, el cuadro 5.2 muestra los resultados obtenidos por nuestro sistema al combinar todas las posibles técnicas, es decir, al utilizar el reconocimiento de la implicación textual (TE), la frecuencia de las palabras (WF), y el principio de la cantidad de codificación (CQP) para generar resúmenes de forma automática, en comparación con el *baseline* del DUC 2002, y con los tres sistemas que mejores resultados obtuvieron en dicha edición. En este caso, los valores mostrados hacen referencia solamente a la cobertura. Esto se debe a que la evaluación para los sistemas y el *baseline* del DUC 2002 ha sido extraída de (Steinberger et al., 2007), donde se utilizaba una versión de ROUGE anterior (versión 1.4.2 frente a la versión 1.5.5 que utilizamos nosotros) que sólo permitía el cálculo para el valor de la cobertura.

Sistema	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
TE+WF+CQP	**0,46008**	**0,20431**	**0,22399**	**0,41744**
S28	0,42776	0,21769	0,17315	0,38645
S21	0,41488	0,21038	0,16546	0,37543
DUC baseline	0,41132	0,21075	0,16604	0,37535
S19	0,40823	0,20878	0,16377	0,37351

Cuadro 5.2: Comparación de resultados con los sistemas del DUC 2002

5.5.1. Análisis de los resultados

De los resultados obtenidos se puede concluir que la combinación de técnicas propuesta es apropiada para la tarea de generación de resúmenes ya que influye de manera positiva. Cuando combinamos las técnicas frecuencia de palabras, reconocimiento de la implicación textual y el principio de la cantidad de codificación se obtiene, en términos medios, una mejora del 7 % y del 12 % respecto a combinar dos de las técnicas o utilizar sólo una, respectivamente (véase el cuadro 5.1). Al comparar nuestra propuesta con los sistemas participantes en el DUC 2002, nuestros resultados mejoran un 10 % los resultados obtenidos por el mejor sistema (S28), y se obtiene una mejora del 14 % sobre el *baseline* (véase el cuadro 5.2).

Si tenemos en cuenta la combinación de dos características, observamos que las técnicas que mejores resultados obtienen son la frecuencia de las palabras combinado con el reconocimiento de la implicación textual. Por otro lado, echando un vistazo general a los resultados obtenidos para cada técnica de manera individual, el principio de la cantidad de codificación es la técnica que peores resultados obtiene. Esto puede deberse a que al tratarse de noticias de tipo periodístico, la información se suele presentar de forma concisa sin que las oraciones contenga sintagmas nominales muy largos. Para comprobar este hecho, decidimos analizar la longitud media de los sintagmas nominales en el corpus de noticias que utilizamos, obteniendo una longitud media para los sintagmas

nominales de 1,5. Por tanto, parece que esto puede ser un motivo por el que el principio de la cantidad de codificación no obtenga tan buenos resultados. En cambio, la frecuencia de las palabras, a pesar de ser una técnica estadística muy simple, obtiene muy buenos resultados, ya sea utilizada de forma individual como en combinación con otras.

En cuanto a resultados cualitativos se refiere, no ha sido posible evaluar de manera objetiva y automática los resúmenes generados para ver si son legibles o si cumplen ciertos criterios lingüísticos que son importantes en un texto, como son la coherencia, cohesión, referencias anafóricas, etc. En total se generaron 3198 resúmenes automáticos (533 para cada aproximación propuesta, teniendo seis aproximaciones distintas), por lo que era muy difícil evaluar de forma manual cada uno de ellos. Por tanto, sólo fue posible evaluar los resúmenes en cuanto a la información contenida en ellos, mediante la comparación con resúmenes de referencia generados por humanos, y no a cómo dicha información está estructurada en el texto.

5.6. Investigación en progreso

Como ya hemos comentado anteriormente, el sistema propuesto genera resúmenes mono-documento (extractos, concretamente) para el inglés a partir de documentos pertenecientes al dominio periodístico. Sin embargo, hemos realizado también algunos experimentos preliminares para extender este sistema a multi-documento, así como también se ha intentado aplicar el sistema propuesto a otro tipo de dominio distinto del periodístico, para ver en qué medida las técnicas propuestas son técnicas fácilmente portables para generar otros tipos de resúmenes.

5.6.1. Resúmenes multi-documento

Partiendo de la hipótesis de que aplicar implicación textual como parte del proceso de generación automática de resúmenes es ventajoso, ya que se elimina información redundante del documento origen, y que, además, la redundancia de información es uno de los problemas más importantes cuando pasamos al ámbito de resúmenes multi-documento (Radev, Hovy, y McKeown, 2002), se decidió realizar una aproximación muy simple y básica para extender el sistema de resúmenes a multi-documento, y aplicar el reconocimiento de la implicación textual para estudiar el comportamiento de la implicación textual para este tipo de resúmenes. Como una primera aproximación, se optó por considerar todos los documentos de la entrada como uno sólo. Aunque esta no es la mejor forma para poder generar resúmenes multi-documento, la finalidad que se perseguía con este experimento era verificar si el reconocimiento de la implicación textual realmente influía positivamente en el proceso de resúmenes, eliminando información redundante, ya que en un conjunto de documentos similares encontraríamos mucha más información repetida que en un único documento. Para este proceso, por

motivos de tiempo, decidimos experimentar con la frecuencia de palabras y combinarla posteriormente con la implicación textual, puesto que los resultados obtenidos para mono-documento evaluados con la herramienta ROUGE fueron resultados muy prometedores (véase el cuadro 5.1).

Para probar el rendimiento del sistema multi-documento, decidimos basarnos en la tarea 2 del DUC 2002, que estaba orientada a la generación de resúmenes multi-documento, y cuyo objetivo era producir automáticamente resúmenes de diferentes longitudes a partir de conjuntos de documentos relacionados. A partir de esta definición, decidimos ejecutar los siguientes experimentos:

- Generación de extractos de 100 palabras de longitud, para cada uno de los grupos del corpus del DUC 2002 considerando solamente la frecuencia de las palabras en el proceso de resumen (multi-documento, WF).

- Generación de extractos de 100 palabras de longitud, para cada uno de los grupos del corpus del DUC 2002 aplicando el reconocimiento de la implicación textual como paso previo a la frecuencia de las palabras en el proceso de resumen (multi-documento, TE+WF).

El cuadro 5.3 muestra los resultados obtenidos para las pruebas realizadas utilizando la herramienta ROUGE, sólo para el valor de la cobertura. En dicho cuadro se ha incluido los resultados para el *baseline* del DUC 2002 para la tarea multi-documento. Dicho *baseline* cogía las 100 primeras palabras del documento más reciente. Los resultados para dicho baseline evaluado con ROUGE se han tomado de (Wan, Yang, y Xiao, 2006), en el que sólo se dispone del valor para la cobertura, de ahí que no hayamos incluido los resultados ni para la precisión ni para la medida-F. Los valores ROUGE seleccionados para esta experimentación han sido ROUGE-1 (unigramas), ROUGE-2 (bigramas) y ROUGE-W. ROUGE-W se basa en la misma idea que ROUGE-L, pero con la salvedad de que memoriza los tamaños de los emparejamientos consecutivos para quedarse con el mayor de ellos.

Aproximación	ROUGE-1	ROUGE-2	ROUGE-W
TE+WF	**0,31333**	**0,05780**	**0,09588**
WF	0,29620	0,05200	0,09266
Lead baseline	0,28684	0,05283	0,09525

Cuadro 5.3: Resultados para la aproximación multi-documento

Analizando los resultados obtenidos, la incorporación de la implicación textual mejora los resultados aproximadamente un 7 % de media respecto a no incorporar el módulo de reconocimiento de implicación textual, es decir respecto al uso de la técnica de frecuencias de palabras solamente. Experimentalmente se ha comprobado que se han eliminado aproximadamente el 72 % de las frases de los

documentos originales del corpus del DUC 2002, por lo que la reducción de los textos al aplicar implicación textual es bastante considerable y significativa. El módulo de implicación textual, a su vez, podría estar eliminando información relevante, pero se observa que, aun trabajando con el 30 % del contenido del documento y generando un resumen a partir de ese fragmento, los resultados que se obtienen son muy prometedores de cara a futuras investigaciones.

5.6.2. Extensión a otros dominios

Para probar si las técnicas propuestas (frecuencia de palabras, implicación textual y el principio de la cantidad de codificación) se pueden portar a otros dominios, además del periodístico, decidimos realizar unos experimentos adicionales sobre un conjunto reducido de documentos de tipo literario narrativo, tratándose concretamente de cuentos infantiles. La razón por la que decidimos experimentar con este dominio, fue porque este tipo de textos no ha sido muy estudiado en la tarea de resúmenes de documentos (sólo en (Kazantseva, 2006) y (Mihalcea y Ceylan, 2007) se proponen enfoques para resumir textos literarios), ya que casi todo el trabajo realizado en tareas de generación de resúmenes se centra en noticias de prensa (Nenkova, Siddharthan, y McKeown, 2005), o en documentos científicos (Morales, Esteban, y Gervás, 2008). Además, los textos de carácter narrativo son muy interesantes porque tienen una estructura muy característica por lo que hay que ser cuidadosos con la manera en la que se genera el resumen, puesto que hay que tener en cuenta qué eventos se desarrollan durante la historia y cómo, qué personajes aparecen y qué papel desempeñan, etc. para poder tomar una decisión de cuáles habrá que extraer como más importantes. Todo esto son aspectos muy importantes a tener en cuenta para generar el resumen, pero lo que se pretende al cambiar de dominio es comprobar, mediante unos experimentos preliminares, la portabilidad de las técnicas propuestas, dejando para más adelante el tratamiento de aspectos concretos para este dominio.

Para la realización de estas pruebas, decidimos experimentar con la frecuencia de las palabras, así como con el principio de la cantidad de codificación combinado con dicha técnica. En este caso, decidimos no utilizar el módulo de implicación textual porque generalmente los cuentos infantiles no suelen contener redundancia de información, ya que narran secuencias de eventos o acciones, y normalmente dichos eventos no suelen estar repetidos. Por esta razón, decidimos no utilizar la implicación textual, porque en estos casos, sí que se podría eliminar información importante del documento, perjudicando los resúmenes generados de forma automática. Por tanto, como ya hemos dicho sólo consideramos la frecuencia de las palabras como característica individual (para ver si en este dominio también proporcionaba buenos resultados), y el principio de la cantidad de codificación combinado con la frecuencia de palabras, ya que dicha combinación considerará más relevantes aquellas frases que contengan sintagmas nominales más largos y que a su vez, las palabras que los forman tengan una alta frecuencia en el documento.

La evaluación se realizó de la misma manera que para los documentos del dominio periodístico, utilizando ROUGE como herramienta de evaluación. Desgraciadamente, es muy difícil conseguir gran cantidad de datos de este dominio con sus correspondientes resúmenes, para llevar a cabo una evaluación más completa, por lo que de momento sólo pudimos trabajar con cinco cuentos infantiles[7], y sus correspondientes resúmenes manuales[8]. Los cuentos seleccionados estaban en el idioma inglés y fueron los siguientes: *Cinderella, Little Red Riding Hood, Jack and the Beanstalk, Rumpelstilskin*, y *Hansel and Gretel*.

Aproximación	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
WF	0,39709	0,07939	0,14093	0,31498
CQP+WF	**0,41797**	**0,10267**	**0,15898**	**0,33742**

Cuadro 5.4: Resultados para los cuentos infantiles

El cuadro 5.4 muestra los resultados obtenidos para el valor de la medida-F. Se puede ver cómo al combinar el principio de la cantidad de codificación con la frecuencia de las palabras, los resultados mejoran un 14 % aproximadamente, mientras que para el dominio periodístico la mejora WF+CQP respecto a WF era del 5 % aproximadamente. Esto nos puede indicar varias cosas. En primer lugar, puede que para este nuevo dominio, la frecuencia de palabras no funcione tan bien como para el dominio anterior, mientras que si la combinamos con el principio de la cantidad de codificación ocurre al contrario. Al contener los cuentos infantiles oraciones con personajes, descripciones de los personajes, eventos que se suceden durante el transcurso de la historia, etc. es más probable que una frase contenga sintagmas nominales más largos, y por tanto, los resultados indican que este principio funciona mejor para este tipo de dominio, y por tanto el incremento de mejora es más alto.

[7] Los cuentos han sido extraídos de: http://www.gutenberg.org/

[8] Los resúmenes de los cuentos se pueden encontrar en: http://www.comedyimprov.com/music/schmoll/tales.html. Nos gustaría agradecer a David Schmoll por dejarnos utilizar sus resúmenes para la realización de estos experimentos.

Capítulo 6

Generación de resúmenes de opiniones

Como ya se introdujo en la sección 3.2 del capítulo 3, una de las aplicaciones de la tarea de generación de resúmenes es la de crear resúmenes a partir de un conjunto de opiniones que versan todas ellas sobre un tema concreto. Por ejemplo, imaginemos que un usuario desea comprar una cámara digital, pero antes de tomar una decisión sobre qué cámara adquirir, le gustaría leer algunas opiniones de otros usuarios que posean cámaras con características similares a la deseada. Sería de gran utilidad, por tanto, disponer de un sistema que, ante una consulta de usuario sobre un determinado producto, devolviera de forma resumida y agrupada, las opiniones que otros usuarios han dejado acerca de dicho producto en distintos blogs, foros, páginas web, etc. De esta forma, ayudaríamos al usuario a tomar decisiones, sin que éste tuviera que mirar por su cuenta cada una de las opiniones que otros usuarios han dado a través de Internet.

6.1. Participación en TAC 2008

Con la finalidad de conseguir un objetivo similar al ejemplo planteado anteriormente, la tarea *Opinion Summarization Pilot Task*, propuesta en el congreso TAC (Text Analysis Conference)[1] del 2008 consistió en producir resúmenes de opiniones. Concretamente, en dicha tarea se proporcionaba un conjunto de blogs (pertenecientes a la colección *Blog06*[2]) y un conjunto de preguntas, de manera que el objetivo consistía en, dada una o varias preguntas sobre un tema, generar un resumen extrayendo la información necesaria de los blogs, que contuviera la respuesta a esas preguntas. Las preguntas eran preguntas de opinión y para simplificar el problema, la propia organización del congreso proporcionó una serie de posibles respuestas, llamadas *snippets*, que los sistemas participantes podían utilizar si lo consideraban oportuno. Estas posibles respuestas eran fragmentos de

[1] www.nist.gov/tac/

[2] http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

textos que sistemas de búsqueda de respuestas reales habían recuperado ante las mismas preguntas, en otra tarea del mismo congreso. En la figura 6.1 se muestra un ejemplo de tema, preguntas, y *snippets*.

Tema:
George Clooney
Preguntas:
Why do people like George Clooney?
Why do people dislike George Clooney?
Snippets:
1050 BLOG06-20060205-018-0000647869 Yes george Clooney is a spoiled punk he judges his own people while having a silver spoon in his mouth .
1050 BLOG06-20060209-006-0013539097 he's a great actor

Figura 6.1: Ejemplo de tema, preguntas y snippets

Nuestra participación en la tarea piloto del TAC 2008 (Balahur et al., 2008) consistió en proponer dos aproximaciones para construir un sistema de resúmenes de opiniones desde dos perspectivas diferentes: una, utilizando los *snippets* proporcionados (aproximación basada en *snippets*) y la otra, buscando directamente las respuestas en los blogs (aproximación basada en blogs), respectivamente. La única diferencia entre las dos aproximaciones es que en la aproximación basada en blogs, como primer paso, se calcula la similitud entre la pregunta formulada y las frases del blog para encontrar las posibles respuestas, mientras que en la aproximación basada en *snippets*, las posibles respuestas ya vienen dadas, y sólo hay que procesar dichos *snippets*. En la figura 6.2 se muestra la arquitectura propuesta para el sistema.

En esta arquitectura se pueden distinguir claramente tres etapas: procesamiento de la pregunta, procesamiento de los *snippets* relacionados con la pregunta (o en el caso de que optemos por no utilizar los *snippets*, el procesamiento de las frases recuperadas directamente de los blogs), y finalmente la generación del resumen final. A continuación, vamos a describir cada una de estas fases en mayor detalle, explicando también las herramientas y recursos que hemos utilizado en cada una de ellas.

- **Procesamiento de la pregunta**

  En esta fase, después de estudiar el tipo de preguntas del corpus, se extraen las palabras clave de la pregunta y la orientación de la misma, es decir, si se trata de una pregunta de carácter positivo o negativo. Para ello, consideramos como palabras clave todas aquellas palabras de la pregunta, excepto las *stopwords*, mientras que consideraremos como foco de la

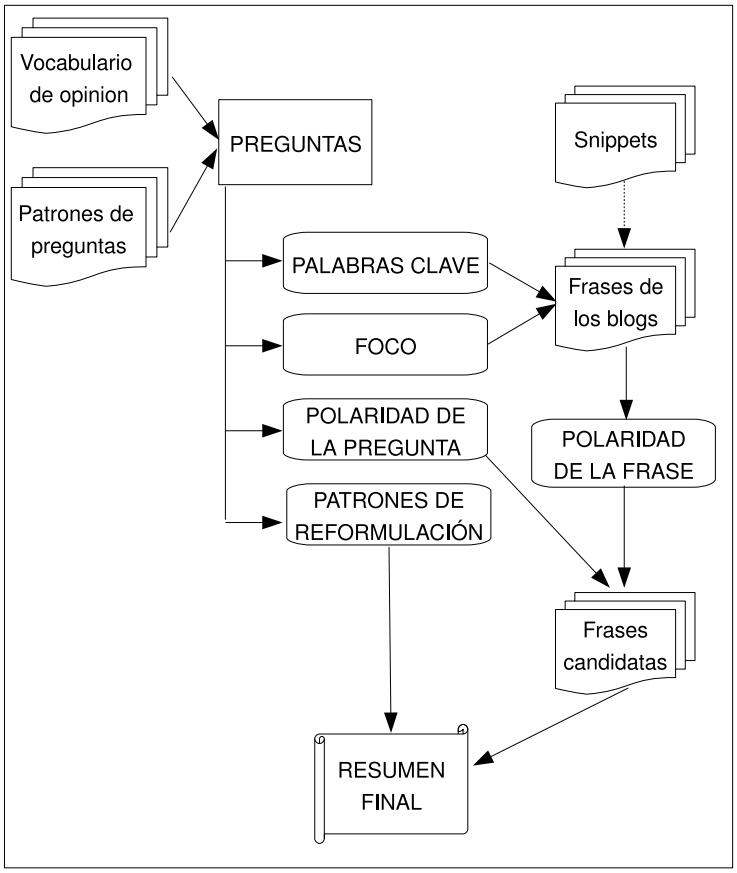

Figura 6.2: Arquitectura del sistema de resúmenes de opiniones

pregunta, aquellas entidades nombradas que aparezcan[3], y en caso de no contener ninguna, asumiremos que el foco es idéntico a las palabras clave. Por otro lado, para determinar la polaridad de la pregunta, primero se definen una serie de patrones para cada pregunta para extraer aquellas palabras de las preguntas (nombres, verbos, adjetivos, etc.) que indiquen opinión, para posteriormente clasificar dichas palabras usando vocabulario específico de emociones, a través de diferentes recursos específicos, entre ellos *Wordnet Affect* (Strapparava y Valitutti, 2004) y *emotion triggers* (Balahur y Montoyo, 2008). Además, se generan también, mediante un conjunto de reglas, unos patrones de reformulación para cada pregunta, que nos servirán para dotar al resumen final de mayor coherencia y adoptar un enfoque abstractivo, en vez de simplemente extraer las frases más relevantes y mostrarlas tal cual aparecen en los blogs originales. Finalmente, como resultado de esta fase, obtenemos por un lado, los patrones de reformulación que acabamos de mencionar y por otro, el foco, las palabras clave y la polaridad de la pregunta.

[3]Utilizamos Freeling (http://garraf.epsevg.upc.es/freeling/) para detectar entidades nombradas.

Por ejemplo, en la figura 6.3 se puede ver lo que se obtendría para la pregunta *"Why do people like George Clooney?"*.

Palabras clave:
people like George Clooney
Foco:
George_Clooney
Polaridad:
positive
Patrones de reformulación:
People like George Clooney because
One reason people like George Clooney is
Another reason people like George Clooney is
It is said that people like George Clooney because
A further motivation people like George Clooney is

Figura 6.3: Resultados obtenidos en la fase "Procesamiento de la pregunta"

- **Procesamiento de los *snippets* o frases recuperadas de los blogs**

  Esta segunda fase del sistema varía según si optamos por la aproximación basada en *snippets* o en blogs. En el caso de que utilicemos directamente los *snippets* proporcionados por la organización del congreso, lo que haremos será buscar y extraer la frase completa del blog donde aparece el *snippet* (si no aparece literalmente, utilizaremos medidas de similitud para extraer las frases más parecidas) y determinar su foco y su polaridad. En la aproximación basada en blogs, para extraer las posibles respuestas a las preguntas formuladas, utilizamos la herramienta Text Similarity (Pedersen y Patwardhan, 2004) para calcular la similitud entre la pregunta y cada una de las frases del blog, de tal manera que finalmente tendremos un conjunto de frases que potencialmente puedan contener la respuesta a la pregunta. Previamente a este paso, realizaremos el siguiente preprocesamiento sobre la colección de blogs: transformaremos los blogs a texto plano, eliminando etiquetas innecesarias que contengan información irrelevante, y segmentaremos el contenido de los mismos en frases.

  Tanto en una aproximación como en otra, habrá que adoptar un mecanismo para eliminar la redundancia, puesto que muchos de los *snippets* proporcionados contienen información repetida o son simplemente idénticos. También puede ocurrir que las frases extraídas directamente de los blogs expresen la misma información. En este punto, la única consideración para tratar este problema, es detectar las frases totalmente idénticas y quedarnos solamente con una de ellas. Una vez que tengamos el conjunto de frases candidatas, a través de una u otra aproximación, extraeremos el foco y la polaridad de cada una, de la misma manera que se hacía para las preguntas. Además,

deberemos asociar cada una de estas frases con su correspondiente pregunta para cada tema. Recordemos que para un mismo tema, podemos tener varias preguntas, siendo una de carácter positiva y otra negativa, por ejemplo. El resumen, por tanto, tiene que contener información acerca de dicho tema, pero respondiendo a cada una de las preguntas. La estrategia empleada para emparejar cada frase con su correspondiente pregunta se puede resumir en las siguientes reglas.

1. Si el tema por el que se pregunta sólo contiene una pregunta, habrá que seleccionar del conjunto de posibles frases candidatas, aquéllas cuya polaridad sea igual a la de la pregunta.
2. Si el tema contiene dos preguntas con el mismo foco, pero una es de carácter positivo y la otra de carácter negativo, para hacer la correspondencia entre la pregunta y las posibles frases, las frases de carácter positivo se asociarán con la pregunta positiva, y las de carácter negativo con la pregunta negativa.
3. Si el tema contiene varias preguntas con diferente foco y además diferente polaridad, habrá que clasificar las posibles frases respuesta atendiendo al foco y a la polaridad de las mismas.
4. Si el tema consta de dos preguntas con el mismo foco y polaridad, la correspondencia se realiza teniendo en cuenta el orden de aparición de las entidades en la pregunta y en la posible respuesta, y verificando, al mismo tiempo, que la polaridad de ambas tiene el mismo carácter.

Una vez que tenemos cada frase asociada con su pregunta, sólo nos queda la última fase, que es la de generar el resumen final, cuyo proceso se explica de forma detallada a continuación.

- **Generación del resumen**

  Esta es la última etapa para el sistema de resúmenes de opinión propuesto. Una vez que sabemos qué frases van con cada pregunta ya podemos construir el resumen final. Pero antes de agrupar las frases, filtramos las frases que aparezcan incompletas, para evitar así que fragmentos de texto sin sentido formen parte del resumen. Para realizar este filtrado, realizamos un análisis sintáctico[4] sobre cada una de las frases y descartamos aquéllas que no estén completas.

  Por otra parte, puesto que en la competición limitaban el tamaño de los resúmenes generados a 7000 caracteres por pregunta formulada, no podíamos incluir en el resumen todas aquellas frases candidatas. El criterio que escogimos para determinar qué frases pertenecerían al resumen final fue el de ordenar las frases según su polaridad de mayor a menor para

[4]Utilizamos la herramienta *Minipar* para realizar el análisis sintáctico: http://www.cs.ualberta.ca/ lindek/minipar.htm

seleccionar así, aquéllas que tuvieran una polaridad, tanto positiva como negativa, bien determinada. Además, al principio de cada frase se introducía uno de los patrones de reformulación generados en la primera fase, para darle al resumen un carácter abstractivo (y no mostrar sólo las frases seleccionadas), y hacer que el resumen fuera más coherente y legible, dándole un carácter general y evitando juicios personales. Relacionado también con la coherencia y la legibilidad, decidimos darle al texto un estilo neutro e impersonal, para que en vez de aparecer frases como *"I think"*, aparecieran frases del estilo *"people think"* o *"they think"*. Esto se realizó generando una serie de reglas para identificar los pronombres personales y sustituyéndolos por el pronombre *they*. Además, se utilizó el etiquetador morfológico *TreeTagger*[5] para identificar qué verbos estaban en tercera persona y transformarlos en estilo impersonal, eliminando la *s* del final de los verbos en tercera persona, ya que el idioma con el que trabajamos es el inglés.

La figura 6.4 muestra un ejemplo con un fragmento[6] de un resumen generado para cada una de las aproximaciones (basada en *snippets*, en la parte superior y basada en blogs, en la parte inferior) para el tema *"George Clooney"* que incluye las preguntas *" Why do people like George Clooney?"* y *"Why do people dislike George Clooney?"*. Como se puede observar en la figura, aunque ambos resúmenes contienen información relacionada con el tema y cada una de las preguntas, la información contenida en cada uno de ellos es bastante diferente. Es importante destacar que los resúmenes no son perfectos y arrastran errores derivados del uso de las herramientas y recursos que se han ido mencionado anteriormente, pero aun así son capaces de obtener buenos resultados en ciertos aspectos, y, aunque en otros aspectos, los resultados no son tan buenos, veremos en las siguientes secciones, cómo se han intentado solventar algunos problemas que presenta el sistema propuesto.

6.1.1. Resultados

La participación en la tarea piloto *Opinion Summarization* del TAC 2008 incluía la evaluación manual de los resúmenes generados por los participantes. Un total de 19 grupos de distintas universidades participaron en esta tarea y aunque, cada grupo podía enviar hasta un máximo de tres aproximaciones distintas, finalmente sólo las dos primeras enviadas pudieron ser evaluadas, dando lugar a 36 evaluaciones en total.

[5] http://www.ims.uni-tuttgart.de/projekte/corplex/TreeTagger/

[6] Ante la imposibilidad de poner el resumen completo, debido a que las limitaciones de longitud impuestas por la organización del TAC 2008 (7000 caracteres por pregunta) daban lugar a resúmenes bastante grandes, se ha optado por poner a modo de ejemplo un fragmento de 100 palabras del resumen original.

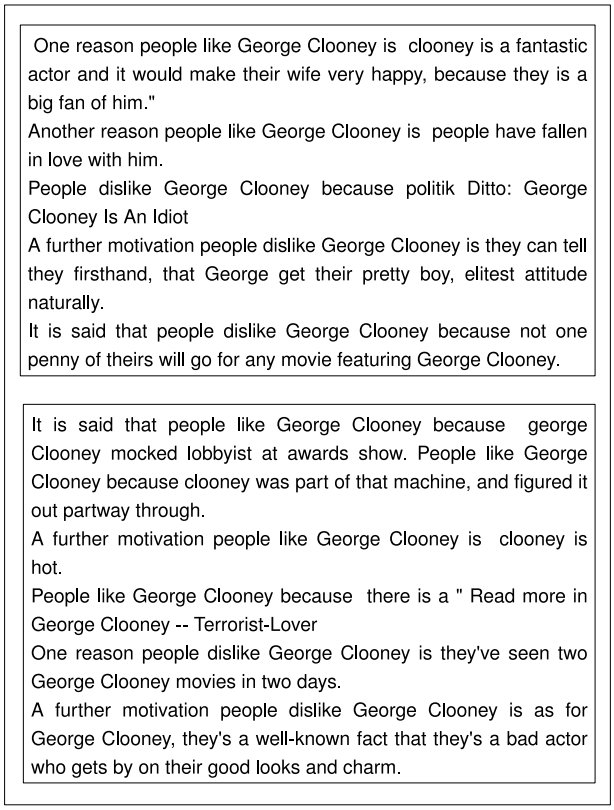

Figura 6.4: Ejemplos de resúmenes generados

Para realizar la evaluación, la organización del congreso decidió utilizar el método *Pyramid* propuesto por (Nenkova, Passonneau, y McKeown, 2007) (introducido en la sección 4 del capítulo 3), que brevemente, lo que propone es establecer un listado de unidades de texto para el resumen, también llamadas *SCUs* o *nuggets*, que indican el grado de importancia que tiene un fragmento de información de cara a pertenecer al resumen, dependiendo del número de humanos que coincidan en que dicho fragmento de texto (no tiene por qué ser frases completas) debe pertenecer al resumen final. Por tanto, cada *nugget* o fragmento de texto lleva asociada una puntuación que indica su importancia. Para ello, se proporcionaron desde la organización del congreso un conjunto de *nuggets* junto a su puntuación, y se calculó cuántos de ellos estaban presentes en los resúmenes generados por los participantes, obteniendo los valores para la precisión, cobertura y medida-F (con beta igual a 1). Además, se evaluaron también los siguientes criterios: *Grammaticality*, que consiste en determinar la corrección gramatical del resumen; *Non-redundancy*, para ver si el resumen tiene información repetida; *Structure/Coherence*, relacionado con la estructura y coherencia del resumen; *Fluency/Readability*, para ver si el resumen es fácil de leer y se entiende; y *Responsiveness*, que mide el grado en el que el resumen contiene información que

responde a las preguntas formuladas para un tema concreto.

El cuadro 6.1 muestra los resultados obtenidos para cada una de las aproximaciones propuestas (basada en *snippets* (A1), y en blogs (A2), respectivamente), así como también la posición de cada aproximación para cada criterio respecto al resto de los grupos participantes (indicada entre paréntesis).

	Medida F	*Grammat.*	*Non-redund.*	*Structure*	*Fluency*	*Respons.*
A1	0,357 (7)	4,727 (8)	5,364 (28)	3,409 (4)	3,636 (16)	5,045 (5)
A2	0,155 (23)	3,545 (36)	4,364 (36)	3,091 (13)	2,636 (36)	2,227 (28)

Cuadro 6.1: Resultados de la participación en TAC 2008

A la vista de los resultados, se puede concluir que la aproximación basada en *snippets* obtiene mejores resultados que la basada en blogs, puesto que es muy difícil obtener directamente de un blog cuáles son las frases más indicativas para que aparezcan en un resumen y que, además, coincidan con los juicios humanos a la hora de evaluarlas. Centrándonos en cada uno de los criterios, merece la pena destacar cómo en lo que respecta a la medida-F y al criterio que evalúa la coherencia (*Structure/Coherence*) ambas aproximaciones obtuvieron resultados muy buenos, por lo que nos hace pensar que la idea de generar patrones de reformulación para los resúmenes es adecuada y ayuda a que el resumen tenga mayor coherencia. Mientras que para la primera aproximación los resultados en lo que a la legibilidad (*Fluency/Readability*) y el contenido (*Responsivenes*) se refiere, fueron también muy buenos, dichos resultados no fueron tan buenos, pero sí aceptables, para la segunda aproximación. Sin embargo, los resultados obtenidos para los criterios relacionados con la redundancia (*Non-redundancy*) y corrección gramatical (*Grammaticality*) fueron bastante bajos para ambas aproximaciones. Por una parte, en cuanto al criterio *Non-redundancy*, los bajos resultados se debieron a que la estrategia utilizada para eliminar información redundante era muy simple (sólo descartábamos frases totalmente idénticas), y por tanto no contemplábamos los casos en los que la información era igual pero expresada de distinta forma. Por otra parte, respecto al criterio *Grammaticality*, al intentar generar los resúmenes en un estilo impersonal, se produjeron errores de concordancia entre sujeto y verbo, que no pudimos controlar en un primer momento, pero que se hubiesen podido corregir utilizando alguna herramienta automática de corrección de textos en una fase posterior.

6.2. Mejora del sistema después del TAC 2008

Posterior al desarrollo del congreso TAC 2008, en vista de que algunos resultados obtenidos se podían mejorar, decidimos realizar varias mejoras sobre el sistema de resúmenes de opiniones propuesto inicialmente. Nuestro objetivo era analizar las posibles causas de los resultados obtenidos, en base al tipo de

evaluación llevada a cabo en la competición, y desarrollar mejoras para los criterios que habían obtenido peores resultados (*Non-redundancy* y *Grammaticality*) (Lloret et al., 2009). Por otro lado, en el seno de la competición, para generar el resumen final nos basábamos principalmente en la agrupación de frases, ordenándolas de acuerdo a la polaridad de cada una de ellas, dando prioridad a aquéllas con un fuerte grado de polaridad, ya fuera positiva o negativa, para establecer los límites del resumen, y empleando mecanismos de postprocesamiento para formar un resumen final coherente y legible. Relacionado con esto, nos planteamos estudiar qué pasaría si utilizásemos un sistema de generación de resúmenes genérico, sin tener en cuenta las frases de opinión ni el carácter de las mismas, para, de esta manera, ver la influencia de cada parte por separado, del sistema de opiniones y de resúmenes.

6.2.1. Problemas en la evaluación

Ya se comentó anteriormente el método que se había seguido en el TAC 2008 para evaluar los resúmenes. Se optó por la evaluación utilizando el método *Pyramid* basándose en un conjunto de *SCUs* (o *nuggets*) definidas manualmente, en función de cuántos expertos humanos pensaban que dicho fragmento de texto debía formar parte del resumen. Examinando en profundidad cada una de las *SCUs* propuestas como fragmentos de texto adecuados para el resumen, nos dimos cuenta de que habían algunos problemas que valía la pena comentar y se podían agrupar en los tipos que, a continuación, se detallan:

1. Algunos *nuggets* con una alta puntuación no aparecían en la lista de *snippets* proporcionada por la organización. Por ejemplo, el siguiente *nugget "When buying from CARMAX, got a better than blue book trade-in on old car"* (0.9) no se podía asociar a ningún *snippet* de los proporcionados, aunque sí que se encontraba en uno de los blogs. Por tanto, confiando solamente en el contenido de los *snippets*, podríamos estar perdiendo información relevante.

2. Algunos *nuggets* para un mismo tema, aunque no eran idénticos, expresaban la misma idea. Un ejemplo lo encontramos en estas dos frases *"NAFTA needs to be renegotiated to protect Canadian sovereignty"* y *"Green Party: Renegotiate NAFTA to protect Canadian Sovereignty"*.

3. Había casos en los que el significado de uno de los *nuggets* se podía deducir de otro: *"reasonably healthy food"* y *"sandwiches are healthy"*.

4. Para algunos *nuggets*, su significado no estaba del todo claro, pudiendo favorecer palabras incorporadas en el resumen por casualidad sin que tuvieran nada que ver. Tal es el caso de los *nuggets*, *"hot"* o *"fun"*, por ejemplo.

5. Un *snippet* podía estar formado por dos *nuggets*, por lo que no quedaba claro si, en dichos casos, se contarían los *nuggets* por separado o, al pertenecer

al mismo *snippet*, éste se contaría sólo una vez, perdiendo la puntuación asociada a uno de ellos. Por ejemplo, los *nuggets "it is an honest book"* y *"it is a great book"* corresponden al mismo snippet *"It was such a great book- honest and hard to read (content not language difficulty)"*.

Estos factores podrían haber influido a la hora de evaluar los resúmenes, puesto que la puntuación asociada a cada *nugget* es la que sirvió para determinar los valores de precisión, cobertura y medida-F. A pesar de que los resúmenes generados por nuestras aproximaciones obtuvieron un valor medio de medida-F bastante bueno, examinando la precisión y la cobertura de manera individual para cada resumen, observamos que para algunos casos, se obtenían valores muy bajos, debidos en parte a los factores previamente comentados. Esto demuestra que el análisis posterior, una vez conocidos los *nuggets* que se consideraban relevantes para componer la información del resumen, fue un buen ejercicio para saber cuáles podían haber sido las causas de los resúmenes con puntuaciones más bajas de lo normal y que a priori, era difícil conocer y entender el porqué de los resultados.

6.2.2. Mejora de los resultados del TAC

De nuestra participación en TAC 2008, los resultados obtenidos para los criterios *Non-redundancy* y *Grammaticality*, no fueron del todo satisfactorios (véase tabla 6.1), por lo que a posteriori, decidimos mejorar el sistema en estos aspectos, para mejorar la calidad de los resúmenes generados y de paso, conseguir un equilibrio de resultados entre todos los criterios.

De las dos aproximaciones propuestas inicialmente, decidimos mejorar solamente la aproximación basada en los *snippets*, puesto que era la que mejores resultados había obtenido. En primer lugar, para evitar la presencia de información redundante en el resumen, decidimos utilizar un módulo de implicación textual similar al propuesto en (Iftene y Balahur-Dobrescu, 2007). El módulo de implicación textual (TE) se utilizó para calcular todas las posibles relaciones de implicación entre pares de frases, una vez que teníamos el conjunto potencial de frases candidatas para incorporar al resumen, en vez de filtrar las frases según el grado de polaridad. De esta manera podíamos eliminar frases, que aunque no contuvieran las mismas palabras, expresasen los mismos conceptos. Después de introducir esta mejora, observamos que los resultados referidos a la medida- F mejoraban. El cuadro 6.2 muestra una comparativa donde se puede apreciar la mejora de los resultados. Puesto que no disponemos información acerca de qué criterios concretos se establecieron en la competición para evaluar la presencia de información redundante en los resúmenes y por tanto, no podemos realizar una comparativa con respecto a los anteriores resultados obtenidos para este criterio, sí comprobamos manualmente que los resúmenes generados en esta ocasión, apenas contenían información redundante. La principal desventaja de incorporar un módulo de reconocimiento de implicación textual es la gran cantidad de tiempo que necesita para poder calcular las relaciones de implicación en ambos sentidos, y con todos

los posibles pares de frases. Para disminuir el tiempo de proceso de este módulo, podíamos haber optado por usar la implicación textual de la misma manera que se usó en (Lloret et al., 2008a) y en (Lloret et al., 2008b), y que fue explicado en la sección 5.2 del capítulo 5.

Sistema	Medida F
Mejor sistema TAC 08	0,534
Segundo mejor sistema TAC 08	0,490
Aprox. *snippets*+TE	0,530
Aprox. *snippets*	0,357

Cuadro 6.2: Mejora del sistema incorporando TE (medida-F)

Respecto al otro criterio, *Grammaticality*, lo que decidimos hacer a posteriori, fue utilizar un módulo externo, concretamente usamos la herramienta *Language Tool*[7], que automáticamente corrige los errores gramaticales que aparecen en un texto y que permite configurar los tipos de errores que queremos detectar. En nuestro caso, los errores a detectar fueron básicamente concordancias entre sujetos y verbos, y entre nombres y determinantes, así como también el uso incorrecto de las mayúsculas, palabras repetidas o mal escritas, o errores de puntuación. Una vez generados los resúmenes, antes de ser presentados como finales, realizamos sobre ellos un proceso de comprobación de este tipo de errores utilizando esta herramienta, de tal manera que se corregían los errores detectados y se mejoraba la calidad del resumen final presentado. Los resultados del análisis del criterio *Grammaticality* usando la citada herramienta sobre los resúmenes finales para la aproximación basada en *snippets* se puede ver en el cuadro 6.3.

Tipo de error	Correctos	Incorrectos
Concordancia Suj-Pred	90 %	10 %
Concordancia Det-Nombre	75 %	25 %
Mayúsculas	80 %	20 %
Palabras repetidas	100 %	0 %
“.” repetidos	80 %	20 %
Errores ortográficos	60 %	40 %
“”/() mal emparejados	100 %	0 %

Cuadro 6.3: Análisis de la corrección gramatical (*Grammaticality*)

A la vista de las mejoras obtenidas, en el futuro sería muy conveniente incorporar dichas mejoras en nuestro sistema de resúmenes de opiniones.

[7]Esta herramienta se encuentra disponible en http://community.languagetool.org/

6.2.3. Sistema de resúmenes genérico

Paralelamente a la mejora de nuestra propuesta del sistema de resúmenes de opiniones, decidimos realizar unas pruebas adicionales, para estudiar y analizar qué hubiera pasado si, en vez de construir un sistema específico para generar resúmenes de opiniones, hubiéramos utilizado el sistema propuesto en el capítulo 5, experimentando con las técnicas frecuencia de palabras (WF) e implicación textual (TE) solamente. La razón por la que decidimos realizar este análisis a posteriori fue porque nos dimos cuenta de que en la competición algunos de los sistemas participantes, por ejemplo los propuestos en (Bossard, Généreux, y Poibeau, 2008) o (Hendrickx y Bosma, 2008), se habían centrado solamente en generar un resumen, sin tener en cuenta para nada las frases que indicaban opiniones ni el carácter de las mismas. Por tanto, esto nos permitiría analizar si nuestra propuesta de sistema de resúmenes genérico podía servir para producir resúmenes de estas características.

Partiendo del conjunto de frases candidatas que obtuvimos para cada una de las aproximaciones propuestas (basada en *snippets* y basada en blogs, respectivamente), ejecutamos el sistema de resúmenes genérico sobre cada una de ellas aplicando, en primer lugar, la frecuencia de las palabras, como característica para extraer la información relevante, y combinando, en segundo lugar, esta característica con el módulo de implicación textual como método para eliminar información redundante. De esta manera, podríamos comprobar si los métodos propuestos para producir resúmenes podrían ser adecuados para otros tipos de resúmenes, en este caso, para resúmenes de opiniones.

El cuadro 6.4 muestra los resultados que obtuvimos para este análisis. En él, representamos los valores obtenidos para la precisión, cobertura y medida-F. Cada una de las filas representa el método empleado. Además, se puede observar cómo se obtienen buenos resultados para la medida-F, en comparación con los obtenidos para la competición. Analizando en profundidad estos resultados, obtenemos que la combinación de la frecuencia de las palabras junto con el módulo de implicación textual es apropiada para detectar información relevante. Aunque a priori se pudiera pensar que al utilizar la implicación textual para eliminar información redundante, podríamos estar eliminando también contenido importante, los resultados nos demuestran que nos conviene utilizar dicho módulo, ya que el valor para la precisión se ve beneficiado al contener mayor cantidad de información relevante en menor cantidad de texto. En lo que respecta a la medida-F, los resultados obtenidos al combinar frecuencia de palabras e implicación textual, mejoran los mejores resultados obtenidos en la competición y mejoran también un 80 % los resultados de nuestra primera aproximación, basada en *snippets*, propuesta para el TAC 2008.

Sin embargo, esto no significa necesariamente que los resúmenes generados con este sistema sean mejores que los generados para la competición, y que debamos prescindir del módulo de minería de opiniones que habíamos incorporado. Lamentablemente, sólo pudimos evaluar el contenido de los mismos usando el conjunto de *nuggets* proporcionados en la competición para la evaluación

Aproximación	Cobertura	Precisión	Medida F
Basada en *snippets*	0,592	0,272	0,357
Basada en blogs	0,251	0,141	0,155
Basada en *snippets*+WF	0,705	0,392	0,486
Basada en *snippets*+TE+WF	0,684	0,630	**0,639**
Basada en blogs+WF	0,322	0,234	0,241
Basada en blogs+TE+WF	0,292	0,282	0,262

Cuadro 6.4: Resultados obtenidos por el sistema genérico

de los resúmenes, y no pudimos evaluar otros criterios también importantes como la coherencia del resumen o la cantidad de errores gramaticales de cada uno de ellos. Por lo tanto, puesto que los resultados obtenidos en la competición y en las mejoras realizadas posteriormente respecto al resto de los criterios eran buenos, en un futuro deberíamos combinar ambos sistemas e introducir estos mismos mecanismos en nuestro sistema genérico, junto con el módulo para el tratamiento de opiniones, para conseguir un sistema de resúmenes de opiniones lo más competitivo posible.

Capítulo 7

Marco de evaluación cualitativa de resúmenes

En el capítulo 4 se presentaron los tipos de evaluación existentes, intrínseca y extrínseca, cuantitativa y cualitativa, describiendo a su vez, diferentes herramientas y métodos para evaluar los resúmenes de manera intrínseca (sección 4.1), comentando también los problemas e inconvenientes que éstos presentaban (sección 4.1.1). Una de las principales desventajas era que todos los métodos eran cuantitativos y estaban orientados a la evaluación del contenido del resumen, y para ello, se necesitaban resúmenes de referencia producidos por personas, tarea costosa y difícil de llevar a cabo. También se explicó, como en las conferencias destinadas a evaluar sistemas de generación de resúmenes, toda la parte de la evaluación de aspectos no relacionados con el contenido del resumen, se realizaba de forma manual.

La finalidad de la tarea de generación de resúmenes es su aplicación final en otras tareas para facilitar el manejo de la información existente. Por tanto, aparte de contener la información más importante, es necesario que dicha información se presente de manera clara e inteligible. A partir de la identificación de estas necesidades, se plantea una propuesta para desarrollar un marco de evaluación que evalúe de forma automática un resumen, atendiendo a diferentes criterios de calidad lingüísticos, como los propuestos en el seno de las conferencias DUC y TAC. La idea para este tipo de evaluación es establecer un umbral de calidad para cada criterio evaluado, que determine si un resumen es apto o no en dicho criterio, y por tanto, que dicho resumen se pueda considerar como aceptable. El objetivo no es sustituir a las herramientas existentes, sino intentar ofrecer una forma de evaluación complementaria y alternativa, investigando la difícil tarea de evaluación de resúmenes desde otro punto de vista.

La propuesta de evaluación cualitativa de resúmenes es todavía una propuesta, y por tanto está en un estado muy inicial. La figura 7.1 muestra de forma gráfica, un primer planteamiento de cómo podría ser este marco de evaluación.

Estamos investigando en el desarrollo de un marco de evaluación cualitativo,

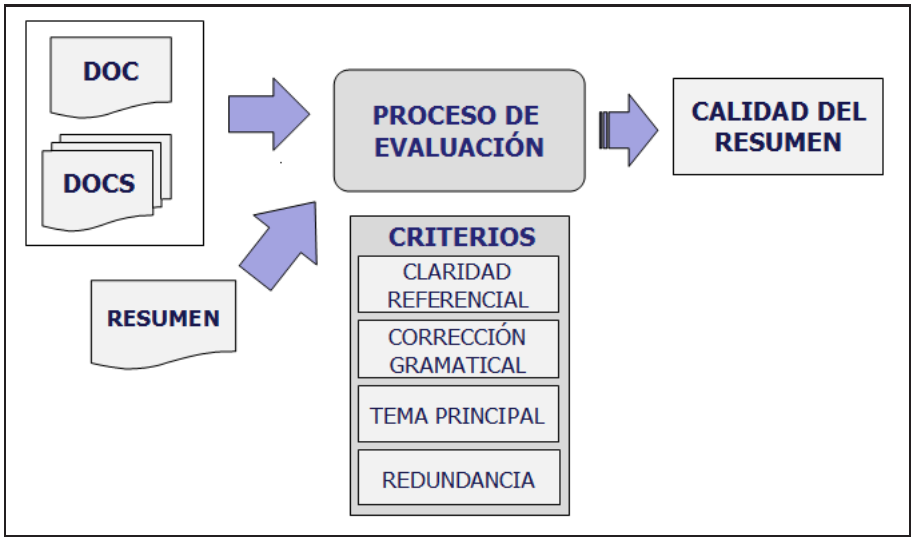

Figura 7.1: Propuesta de evaluación cualitativa

donde básicamente, la idea sería establecer unos criterios de calidad para medir en qué grado el resumen es correcto o no en dicho criterio. Inicialmente, basándonos en las evaluaciones de calidad realizadas manualmente en las conferencias DUC y TAC, se establecen los siguientes criterios: corrección gramatical, identificación del tema principal, claridad en el uso de referencias (correferencia, uso de pronombres personales, etc.) y presencia de información redundante. Nuestra idea es que se trate de una arquitectura modular, y por tanto, tanto la adición como la eliminación de criterios podría ser posible. Como la evaluación de estos criterios lingüísticos no implica la comparación con resúmenes modelos, no sería necesario disponer de resúmenes producidos por humanos para llevar a cabo este tipo de evaluación, y nos ahorraríamos esta difícil tarea. Por el contrario, para algunos de los criterios propuestos, como por ejemplo el de la claridad en cuanto a las referencias que aparecen en el resumen, o para saber si el tema principal del documento está correctamente identificado, sí que necesitaríamos el documento o documentos originales a partir de los que se ha generado el resumen. Actualmente, este marco de evaluación es solamente una propuesta, ya que cada uno de los criterios sugeridos necesita ser analizado y estudiado en detalle, para determinar, por un lado, la viabilidad o no de cada criterio en sí y, por otro, cómo afecta dicho criterio en el resumen y si es posible evaluarlo de forma automática. También deberemos estudiar los valores óptimos para los respectivos umbrales de calidad a partir de los cuales, el resumen sería aceptable en cada criterio, y por tanto de forma global.

Nuestra hipótesis de trabajo es que cada uno de los criterios propuestos puede ser automatizado, y para justificar la viabilidad general de la propuesta de evaluación cualitativa, vamos a presentar a continuación un breve análisis de cada uno de los criterios de calidad comentados anteriormente. La corrección

gramatical de un texto tiene que ver con que dicho texto carezca de errores ortográficos y gramaticales, por ejemplo, que no tenga palabras mal escritas, con faltas de ortografía, y a su vez, que no presente fallos en cuanto a la concordancia de sujeto-verbo, determinante-pronombre, etc. se refiere. Esto varía en función del idioma que se trate, puesto que cada idioma tiene sus reglas ortográficas y gramaticales concretas. Existen herramientas, como *Language Tool*[1] o simplemente los correctores ortográficos incorporados en la mayoría de los procesadores de texto, que pueden usarse para llevar a cabo este cometido. La ausencia de redundancia en un resumen se refiere a que éste no debe contener información duplicada, puesto que si no, el resumen estaría perdiendo capacidad informativa. El problema de la redundancia de información es un problema importante en el proceso de generación de resúmenes también. Por ello, en (Newman et al., 2004) se proponen varios métodos que se pueden emplear para detectar información redundante en un documento. Además, la propuesta de generación de resúmenes presentada en este trabajo utiliza la técnica de implicación textual para abordar este problema. Por otra parte, en relación al criterio de claridad referencial, lo que se pretende es comprobar, de alguna manera, que el antecedente de los pronombres o referencias que aparecen en un resumen sean realmente el elemento o elementos pertinentes y no se haga referencia a otro de forma errónea, dando lugar a incongruencias o malentendidos en el texto. Esto se podía abordar por ejemplo, analizando las herramientas existente para resolver la anáfora o correferencia, y siguiendo un enfoque similar al propuesto en (Steinberger et al., 2007), donde se comprueba que las cadenas de correferencia en un resumen se deben corresponder con las cadenas de correferencia identificadas en el documento original. Finalmente, respecto al criterio de identificación del tema principal de un documento, la evaluación consistiría en determinar si el resumen contempla claramente el tema principal. La evaluación de este criterio se podría realizar mediante el análisis de las palabras clave, o bien utilizando algoritmos existentes para la segmentación de documentos en temas y subtemas (como por ejemplo, *TextTiling* (Hearst, 1997) o *C99* (Choi, 2000)).

Al tratarse sólo de una propuesta, lo que se pretende realizar en el futuro es estudiar hasta qué punto y de qué manera cada uno de esos criterios están presentes en resúmenes humanos, identificando las relaciones existentes respecto a los documentos originales, asumiendo que un resumen producido por un humano será mejor desde el punto de vista lingüístico que uno generado automáticamente. Una vez analizados los criterios en resúmenes humanos y de analizar los métodos o características que mejor identifiquen cada uno de ellos, se procederá a aplicarlos en resúmenes automáticos para estudiar su comportamiento y desarrollar el marco de evaluación cualitativa.

[1] http://community.languagetool.org

Capítulo 8

Conclusiones y trabajo en progreso

Recapitulando todo el trabajo realizado y presentado en esta memoria de investigación, podemos extraer algunas conclusiones interesantes respecto al trabajo desarrollado. Sin embargo, también hay que tener en cuenta que esta memoria constituye sólo un primer paso en la tarea de generación automática de resúmenes, quedando muchos aspectos para mejorar. Algunos de ellos se expondrán como trabajos para un futuro; otros, sin embargo, se quedarán en el aire, debido a la imposibilidad de abarcar todos los problemas asociados con esta tarea del PLN.

8.1. Conclusiones

En esta memoria de investigación se ha realizado un repaso a las distintas técnicas y sistemas que se han propuesto y desarrollado con la finalidad de generar resúmenes de forma automática. También se han descrito las herramientas existentes para la evaluación de resúmenes, destacando la dificultad que conlleva esta tarea, e identificando los problemas e inconvenientes que tienen los métodos de evaluación actuales.

Además, en este trabajo de investigación se ha presentado una propuesta de un sistema de resúmenes automático mono-documento, que integra y combina varias técnicas que influyen de manera positiva en la tarea de generación de resúmenes. Las técnicas estudiadas han sido: la frecuencia de las palabras, el reconocimiento de la implicación textual y el principio de la cantidad de codificación. En primer lugar, decidimos realizar un análisis individual de cada una de ellas, para posteriormente combinarlas y analizar los resultados obtenidos. Básicamente, el objetivo de estas técnicas es detectar las frases relevantes de un documento, para establecer un ránking y extraer las más importantes según la clasificación que cada técnica establece. Además del citado módulo de detección de información relevante, el sistema propuesto necesita una fase de preprocesamiento

para eliminar la posible información ruidosa del documento, así como una fase posterior de postprocesamiento, que agrupará las frases seleccionadas, de tal manera que se genere el resumen final teniendo en cuenta factores como la coherencia. La evaluación de la propuesta se realizó utilizando el conjunto de documentos proporcionado en la edición del DUC 2002. Concretamente, se trataba de noticias periodísticas en inglés. A partir de dichos datos, se realizó un conjunto de experimentos para probar el funcionamiento del sistema y compararlo con otros sistemas participantes en dicha edición. Adicionalmente, se ha presentado también el trabajo que se está desarrollando para generar resúmenes multi-documento, por un lado, y analizar su portabilidad a otros dominios, como es el caso del dominio literario (a través de cuentos infantiles) por otro.

Por otro lado, se ha analizado también la tarea de generación de resúmenes mediante su aplicación en la tarea de minería de opiniones. A través de la participación en el TAC 2008, se ha presentado un sistema capaz de producir resúmenes de opiniones, localizando la información deseada en una colección de blogs proporcionados por la organización de dicho congreso. La información a incluir en el resumen venía marcada por la formulación de una pregunta que requería respuestas que fueran opiniones sobre un tema, persona, organización, etc. Dicho sistema de resúmenes se construyó a partir de un módulo de tratamiento de la pregunta, mediante el cuál se identificaba el foco de la pregunta y su polaridad; un módulo de tratamiento de los fragmentos de textos que contenían la respuesta a la pregunta (bien fuera a partir de los *snippets* proporcionados por el TAC, o bien a partir de fragmentos localizados directamente en los blogs); y finalmente, un módulo de generación del resumen, en el que se agrupaban las oraciones candidatas, y se realizaban una serie de modificaciones, como por ejemplo, transformar las oraciones a estilo indirecto, cambiar frases subjetivas por impersonales, o añadir fragmentos de texto, para dar al resumen un carácter abstractivo. Los resultados obtenidos demostraron que nuestro sistema de resúmenes de opiniones funcionaba bien en algunos aspectos, como la coherencia o medida-F, pero fallaba en la corrección gramatical, conteniendo además, gran cantidad de información redundante. Como consecuencia, decidimos mejorar el sistema en estos aspectos, utilizando el reconocimiento de la implicación textual para eliminar la redundancia de información, y un corrector automático para solventar los problemas de errores gramaticales. Además, decidimos investigar hasta qué punto, la propuesta del sistema genérico descrita en el capítulo 5 podía servir para producir resúmenes de este tipo, viendo que era conveniente incorporar un módulo de identificación y clasificación de opiniones.

Finalmente se ha propuesto un marco de evaluación para determinar la corrección de un resumen automático en función de unos criterios de calidad lingüísticos preestablecidos. Esta propuesta, motivada por la dificultad y la escasez de recursos para evaluar los resúmenes en cuanto a aspectos de carácter lingüísticos se refiere, está todavía en un estado preliminar y pretende servir como marco para poder definir la calidad de un resumen. La idea básica que se ha presentado consiste en definir unos criterios, como la redundancia de información o la corrección

gramatical, analizarlos, ver cómo están presentes en resúmenes humanos respecto a los correspondientes documentos originales, y establecer un umbral a partir del cual se pueda considerar que un resumen es aceptable para cada uno de los criterios.

De cada uno de los estudios se extraen conclusiones en diferentes líneas:

- **Identificación de las dificultades en la tarea de evaluación automática de resúmenes**

  Los sistemas actuales de evaluación automática de resúmenes, aunque ayudan y facilitan la difícil tarea de decidir si un resumen es correcto o no, presentan todavía ciertas limitaciones. Entre ellas, destacamos la necesidad de tener resúmenes humanos que sirvan como modelo para realizar la comparación con los resúmenes automáticos. Disponer de resúmenes humanos es una tarea muy costosa, y el hecho de que un resumen no tenga parecido con uno manual, no implica necesariamente que el resumen esté mal. Otro de los inconvenientes de estos sistemas es que sólo se preocupan de la información contenida en el resumen, y no de cómo ésta está organizada y estructurada. La evaluación de aspectos lingüísticos, como la coherencia o la corrección gramatical se realizan de forma manual.

- **Análisis y comparación de tres técnicas para la generación de resúmenes automáticos**

  En lo que respecta al sistema de generación de resúmenes propuesto, las tres técnicas estudiadas (frecuencia de palabras, implicación textual y el principio de la cantidad de codificación) tienen una influencia positiva en el sistema, tal y como demuestran los resultados obtenidos. Además, la combinación de técnicas es apropiada, obteniendo una mejora del 12 % respecto a utilizar sólo una técnica, y superando los resultados obtenidos por otros sistemas similares en un 10 %. De forma individual, la técnica que mejores resultados obtiene para el dominio periodístico a partir de los datos del DUC 2002 es la frecuencia de las palabras, que da preferencia a las frases que contengan palabras que aparezcan mucho en el documento.

- **Análisis de la validez de las técnicas estudiadas para generar otros tipos de resúmenes**

  Como experimentos adicionales, se ha propuesto extender el sistema propuesto a multi-documento (proponiendo una sencilla aproximación y viendo los resultados obtenidos), y también a otro tipo de documentos (cuentos infantiles). Los resultados obtenidos para los resúmenes multi-documento demuestran que la técnica de implicación textual es válida y eficaz para eliminar información redundante y por tanto, ayuda a que los resúmenes no contengan información repetida, mientras que se ha observado que el Principio de cantidad de codificación funciona mejor para el dominio de cuentos infantiles, ya que se trata de textos narrativos, y por tanto, incorporan secuencias de acciones, descripción de personajes, de

escenarios, etc. incluyendo mayor número de sintagmas nominales que en los documentos pertenecientes al domino periodístico.

- **Aplicación de los resúmenes a la tarea de minería de opiniones**

  La tarea de generación de resúmenes se puede aplicar a distintos escenarios reales, entre ellos, la generación de resúmenes de opiniones. A través de la competición propuesta en el TAC 2008, se desarrollaron dos aproximaciones distintas para lograr la obtención de resúmenes de opiniones. Para ello fue necesario combinar técnicas de identificación y clasificación de opiniones (propias de la tarea de minería de opiniones), junto con técnicas de generación de resúmenes. Los resultados derivados de la participación revelaron que se trata de una tarea difícil, pero aun así, nuestros enfoques consiguieron buenos resultados en algunos aspectos (medida-F, estructura y coherencia), mientras que no tan buenos en otros (redundancia, corrección gramatical), pero que fueron corregidos y mejorados a posteriori.

- **Marco de evaluación cualitativo en base a criterios de calidad**

  Las herramientas existentes actualmente se centran sólo en el contenido del resumen, evaluando su similitud respecto a resúmenes de referencia escritos por humanos. Sin embargo, para que un resumen automático sea útil y se pueda aplicar a otras tareas, es necesario que el resumen, además de contener la información relevante, esté organizado y estructurado de manera clara y concisa, respetando ciertos criterios lingüísticos para que se pueda leer y entender. Por lo tanto, surge la necesidad de poder evaluar automáticamente un resumen atendiendo a diversos criterios como la redundancia de información, corrección gramatical, etc. La propuesta planteada persigue esta finalidad, proponiendo unos criterios específicos que tendrán que ser analizados con mayor detenimiento y ver si realmente se pueden automatizar para determinar la calidad de un resumen automático.

8.2. Trabajo futuro

Como ya se ha comentado, esta memoria de investigación abarca sólo una ínfima parcela de todo lo que brinda la tarea de generación de resúmenes, quedando muchos aspectos por mejorar todavía. Por consiguiente, como trabajo futuro se propone continuar la investigación en base a tres objetivos bien diferenciados:

- Construcción de un sistema de generación de resúmenes:

  - Estudiar y analizar otras técnicas, métodos o tareas del PLN que se puedan aplicar a la generación de resúmenes, para detectar información relevante en un documento, tales como resolución de la correferencia.
  - Experimentar con otros tipos de documentos (páginas web, blogs, etc.) y dominios (literario, científico, etc.).

  - Analizar la bondad del sistema propuesto y de las técnicas empleadas para su aplicación al español.
  - Profundizar en el módulo de postprocesamiento, investigando en técnicas que ayuden a producir resúmenes siguiendo un paradigma abstractivo, presentando el resumen final de forma organizada y coherente.
  - Realizar una evaluación extrínseca del sistema propuesto, analizando hasta qué punto los resúmenes generados son útiles para ayudar a tareas de recuperación de información y búsqueda de respuestas.

- Aplicación de resúmenes a tareas concretas:

  - Investigar en técnicas para generar resúmenes de opiniones, combinando las tareas de minería de opiniones y de resúmenes.
  - Estudiar otros posibles escenarios de aplicación.
  - Analizar las ventajas e inconvenientes derivados del uso de resúmenes automáticos en diferentes tareas frente a utilizar los documentos completos.

- Análisis y desarrollo de un marco de evaluación automático de resúmenes cualitativo:

  - Investigar distintos criterios que permitan evaluar aspectos lingüísticos de un resumen.
  - Estudiar cada uno de los criterios propuestos, analizando la relación documento-resumen, tanto para resúmenes humanos como para resúmenes automáticos.
  - Establecer un umbral de forma experimental para determinar la calidad del resumen de acuerdo a cada uno de los criterios definidos.
  - Integrar los distintos criterios dentro de un marco de evaluación automático.

Bibliografía

Amigó, Enrique, Julio Gonzalo, Anselmo Peñas, y Felisa Verdejo. 2005. QARLA: a framework for the evaluation of text summarization systems. En *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, páginas 280–289.

Angheluta, Roxana, Rik De Busser, y Marie-Francine Moens. 2002. The use of topic segmentation for automatic summarization. En *Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization*, páginas 66–70.

Aone, Chinatsu, Mary Ellen Okurowski, y James Gorlinsky. 1998. Trainable, scalable summarization using robust NLP and machine learning. En *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, páginas 62–66, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Azzam, S., K. Humphreys, y R. Gaizauskas. 1999. Using coreference chains for text summarization. En *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, Baltimore, June.

Balahur, A., E. Lloret, O. Ferrández, A. Montoyo, M. Palomar, y R. Muñoz. 2008. The DLSIUAES team's participation in the TAC 2008 tracks. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 177–188.

Balahur, A. y A. Montoyo. 2008. An incremental multilingual approach to forming a culture dependent emotion triggers database. En *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE 2008), Copenhagen*.

Baldwin, Breck y Thomas S. Morton. 1998. Dynamic coreference-based summarization. En *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*.

Barzilay, Regina y Michael Elhadad. 1999. Using lexical chains for text summarization. En *Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization, páginas 111–122. MIT Press.

Barzilay, Regina y Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, páginas 141–148, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Beineke, Philip, Trevor Hastie, Christopher Manning, y Shivakumar Vaithyanathan. 2004. An exploration of sentiment summarization. En *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US.

Bellemare, S., S. Bergler, y R. Witte. 2008. ERSS at TAC 2008. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 122–124.

Biadsy, Fadi, Julia Hirschberg, y Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. En *Proceedings of ACL-08: HLT*, páginas 807–815, Columbus, Ohio, June. Association for Computational Linguistics.

Boguraev, Branimir K. y Mary S. Neff. 2000. Discourse segmentation in aid of document summarization. En *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3*, página 3004.

Bossard, A., M. Généreux, y T. Poibeau. 2008. Description of the LIPN systems at TAC 2008: Summarizing information and opinions. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 282–300.

Burges, Chis, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, y Greg Hullender. 2005. Learning to rank using gradient descent. En *ICML '05: Proceedings of the 22nd international conference on Machine learning*, páginas 89–96, New York, NY, USA. ACM.

Cesarano, Carmine, Antonino Mazzeo, y Antonio Picariello. 2007. A system for summary-document similarity in notary domain. *International Workshop on Database and Expert Systems Applications*, 0:254–258.

Choi, Freddy Y. Y. 2000. Advances in domain independent linear text segmentation. En *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics (NAACL)*, páginas 26–33.

Conroy, J.M. y J.D. Schlesinger. 2008. CLASSY at TAC 2008 metrics. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 132–141.

Conroy, John M. y Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. En *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, páginas 145–152, Manchester, UK, August. Coling 2008 Organizing Committee.

Conroy, John M. y Dianne P. O'leary. 2001. Text summarization via Hidden Markov Models. En *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 406–407.

Díaz, Alberto y Pablo Gervás. 2007. User-model based personalized summarization. *Information Processing & Management*, 43(6):1715–1734.

Donaway, Robert L., Kevin W. Drummey, y Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. En *Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization*, páginas 69–78.

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

Dunlavy, Daniel M., Dianne P. O'Leary, John M. Conroy, y Judith D. Schlesinger. 2007. QCS: A system for querying, clustering and summarizing documents. *Information Processing & Management*, 43(6):1588–1605.

Edmundson, H. P. 1969. New methods in automatic extracting. En *Inderjeet Mani and Mark Maybury, editors*, Advances in Automatic Text Summarization, páginas 23–42. MIT Press.

Elsner, Micha y Eugene Charniak. 2008. Coreference-inspired coherence modeling. En *Proceedings of ACL-08: HLT, Short Papers*, páginas 41–44, Columbus, Ohio, June. Association for Computational Linguistics.

Ercan, Gonenc y Ilyas Cicekli. 2008. Lexical cohesion based topic modeling for summarization. En *Proceedings of the 9th International Conference in Computational Linguistics and Intelligent Text Processing, CICLing 2008, Haifa, Israel, February 17-23*, páginas 582–592.

Ferrández, Óscar, Daniel Micol, Rafael Muñoz, y Manuel Palomar. 2007. A perspective-based approach for solving textual entailment recognition. En *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, páginas 66–71, Prague, June. Association for Computational Linguistics.

Fuentes, María, Edgar González, Daniel Ferrés, y Horacio Rodríguez. 2005. QASUM-TALP at DUC 2005 automatically evaluated with a pyramid based metric. En *Document Understanding Workshop (presented at the* HLT/EMNLP *Annual Meeting), Vancouver, B.C., Canada*.

Fuentes, María, Horacio Rodríguez, y Daniel Ferrés. 2007. FEMsum at DUC 2007. En *Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*.

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, y Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. En *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, páginas 1–9, Prague, June. Association for Computational Linguistics.

Giannakopoulos, G., V. Karkaletsis, y G. Vouros. 2008. Testing the use of n-gram graphs in summarization sub-tasks. En *Proceedings of the Text Analysis Conference (TAC)*.

Giannakopoulos, George, Vangelis Karkaletsis, George Vouros, y Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing*, 5(3):1–39.

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

Glickman, Oren. 2006. *Applied Textual Entailment*. Ph.D. tesis, Bar Ilan University.

Gotti, Fabrizio, Guy Lapalme, Luka Nerima, y Eric Wehrli. 2007. GOFAISUM: A symbolic summarizer for DUC. En *Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*.

Harabagiu, Sanda, Andrew Hickl, y Finley Lacatusu. 2007. Satisfying information needs with multi-document summaries. *Information Processing & Management*, 43(6):1619–1642.

Harabagiu, Sanda y Finley Lacatusu. 2005. Topic themes for multi-document summarization. En *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 202–209.

Hassel, Martin. 2007. *Resource Lean and Portable Automatic Text Summarization*. Ph.D. tesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden.

He, T., J. Chen, Z. Gui, y F. Li. 2008. CCNU at TAC 2008: Proceeding on using semantic method for automated summarization. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 100–107.

Hearst, Marti A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hendrickx, I. y W. Bosma. 2008. Using coreference links and sentence compression in graph-based summarization. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 429–435.

Hovy, Eduard, Chin-Yew Lin, Liang Zhou, y Junichi Fukumoto. 2006. Automated summarization evaluation with Basic Elements. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy*.

Hovy, E.H. y C-Y. Lin. 1999. Automated multilingual text summarization and its evaluation. Informe técnico, Information Sciences Institute, University of Southern California.

Iftene, A. y A Balahur-Dobrescu. 2007. Hypothesis transformation and semantic variability rules for recognizing Textual Entailment. En *Proceedings of the ACL 2007 Workshop on Textual Entailment and Paraphrasis*.

Jaoua, Maher y Abdelmajid Ben Hamadou. 2003. Automatic text summarization of scientific articles based on classification of extract's population. En *Proceedings of Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing 2003, Mexico City, Mexico, February 16-22, 2003*, páginas 623–634.

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

Jones, Karen Sparck y Julia Rose Galliers, editores. 1996. *Evaluating Natural Language Processing Systems, An Analysis and Review*, volumen 1083 de *Lecture Notes in Computer Science*. Springer.

Kazantseva, Anna. 2006. An approach to summarizing short stories. En *Proceedings of the Student Research Workshop at the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Khan, Afnan Ullah, Shahzad Khan, y Waqar Mahmood. 2005. MRST: A new technique for information summarization. En *The Second World Enformatika Conference, WEC'05, February 25-27, 2005, Istanbul, Turkey, CDROM*, páginas 249–252.

Kuo, June-Jei y Hsin-Hsi Chen. 2008. Multidocument summary generation: Using informative and event words. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(1):1–23.

Kupiec, Julian, Jan Pedersen, y Francine Chen. 1995. A trainable document summarizer. En *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 68–73.

Lapata, Mirella y Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. En *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK*, páginas 1085–1090.

Li, S., W. Wan, y C. Wang. 2008. TAC 2008 update summarization task of ICL. En *Proceedings of the Text Analysis Conference (TAC)*, páginas 221–226.

Li, Sujian, You Ouyang, Wei Wang, y Bin Sun. 2007. Multi-document summarization using support vector regression. En *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*.

Lin, Chin-Yew. 2004. ROUGE: a package for automatic evaluation of summaries. En *Proceedings of ACL Text Summarization Workshop*, páginas 74–81.

Lin, Chin-Yew y Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. En *Proceedings of the 18th conference on Computational linguistics*, páginas 495–501.

Liu, Feifan y Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. En *Proceedings of ACL-08: HLT, Short Papers*, páginas 201–204, Columbus, Ohio, June. Association for Computational Linguistics.

Lloret, Elena, Alexandra Balahur, Andrés Montoyo, y Manuel Palomar. 2009. Towards building a competitive opinion summarization system: Challenges and keys. En *Proceedings of the NAACL Student Research Workshop*, páginas 72–77.

Lloret, Elena, Óscar Ferrández, Rafael Muñoz, y Manuel Palomar. 2008a. Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, (41):183–190.

Lloret, Elena, Óscar Ferrández, Rafael Muñoz, y Manuel Palomar. 2008b. A Text Summarization Approach Under the Influence of Textual Entailment. En *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) 12-16 June, Barcelona, Spain*, páginas 22–31.

luc Minel, Jean, Sylvaine Nugier, y Gérald Piat. 1997. How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. En *Proceedings of Intelligent Scalable Text Summarization Workshop in conjunction with the European Chapter of the Association of Computational Linguistics (EACL)*, páginas 25–30.

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.

Mani, Inderjeet. 2001a. *Automatic Summarization*. John Benjamins Pub Co.

Mani, Inderjeet. 2001b. Summarization evaluation: An overview. En *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL). Workshop on Automatic Summarization*.

Mani, Inderjeet y Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. The MIT Press.

Mann, William C. y Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, Daniel. 1999. Discourse trees are good indicators of importance in text. En *Inderjeet Mani and Mark Maybury, editors*, Advances in Automatic Text Summarization, páginas 123–136. MIT Press.

Medelyan, Olena. 2007. Computing lexical chains with graph clustering. En *Proceedings of the ACL 2007 Student Research Workshop*, páginas 85–90, Prague, Czech Republic, June. Association for Computational Linguistics.

Mihalcea, Rada. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. En *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, página 20.

Mihalcea, Rada y Hakan Ceylan. 2007. Explorations in automatic book summarization. En *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, páginas 380–389.

Mitkov, Ruslan, Richard Evans, Constantin Orasan, Le An Ha, y Viktor Pekar. 2007. Anaphora resolution: To what extent does it help nlp applications? En *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007, Lagos, Portugal, March 29-30*, páginas 179–190.

Mittal, Vibhu, Mark Kantrowitz, Jade Goldstein, y Jaime Carbonell. 1999. Selecting text spans for document summaries: heuristics and metrics. En *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial Intelligence and the eleventh Innovative Applications of Artificial Intelligence conference*, páginas 467–473.

Morales, Laura Plaza, Alberto Díaz Esteban, 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 (SEPLN)*, (41):191–198.

Moreno Boronat, Lidia, Manuel Palomar Sanz, Antonio Molina Marco, y Antonio Ferrández Rodríguez. 1999. *Introducción al procesamiento del lenguaje natural*. Universidad de Alicante.

Nenkova, Ani. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. En *Proceedings of theAmerican Association fro Artificial Intelligence (AAAI)*, páginas 1436–1441.

Nenkova, Ani. 2006. Summarization evaluation for text and speech: issues and approaches. En *INTERSPEECH-2006*, páginas 1527–1530.

Nenkova, Ani, Rebecca Passonneau, y Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2):4.

Nenkova, Ani, Advaith Siddharthan, y Kathleen McKeown. 2005. Automatically learning cognitive status for multi-document summarization of newswire. En *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, páginas 241–248.

Nenkova, Ani, Lucy Vanderwende, y Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. En *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 573–580.

Neto, Joel Larocca, Alexandre Santos, Celso A. A. Kaestner, y Alex Alves Freitas. 2000. Generating text summaries through the relative importance of topics. En *IBERAMIA-SBIA '00: Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI*, páginas 300–309.

Newman, Eamonn, William Doran, Nicola Stokes, Joe Carthy, y John Dunnion. 2004. Comparing redundancy removal techniques for multi-document summarisation. En *Proceedings of STAIRS, August*, páginas 223–228.

Nunes Gonçalves, P., Lucia Rino, y Renata Vieira. 2008. Summarizing and referring: towards cohesive extracts. En *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, páginas 253–256.

Orasan, Constantin. 2004. The influence of personal pronouns for automatic summarisation of scientific articles. En *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2004, Furnas, S. Migue, Azores, Protugal, September, 23 - 24*, páginas 127–132.

Orasan, Constantin. 2007. Pronominal anaphora resolution for text summarisation. En *Proceedings of the Recent Advances , RANLP 2007, Borovets, Bulgaria, September, 27 - 29*, páginas 430–436.

Pedersen, Ted y Siddharth Patwardhan. 2004. Wordnet::similarity - measuring the relatedness of concepts. En *In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04*, páginas 1024–1025.

Pitler, Emily y Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. En *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, páginas 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Plaza, Laura, Alberto Díaz, y Pablo Gervás. 2008. Concept-graph based biomedical automatic summarization using ontologies. En *Coling 2008: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, páginas 53–56, Manchester, UK, August. Coling 2008 Organizing Committee.

Qiu, Li-Qing, Bin Pang, Sai-Qun Lin, y Peng Chen. 2007. A novel approach to multi-document summarization. En *Proceedings of the 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), 3-7 September 2007, Regensburg, Germany*, páginas 187–191.

Radev, Dragomir R., Sasha Blair-Goldensohn, y Zhu Zhang. 2001. Experiments in single and multi-document summarization using MEAD. En *First Document Understanding Conference, New Orleans, LA*, páginas 1–7.

Radev, Dragomir R., Eduard Hovy, y Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Radev, Dragomir R. y Daniel Tam. 2003. Summarization evaluation using relative utility. En *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, páginas 508–511.

Ramshaw, Lance A. y Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. En *Proceedings of the Third ACL Workshop on Very Large Corpora, Cambridge MA, USA*.

Saggion, Horacio y Guy Lapalme. 2000. Selective analysis for automatic abstracting: Evaluating indicativeness and acceptability. En *Proceedings of Content-Based Multimedia Information Access (RIAO)*, páginas 747–764.

Saravanan, M., B. Ravindran, y S. Raman. 2006. Improving legal document summarization using graphical models. En *Proceedings of the Nineteenth Annual Conference on Legal Knowledge and Information Systems (JURIX 2006), Paris, France, 7-9 December*, páginas 51–60.

Schilder, Frank y Ravikumar Kondadadi. 2008. FastSum: Fast and accurate query-based multi-document summarization. En *Proceedings of ACL-08: HLT, Short Papers*, páginas 205–208, Columbus, Ohio, June.

Schlesinger, J. D., M. E. Okurowski, J. M. Conroy, D. P. O'Leary, A. Taylor, J. Hobbs, y H. Wilson. 2002. Understanding machine performance in the context of human performance for multi-document summarization. En

Proceedings of the DUC 2002 Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia.

Sjöbergh, Jonas. 2007. Older versions of the ROUGEeval summarization evaluation system were easier to fool. *Information Processing & Management*, 43(6):1500–1505.

Spärck Jones, Karen. 1999. Automatic summarizing: factors and directions. En *Inderjeet Mani and Mark Maybury, editors*, Advances in Automatic Text Summarization, páginas 1–14. MIT Press.

Spärck Jones, Karen. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Steinberger, Josef, Karel Jezek, y Martin Sloup. 2008. Web topic summarization. En *Proceedings of the 12th International Conference on Electronic Publishing (ELPUB, Toronto, Canada, 25-27 June*, páginas 322–334.

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.

Strapparava, C. y A. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. En *Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May*, páginas 1083–1086.

Svore, Krysta M., Lucy Vanderwende, y Christopher J.C. Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. En *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, páginas 448–457.

Svore, Krysta Marie, Lucy Vanderwende, y Christopher J. C. Burges. 2008. Using signals of human interest to enhance single-document summarization. En *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17*, páginas 1577–1580.

Sweeney, Simon, Fabio Crestani, y David E. Losada. 2008. 'Show me more': Incremental length summarisation using novelty detection. *Information Processing & Management*, 44(2):663–686.

Tatar, Doina, Emma Tamaianu-Morita, Andreea Mihis, y Dana Lupsa. 2008. Summarization by logic segmentation and text entailment. En *Proceedings of Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008)*, páginas 15–26.

Teng, Zhi, Ye Liu, Fuji Ren, Seiji Tsuchiya, y Fuji Ren. 2008. Single document summarization based on local topic identification and word frequency. En *MICAI '08: Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence*, páginas 37–41, Washington, DC, USA. IEEE Computer Society.

Teufel, S. y H. van Halteren. 2004. Evaluating information content by factoid analysis : Human annotation and stability. En *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 25 - 26 July, Barcelona, Spain*, páginas 419–426.

Teufel, Simone y Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistis*, 28(4):409–445.

Titov, Ivan y Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. En *Proceedings of ACL-08: HLT*, páginas 308–316, Columbus, Ohio, June.

Toutanova, Kristina, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamudi, Hisami suzuki, y Lucy Vanderwende. 2007. The PYTHY summarization system: Microsoft research at duc 2007. En *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*.

Van Rijsbergen, C. J. 1981. *Information Retrieval*. Elsevier Science & Technology. URL: *http://www.dcs.gla.ac.uk/Keith/Preface.html*.

Wan, Xiaojun, Jianwu Yang, y Jianguo Xiao. 2006. The great importance of cross-document relationships for multi-document summarization. En *Proceedings of the 21st International Conference Computer Processing of Oriental Languages (ICCPOL 2006), Singapore*, páginas 131–138.

Wan, Xiaojun, Jianwu Yang, y Jianguo Xiao. 2007. Towards a unified approach based on affinity graph to various multi-document summarizations. En *Proceedings of the 11th European Conference, ECDL 2007, Budapest, Hungary*, páginas 297–308.

Wang, Chan, Lixia Long, y Lei Li. 2008. HowNet based evaluation for chinese text summarization. En *Proceedings of the International Conference on Natural Language Processing and Software Engineering (IEEE-NLPKE 2008, Beijing, China, October 19-22*, páginas 82–87.

Witte, René, Ralf Krestel, y Sabine Bergler. 2005. ERSS 2005: Coreference-based summarization reloaded. En *Proceedings of Document Understanding Workshop (DUC)*, Vancouver, B.C., Canada, October 9–10.

Witte, René, Ralf Krestel, y Sabine Bergler. 2007. Generating update summaries for DUC 2007. En *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*.

Wong, Kam-Fai, Mingli Wu, y Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. En *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, páginas 985–992, Manchester, UK, August.

Yang, Xiao-Peng y Xiao-Rong Liu. 2008. Personalized multi-document summarization in information retrieval. *International Conference on Machine Learning and Cybernetics*, 7:4108–4112, July.

Zhou, Liang, Chin-Yew Lin, Dragos Stefan Munteanu, y Eduard Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. En *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY*, páginas 447–454.

Zhou, Liang, Miruna Ticrea, y Eduard Hovy. 2004. Multi-document biography summarization. En *Proceedings of the International Conference on Empirical Methods in NLP (ENMLP). Barcelona Spain*, páginas 434–441.

Zhuang, Li, Feng Jing, y Xiao-Yan Zhu. 2006. Movie review mining and summarization. En *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*, páginas 43–50.

Apéndice A

Aportaciones científicas a través de publicaciones

Publicaciones

- Lloret, E., Ferrández, O., Muñoz, R., Palomar, M.: A Text Summarization Approach Under the Influence of Textual Entailment. En: Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) 12–16 June, Barcelona, Spain. (2008) 22–31
- Lloret, E., Ferrández, O., Muñoz, R., Palomar, M.: Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) (41) (2008) 183–190
- Balahur, A., Lloret, E., Ferrández, O., Montoyo, A., Palomar, M., Muñoz, R.: The DLSIUAES team's participation in the TAC 2008 tracks. En: Proceedings of the Text Analysis Conference (TAC). (2008) 177–188
- Lloret, E., Balahur, A., Montoyo, A., Palomar, M.: Towards building a competitive opinion summarization system: Challenges and keys. En: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium. (NAACL 2009) 72–77
- Lloret, E., Palomar, M.: Challenging issues of automatic summarization: Relevance detection and quality-based evaluation. Informatica (2009) Aceptado. Pendiente de publicación.
- Lloret, E., Palomar, M.: A Gradual Combination of features for Building Automatic Summarisation Systems. En: Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD 2009) Aceptado. Pendiente de publicación.

Becas de investigación

- Desde 01/09/2007: Beca predoctoral de Formación de Personal Investigador (FPI) del Ministerio de Ciencia e Innovación. Asociada al proyecto “TEXT-MESS” (TIN2006-15265-C06-01). Referencia de la beca: BES-2007-16268.