Universidad de las Ciencias Informáticas

Facultad 3

Algoritmos para la construcción automática de resúmenes de documentos de texto.

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autores: Yordis Monteserin Matos
Alex Rosales Hechavarría

Tutor: Lic. Enrique Matos Alfonso

Asesor: Ing. Daniel Mariano García Fernández

La Habana, Junio 23, 2007

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo al <nombre área> de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

"[Insertar nombre(s) de autor(es)]"

_____

"[Insertar nombre(s) de tutor(es)]"

_____

AGRADECIMIENTOS

Deseamos agradecer a todos aquellos que han colaborado, responsable o irresponsablemente, de forma directa o indirecta, concientes o no, a que esta investigación llegara a feliz término. En especial, agradecemos a:

Al Dave, sin cuya ayuda, tecleo e insomnios no hubiera sido posible terminar este trabajo.

Al Cesar, cuyas consultas telefónicas cimentaron nuestro diseño teórico.

A Yoansy y Z., cuya colaboración sirvió para culminar las pruebas inferenciales de la evaluación.

A Yari y Dailín, que ayudaron a encontrar bibliografía valiosa.

A Ari, por su paciencia infinita, confianza y apoyo.

A Enrique Matos, nuestro tutor, por su apoyo y consejos.

DEDICATORIA

Quiero dedicar esta tesis a mis padres Lidia y César por dedicarse a mi formación y educación desde que nací, para convertirme en lo que soy hoy; a mis hermanos Alaín, Arianna y Aliuska por brindarme su apoyo siempre que lo he necesitado; a mi cuñado Teuris por ser como un hermano para mí, a mis amigos por brindarme su amistad. Y en especial a mi abuela.

Alex

A mis padres, por su dedicación, empeño y fe inquebrantable.

A los Joses, cuya fidelidad sigue esperando por un libro.

A la memoria de mi abuelo, tronco que no se destiñe.

A Yadira, por simplemente ser.

A Lucas, mi otro yo.

A tantos y tantos amigos.

Yordis

RESUMEN

Producto de la inmensa cantidad de texto existente en la actualidad y la incapacidad de las personas de procesar tanta información, el presente trabajo desarrolla un estudio crítico sobre la construcción automática de resúmenes de texto basados en las técnicas de extracción, limitado a la generación de resúmenes monodocumentos. La finalidad de esta investigación es la de analizar las posibilidades de establecer algunas mejoras a algoritmos conocidos en este campo de la literatura, con el objetivo de brindar una referencia robusta y eficaz para aquellos que intenten construir un sistema de este tipo.

Primeramente se muestra una panorámica histórica de las investigaciones desarrolladas en el campo y se selecciona el algoritmo de Edmundson como el más adecuado para someterlo al experimento. Seguidamente se implementa el original y dos versiones del mismo: una variante basada en el Algoritmo de Hoey y otra en aportes propuestos por los autores de este trabajo.

Finalmente se evalúan los algoritmos implementados mediante técnicas automáticas, estadísticas y de inferencia, asegurando así un resultado más confiable.

PALABRAS CLAVE

Algoritmo Edmundson de Sumarización basado en Extracción

TABLA DE CONTENIDOS

INTRODUCCIÓN

La época actual, también denominada “Sociedad de la Información”, se ha caracterizado por un crecimiento vertiginoso de los avances tecnológicos. El surgimiento y desarrollo de la World Wide Web y de los servicios de información digitales ha provocado que el volumen de información disponible sea monstruoso, y de muy diversas naturalezas: gráficos, imágenes, videos, etc. Sin embargo, sigue siendo el texto el elemento fundamental. En este marco de sobrecarga de información, parece conveniente el estudio de técnicas que ayuden a los usuarios a organizar, buscar y comprender la información. Sin duda, los sistemas de generación automática de resúmenes de texto pueden jugar un papel importante en todas esas tareas.

La Universidad no es una excepción; la misma cuenta con un gran cúmulo de información en forma de texto, y los usuarios no tienen a su alcance con una herramienta personalizada de acuerdo a sus necesidades que les posibilite solucionar lo anterior; por otro lado, en idioma español se cuenta con un ínfimo material de estudio sobre este campo. La anterior **situación problemática** conlleva al siguiente **problema científico**: el insuficiente estudio crítico existente en la Universidad sobre los algoritmos de sumarización dificulta la implementación de una herramienta capaz de realizar resúmenes de textos. El **objeto de estudio** de esta investigación es la Minería de Texto. Para dar solución al problema científico, se propone, como **objetivo general,** estudiar los algoritmos de generación automática de textos. Teniendo en cuenta los siguientes **objetivos específicos**:

1.- Seleccionar un algoritmo de generación automática de textos existente en la literatura.

2- Proponer dos variantes a dicho algoritmo, una de las cuales suponga una propuesta auténtica por parte de los autores de la presente investigación.

3.- Someter a análisis los resúmenes generados por dichos algoritmos.

El **campo de acción** son los algoritmos de construcción de resúmenes de documentos.

La **hipótesis** de partida de esta tesis es que la realización de un estudio crítico de los algoritmos de sumarización facilite el desarrollo en la Universidad de herramientas con funcionalidades para entregar al usuario resúmenes personalizados de grandes volúmenes de texto.

Las **tareas de investigación** del presente trabajo son:

1.- Analizar el estado del arte de los algoritmos de construcción de resúmenes de documentos así como de sus técnicas de evaluación.

2.- Diseñar la implementación de los algoritmos propuestos.

3.- Aplicar técnicas de evaluación a los extractos obtenidos.

4- Comparar los resultados de la evaluación.

Para desarrollar nuestra investigación utilizamos los siguientes **métodos de la investigación**: *Método hipotético-deductivo* para la elaboración de la hipótesis central de la investigación y para proponer nuevas líneas de trabajo a partir de los resultados parciales. *Método histórico-lógico* y *el dialéctico* para el estudio crítico del los trabajos anteriores, y para utilizar estos como punto de referencia y comparación de los resultados alcanzados. *Método experimental* para comprobar la utilidad de los resultados obtenidos a partir del modelo definido.

La tesis está **estructurada** en una introducción y tres capítulos. El primer capítulo está dirigido a dar una panorámica de la construcción automática de resúmenes de texto. Para ello se darán a conocer las nociones básicas relacionadas con los resúmenes, las técnicas de generación de resúmenes y las técnicas para su evaluación. En el segundo capítulo se presentan primeramente los algoritmos estudiados y, teniendo en cuenta las fases en las que se divide el proceso de construcción de resúmenes, se exponen además los algoritmos propuestos. El capítulo tres describe la técnica de evaluación escogida y muestra los resultados de la evaluación de los extractos obtenidos por los algoritmos implementados. Finalmente se presentan las conclusiones de nuestro trabajo con los resultados alcanzados.

CAPÍTULO 1: ESTADO DEL ARTE Y NOCIONES BÁSICAS

Un *resumen* es la condensación de los conceptos principales del contenido del texto al que hace referencia (Burgos, 1994). Luego, puede definirse un *resumen automático* como el conjunto de técnicas que producen a partir de un texto de entrada un documento de salida de menor extensión pero que aún contiene los puntos más relevantes del original. A continuación se explican los factores que intervienen en la generación automática de los resúmenes.

1.1. Factores que afectan la construcción automática de resúmenes

Según Mani y Sparck-Jones (1993), existen tres factores que afectan directamente el proceso de la generación de resúmenes y a los que es necesario tener en cuenta: el factor de *entrada* (la fuente de información), *propósito* (la aplicación o el usuario al cual va dirigido el resumen), y *salida* (la forma en que se presentará el contenido). Cada uno de estos factores está integrado por otros elementos específicos, como se muestra a continuación en la figura 1.1.

Fig. 1.1. Factores que intervienen en el proceso de generación de resúmenes.

Los factores de *entrada* pueden clasificarse, a su vez, en *forma* del documento, *especificidad* del contenido y *alcance.* Respecto a la *forma*, deben tenerse en cuenta la *estructura, escala, medio* y *género* del documento.

La *estructura* hace referencia tanto a la organización externa del documento (en capítulos o secciones) como a la propia estructura del discurso. La *estructura* es importante porque puede ayudar a identificar los temas o aspectos de un tema que trata un documento. El resumen puede respetar la estructura del texto original o centrarse en una parte de él. La *escala* alude al tamaño del documento fuente. Los resúmenes de un artículo de periódico y un libro requerirán distintos grados de reducción, pero también las necesidades de transformar el contenido son distintas. Las posibilidades en el resumen del libro son mucho mayores que en la noticia del periódico. Por *medio* entendemos tanto el idioma en que está escrito el texto como la jerga empleada en el mismo. Aunque el resumen puede no preservar ninguna de las dos cosas. Puede requerirse un resumen en un idioma distinto al del texto original o en el que se hayan sustituido expresiones demasiado específicas del dominio (por ejemplo, términos jurídicos o económicos). El *género* del texto fuente, es decir, su estilo literario o su tipo de contenido, también afecta al proceso de generación de un resumen. Tratándose de un artículo científico, su resumen podría hacer referencia a elementos como el propósito del trabajo, el procedimiento de investigación seguido y las conclusiones a las que se llegan. En un resumen de una noticia, el objetivo podría ser preservar la información sobre el que, como, cuando, donde y por qué de lo sucedido.

La *especificidad* hace referencia al nivel de especialización del tema tratado en el documento fuente o a la presencia en el mismo de referencias locales. Puede presumirse que el lector del resumen dispone de los conocimientos adecuados o, en caso contrario, pueden evitarse las cuestiones demasiado específicas, de forma que el rango de posibles lectores sea el más amplio posible.

El factor *alcance* alude al número de fuentes utilizadas para confeccionar el resumen. Cuando intervienen varias fuentes, deben tenerse en cuenta aspectos como la redundancia de información o los cambios en la misma debido al paso del tiempo.

Los *factores de propósito* son los más importantes y su participación en la estrategia del proceso de generación de resúmenes, crítica. Los factores de *propósito* pueden clasificarse en *situación*, *audiencia* y

función. La *situación* hace referencia al contexto en que se va a usar el resumen. La *audiencia* alude a los conocimientos del dominio, la competencia lingüística o los intereses de los lectores a los que está destinado el resumen; de esta forma buscamos la posibilidad de que un texto pueda adaptarse a distintos intereses. En cuanto a la *función*, un resumen puede tener, entre otros, los siguientes destinos: sustituir al texto fuente, ayudar a recordar el contenido de un texto ya leído o facilitar la decisión sobre el interés de la fuente, proporcionando una pista del tema tratado en la misma.

En cuanto a los *factores de salida*, se debe tener en cuenta el *formato* del resumen y la *extensión*. Respecto al *formato*, el resumen puede presentarse como texto continuo, como en el caso de los resúmenes de artículos científicos, u organizado en secciones, que se corresponderán con la estructura del documento. Dependiendo de la *extensión* máxima requerida, el resumen puede intentar recoger todos los aspectos importantes del texto fuente o centrarse solo en uno de ellos.

Antes de entrar en un análisis más profundo de la generación de resúmenes es necesario presentar algunas de las clasificaciones habituales que se realizan sobre los mismos. Estas clasificaciones se llevan a cabo en función de algunos de los factores que intervienen en el proceso de generación y que fueron mencionados anteriormente.

1.2. Clasificación de los extractos generados

Atendiendo a su función

Según su *función*, los resúmenes se clasifican en: *indicativos e informativos* (Borko y Bernier, 1975). Los denominados *indicativos* tienen como propósito anticipar al lector del contenido del texto y ayudarle a decidir sobre el interés del documento fuente. Son usados para condensar textos de poca estructuración y gran extensión tales como editoriales, ensayos, libros, etc. Por el contrario, los *informativos* pretenden sustituir al texto completo incorporando toda la información trascendente; a partir de ellos se puede reconstruir por completo el texto de la fuente. Es decir, los informativos pueden entenderse como un subtipo de los indicativos.

Atendiendo a su audiencia

Otra clasificación habitual de los resúmenes es la que los agrupa, en función de la *audiencia*, en resúmenes *genéricos* y *adaptados al usuario*. Los resúmenes *genéricos* intentan recoger los temas principales de un documento, y están destinados a una amplia comunidad de lectores, mientras que los *adaptados al usuario* confeccionan el resumen de acuerdo a los intereses del usuario al que va dirigido, esto es, sus conocimientos previos, sus ámbitos de interés o sus necesidades de información. Constituyen ejemplos de resúmenes enfocados a un usuario aquellos que son mostrados por los buscadores de Internet para cada documento recuperado.

Atendiendo a su alcance

Finalmente, atendiendo al alcance, se distingue entre resúmenes *monodocumento y multidocumento*, en función de si el resumen se refiere a un único texto fuente o a varios. En el caso de los resúmenes *multidocumento* la concepción de resumen se extiende, de una única fuente de información, a un conjunto de documentos relacionados semánticamente. Para la generación del resumen debe tenerse en cuenta la información común que comparte el conjunto de documentos, de forma tal que sea posible eliminar la redundancia, así como que se mencionen los aspectos significativos que puede aportar cada documento. Sin embargo, este método presenta problemas adicionales, como puede ser el hecho de que los documentos seleccionados no guarden suficiente similitud semántica entre ellos, sin mencionar el hecho de que, al tratarse de un nivel de información mucho mayor, se multiplican los problemas típicos existentes en los métodos para crear resúmenes *monodocumentos*. El resultado puede ser un resumen de difícil lectura. El estudio de este trabajo se limitará al campo de las técnicas para generar resúmenes *monodocumento*.

1.3. Clasificación de las técnicas de obtención de extractos

Durante todos estos años se han desarrollado un número importante de sistemas de generación automática de resúmenes de texto que, en función del nivel del análisis lingüístico realizado sobre la fuente, podemos clasificar en dos categorías: *extracción* y *abstracción* (Maña, 2003).

Los sistemas basados en técnicas de *extracción* se caracterizan por un análisis superficial del texto fuente, no profundizando más allá del nivel sintáctico. El resultado es un resumen generado a partir de la extracción de elementos significativos del texto original. Estos elementos pueden ser, por ejemplo, palabras, oraciones o párrafos. Generalmente es preferible seleccionar oraciones, puesto que la selección de palabras produce resúmenes bastante incoherentes, y la selección de párrafos ocasiona muchos problemas con la tasa de compresión. Los problemas de incoherencia, que probablemente se producirán al separar estos elementos de su contexto, pueden mitigarse mediante un proceso de revisión del texto seleccionado.

En un caso opuesto estarían sistemas de resumen automático que construyen un extracto completamente nuevo que recoge las ideas principales del documento original sin incluir necesariamente fragmentos literales del mismo. En otras palabras, estos sistemas (de resumen *abstractivo*) trabajan de manera similar a un ser humano. Las fases de análisis y generación requieren gran cantidad de conocimiento del dominio, por lo que este tipo de sistemas solo es aplicable a ámbitos muy concretos y enteramente conocidos (Spärck-Jones, 1999) y, de hecho, no es previsible la existencia de sistemas prácticos de resumen por abstracción a corto plazo (Hovy, 1999).

Ante la variedad de tipos y dominios de los documentos disponibles, las técnicas de selección y extracción de frases resultan muy atractivas por su independencia del dominio y del idioma (Mitra, Singhal y Buckley, 1997). Teniendo en cuenta estas razones, el presente trabajo se inclina por el criterio de emplear técnicas de extracción para obtener resúmenes; es por ello que en el *Capítulo 2* los algoritmos sometidos a estudio son de este tipo. A continuación damos una panorámica histórica de las investigaciones desarrolladas en este campo.

1.4. Estado del arte de los algoritmos basados en extracción

Los orígenes de este campo de estudio pueden remontarse a los trabajos de Luhn (1958) y Edmundson (1969) que desarrollaron los primeros sistemas de “extracción de resúmenes”. Luhn fue el primero en proponer un método estadístico para extraer las oraciones más significativas de un texto y construir un resumen del mismo. Según su propuesta, debía determinarse, en primer lugar, la importancia de cada una de las palabras (luego de haber eliminado las stop words, o palabras vacías). Luhn se basaba en la

suposición de que las más frecuentes indicarían una mayor relevancia, puesto que las oraciones en las que estuvieran incluidas estas palabras tendrían mayor probabilidad de contener información relevante para el resumen. Posteriormente se asignaría a cada oración un peso en función del número de palabras relevantes que incluyese, el peso de las mismas y la distancia entre ellas dentro de la oración. Una vez puntuadas todas las oraciones del documento se procedería a ordenarlas de mayor a menor, y a seleccionar un subconjunto de las más significativas como resumen del texto original. Este método de selección de oraciones es bastante acertado pese a ser un método simplista; como demostrara (Kupiec, 1995), aproximadamente el 80% de las oraciones en resúmenes creados por humanos están copiadas tal cual o con pequeñas modificaciones a partir del texto original. Sin embargo, el método de Luhn, como único medidor de importancia, es insuficiente.

Edmundson (1969), en cambio, emplea cuatro métodos distintos para asignar pesos a las oraciones del documento: una colección de frases que proporcionan pistas sobre la relevancia de las oraciones, la utilización de palabras frecuentes como indicadores de relevancia (semejante al método de Luhn), el uso de palabras del título del documento como indicadores positivos, y heurísticos basados en la posición de las oraciones en el texto. Cada una de estas métricas contribuía a la asignación final del peso de las oraciones. El trabajo de Edmundson definió un paradigma de solución que ha sido incluido en buena parte de las investigaciones realizadas hasta el momento en el campo de la generación de resúmenes mediante extracción. Los trabajos posteriores han incidido, fundamentalmente, en nuevas métricas o en definiciones más sofisticadas o especialmente adaptadas a las características del corpus. Edmundson, además, fue uno de los primeros en señalar la necesidad de evaluar los sistemas de extracción de resúmenes comparando sus resultados con extractos producidos por evaluadores humanos.

Durante la década de 1970 y primeros años de 1980 la investigación en sistemas de resumen automático disminuyó considerablemente (Spärck-Jones, 1999) y no volvería a convertirse en un campo activo hasta principios de la década del 1990, potenciada por el surgimiento de Internet y de inmensos volúmenes de textos.

Hoey (1991) demostró que la repetición léxica es una relación que sirve para ilustrar de manera bastante realista cuán conectados están los elementos de un texto. En este algoritmo se buscan los vínculos y

lazos que existan entre las oraciones de un documento. Existirá un vínculo entre cada par de oraciones por cada término que tengan en común.

Salton y Singhal (1994) señalan la necesidad de identificar, en primer lugar, los distintos temas tratados en un documento, así como los párrafos del texto que se refieren al mismo asunto para, posteriormente, emplear una selección de párrafos como resumen del documento. Básicamente se trata de realizar una clasificación automática de párrafos controlada mediante un umbral de similitud que será inversamente proporcional al número de temas que se deseen "descubrir" en el documento. Sin embargo, como habíamos mencionado antes en la *Sección 1.3*, la extracción de párrafos íntegros del texto fuente crea serios problemas con la tasa de compresión del extracto deseado.

En 1996, estos mismos investigadores ahondan en el tema anterior, pero tratando de avanzar hacia la identificación de pasajes, o sea, "fragmentos del texto que exhiben consistencia interna y que pueden distinguirse del resto del texto circundante" (Salton, 1996).

Mitra, Singhal y Buckley (1997) evalúan el anterior sistema comparándolo con resúmenes creados manualmente. Las conclusiones a las cuales llegan en su investigación aún están vigentes. Según ellos, los primeros párrafos de un texto resultan casi tan efectivos como un resumen obtenido mediante métodos extractivos "inteligentes". Esto puede ser cierto, al menos en el caso de los documentos que normalmente se emplean para este tipo de experimentos (artículos periodísticos, técnicos o científicos), ya que están estructurados de tal modo que los primeros párrafos son una suerte de extracto del contenido. Según Zechner (1997) y este trabajo se acoge a su criterio-, aunque tal vez los resultados de las anteriores técnicas automáticas fueran menos legibles, sí eran mejores en cuanto a precisión y exhaustividad.

En el trabajo de Kupiec, Pedersen y Chen (1995) se emplea un *corpus* de documentos y resúmenes creados manualmente como datos de entrenamiento para un clasificador bayesiano que debía determinar cuáles oraciones de un documento deberían formar parte de un resumen y cuáles no. El sistema propuesto determinaba para cada oración la probabilidad de pertenencia al resumen final y extraía las más probables. Al reducir los documentos a un 25% del tamaño original, seleccionaba un 84% de las oraciones elegidas por los expertos humanos y para resúmenes más cortos, y resultaba sustancialmente superior al estudio anterior, consistente en presentar el inicio del documento.

Kraaij, Spitters y van der Heijden (2001) y Kraaij, Spitters y Hulth (2002) también han utilizado clasificadores bayesianos para la extracción de resúmenes. Por su parte, Conroy (2001) y Dunlavy (2003) han implementado sistemas extractivos mediante modelos de Markov que hacen menos suposiciones que los clasificadores bayesianos sobre la independencia entre elementos y Fuentes (2003) o Doran (2004) utilizan árboles de decisión. Alfonseca y Rodríguez (2003), Jaoua y Ben Hamadou (2003) y Alfonseca, Guirao y Moreno Sandoval (2004) han utilizado algoritmos genéticos para la selección de las oraciones.

Erkan y Radev (2004) han desarrollado una nueva medida de "centralidad" para las oraciones, denominada *LexPageRank*, basada en la idea de "prestigio" de las redes sociales, y análoga al *PageRank* (Page, 1998) de *Google*. El valor de *LexPageRank* para una oración *S* se define como la suma de los valores *LexPageRank* de aquellas oraciones similares a *S*, donde la similitud se determina mediante la función del coseno. La última versión resultó uno de los mejores participantes en cuatro de las cinco tareas de *DUC 2004*. Por su parte, Vanderwende, Banko y Menezes (2004) utilizan *PageRank* para determinar qué elementos de un documento son los más relevantes aunque sus resúmenes son generados y no construidos a partir de oraciones extraídas literalmente de los documentos. Ambos trabajos guardan cierta relación con los desarrollados por Salton (1996) que también emplearon grafos para analizar los contenidos de un texto.

Así pues, la mejora de los métodos de resumen extractivo es un campo de investigación activo debido, por un lado, a las menores exigencias de partida (requieren un conocimiento lingüístico nulo o mínimo) y, por otro, al hecho de que la mayor parte de documentos con los que tratan los usuarios en la actualidad no tienen una estructura fija ni pertenecen a un dominio concreto. En semejante escenario la sencillez, flexibilidad y robustez de los métodos de extracción son aspectos valiosos.

1.5. Características generales de los algoritmos basados en técnicas de extracción

Los sistemas basados en extracción siguen una arquitectura estructurada en dos procesos: *análisis* y *síntesis* (Hahn y Mani, 2000). En la fase de análisis se identifican los segmentos de texto que contienen la información más significativa. Durante esta fase se aplica un conjunto de heurísticas (métricas) a cada una de las unidades a evaluar. Las heurísticas que se suelen utilizar en la función de peso pueden clasificarse

en *posicionales*, si tienen en cuenta la posición que ocupa la frase, *lingüísticas*, si buscan ciertos patrones de expresiones indicativas, o *estadísticas*, si incluyen frecuencias de aparición de ciertas palabras. En la fase de síntesis se extraen las frases con mejor puntuación de acuerdo a la tasa de compresión deseada.

Los algoritmos basados en extracción se dividen en tres grupos, según (Maña, 2003): *algoritmos supervisados, semi-supervisados* y *no supervisados.*

1.5.1. Extracción basada en aprendizaje

Cuando se cuenta con un corpus de documentos junto con sus respectivos extractos “ideales” (generados por humanos) es posible aplicar técnicas de aprendizaje automático supervisado para asignar los pesos asociados a cada una de las métricas del algoritmo. Esta técnica es independiente del dominio de los documentos, sin embargo, convierte al sistema generador de resúmenes en un sistema dependiente por completo del corpus. Su aplicación a un corpus de un dominio diferente requerirá que se disponga de otra colección de pares de texto fuente-extracto “ideal” con la cual entrenar al sistema.

Está considerado a (Kupiec, 1995) como el paradigma de solución para este tipo de sistemas. Los elementos del texto fuente a extraer (generalmente oraciones) se analizan en términos de las características que son de interés. Así, se obtiene un vector que representa a cada oración del texto fuente. Dichos vectores se puntúan según sea su semejanza con el contenido del resumen del documento. El algoritmo de aprendizaje utiliza los vectores puntuados para obtener un clasificador que puede emplearse para averiguar si cada una de las oraciones que componen un documento debe formar parte o no del extracto, tal y como puede apreciarse en la *Fig. 1.2*. La tasa de compresión limitará el número de oraciones que se seleccionan a las mejor valoradas por el clasificador.

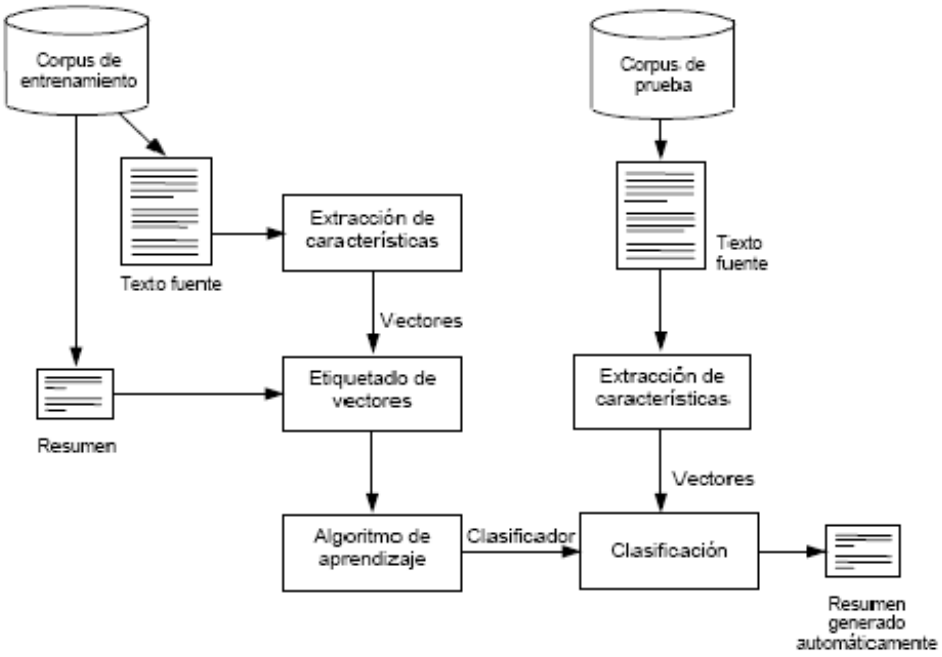

Fig. 1.2. Representación de un algoritmo de extracción basado en aprendizaje según (Kupiec, 1995)

El clasificador más utilizado ha sido denominado “Bayes ingenuo” o Naive Bayes (Maña, 2003), el cual puede ser expresado mediante la siguiente expresión:

$$P(s \in R \mid C_1 \ldots C_n) = \frac{\prod_{i=1}^{n} P(C_i \mid s \in R) P(s \in R)}{\prod_{i=1}^{n} P(C_i)}$$

Donde P ($s \in R$) es la probabilidad de que una oración s del texto fuente esté incluida en el resumen R, y está dada por la tasa de compresión. P($C_i \mid s \in R$) es la probabilidad de aparición de la característica

C_i, una vez conocido que la oración s está incluida en el resumen. Por último, P (C_i) es la probabilidad de que la característica C_i se dé en las frases del corpus fuente.

El principal inconveniente de la aproximación basada en aprendizaje está en la dificultad que puede suponer la obtención de una colección de documentos de entrenamiento de tamaño suficiente (Maña, 2003). El elevado coste que esto puede significar impone, en la práctica, severas limitaciones a la aplicación de este tipo de técnicas. Por esa razón, la utilización de procedimientos de aprendizaje *no supervisado* o *semi-supervisado* resulta de gran interés.

Como ejemplo podemos citar los trabajos de Hoey (1991), Mani y Bloedorn (1998), Chiang y Yang (2000), entre otros. Estos investigadores han utilizado características de *cohesión* y *coherencia* del texto fuente en la representación de las oraciones.

Por *cohesión* del texto se entiende las relaciones semánticas que se establecen entre las palabras, los sentidos de las palabras o las expresiones que remiten a otros elementos del texto (Halliday y Hasan, 1996). Estas relaciones determinan el grado de conexión del texto, y son de dos clases: gramaticales y léxicas. Entre las primeras están las relaciones anafóricas y las de elipsis, mientras que las relaciones de sinonimia, hiponimia y repetición, entre otras, se encuentran en la segunda.

La *coherencia* representa las relaciones que se establecen a más alto nivel entre las cláusulas o frases de un texto (Mann y Thompson, 1988). Estos vínculos determinan la estructura argumentativa del texto. En este caso la estrategia es que los principales núcleos del discurso sean usados para la composición del extracto.

Los sistemas que emplean el nivel de discurso del texto como técnica, construyen un grafo que luego es usado para clasificar y extraer los elementos que formarán parte del resumen.

1.6. Problemas de las técnicas de extracción

El principal inconveniente que se presenta cuando se utilizan técnicas basadas en extracción es que, al extraer elementos aislados del texto fuente sin considerar las relaciones existentes entre ellos, se corre el

riesgo de provocar la aparición de *inconsistencia* en los resúmenes. Existen diversas causas. Según (Paice, 1990) puede deberse a la presencia de referencias anafóricas a elementos del discurso que no se han incluido en el resumen o, simplemente, a que las oraciones extraídas no aparecían de manera consecutiva en el texto original. El mismo autor añade que también pueden resultar *desequilibrados*, por no considerar todos los aspectos importantes que se tratan en el documento o por no reflejar su organización estructural.

Sin embargo, experimentos centrados en comprensión de lectura muestran que no hay diferencias significativas entre el texto fuente y los resúmenes construidos mediante extracción. En (Morris, 1992) se demuestra que la falta de fluidez o cohesión de los resúmenes generados mediante extracción no afecta a la comprensión de la lectura.

Inconsistencia

Como ya habíamos mencionado, los problemas de inconsistencia aparecen cuando el resumen contiene oraciones que no pueden comprenderse debido a que incluyen referencias a elementos que no aparecen en el mismo. En (Maña, 2003) se analizan los dos tipos de anáforas que pueden presentarse cuando se genera un resumen mediante extracción. El primer caso sucede cuando la referencia anafórica es un sujeto en el que aparece un demostrativo o artículo determinado. Este caso se caracteriza porque puede hacer referencia a información que aparezca lejana respecto a la frase donde se encuentra la anáfora. Este es, por tanto, un tipo de anáfora difícil de resolver de manera automática. El segundo caso ocurre cuando se incluyen, de manera consecutiva, dos oraciones que no son adyacentes. Generalmente, la información a la cual hace referencia el pronombre sujeto anafórico se encuentra en la oración precedente a la que contiene la anáfora en el texto original.

Desequilibrio

Según (Paice, 1990) el concepto de desequilibrio hace referencia a la ausencia en el resumen de:

- Alguno de los temas importantes que aborda el documento original. Muchos documentos no tratan un único tema, sino un conjunto de temas relacionados. De esta manera, un resumen realizado mediante

extracción que no tenga en cuenta la distribución temática del documento fuente puede estar omitiendo aspectos relevantes del contenido.

- Alguno de los aspectos referentes a su organización estructural. Así, por ejemplo, un resumen equilibrado sobre un artículo de investigación debería proporcionar información sobre el “Propósito del trabajo", “Procedimientos de investigación" y “Conclusiones".

A pesar de todo lo anterior, existen técnicas que permiten minimizar estos problemas. A continuación nos ocuparemos de las mismas.

1.7. Soluciones a los problemas de las técnicas basadas en extracción

Las técnicas de revisión de los extractos generados que se mencionarán a continuación inciden en la fase de síntesis, permitiendo mejorar la calidad de los resúmenes. Es válido aclarar que el objetivo de las mismas no es realizar abstracción, puesto que no se genera nuevo texto, tan sólo es revisada la selección de las oraciones, añadiendo criterios adicionales al peso alcanzado por cada oración.

Inconsistencia

(Rush, 1971) propone una de las primeras soluciones al problema de la inconsistencia de los resúmenes. En el caso de que sea detectada una referencia anafórica en una de las oraciones que debe ser extraída para conformar el resumen, debe incluirse también la oración que le precede (donde supuestamente debe resolverse la anáfora), aunque esta oración no sea una de las seleccionadas. El proceso se repite con cada oración precedente que se añade. Sin embargo, si una oración requiere que se añadan más de tres oraciones precedentes, no se incluirá en el resumen final, con el objetivo de evitar que se seleccionen pasajes innecesariamente extensos.

En cambio, (Brandow, 1995; Johnson, 1993) proponen sencillamente eliminar las oraciones que contengan palabras o expresiones anafóricas.

Por su parte, (Paice, 1990) propone una solución análoga a la primera, aunque más compleja. Mediante un conjunto de reglas no sólo son localizadas las anáforas, sino también es decidido si el antecedente se encuentra en la misma oración o en otra precedente.

El principal inconveniente de la inclusión de nuevas oraciones en el documento está en la posible pérdida de significación semántica del resumen. Si se incluyen estas nuevas oraciones y se desea mantener constante la tasa de compresión del resumen es necesario excluir otras oraciones que podrían tener un mayor peso. El resultado será un resumen con un contenido menos significativo (Maña, 2003).

Con el objetivo de evitar lo anterior, en (Nanba y Okumura, 2000) se opta por eliminar las anáforas que aparecen al comienzo de una oración cuando la oración que la antecede en el texto fuente no ha sido seleccionada para formar parte del resumen.

En otros trabajos, la unidad de extracción seleccionada es el párrafo (Salton, 1994; Abracos y Lopes, 1997; Mitra, 1997; Salton, 1997; Fukumoto y Suzuki, 2000). El hecho de que un párrafo proporciona más contexto hace pensar que los problemas de cohesión y legibilidad serán menores que cuando se extraen oraciones. Sin embargo, como ya se ha mencionado anteriormente, la extracción de párrafos puede resultar inadecuada cuando se desean confeccionar resúmenes con tasas de compresión bajas (la elección de un único párrafo podría exceder la longitud deseada). Por otra parte, el párrafo puede contener oraciones de escaso peso, por lo que, un resumen de la misma longitud generado mediante extracción de frases podría tener una mayor significación semántica (Maña, 2003).

Desequilibrio

Ya en (Luhn, 1958) este investigador se percata del problema del desequilibrio de los resúmenes generados mediante extracción. “El poder de resolución de las palabras significativas decrece cuando aumenta el número total de palabras del documento”. Luhn propone que el documento sea dividido, de manera que se tengan en cuenta los diferentes temas que se tratan. Así, al generar el extracto puede tomarse información de cada una de las partes, de forma que quede reflejada la estructura del texto fuente.

Incluso en los textos más sencillos es posible encontrar algún tipo de organización estructural de los temas. En (Paice, 1990) se propone utilizar tanto la división en secciones del documento como la información de orientación que incluye el autor. Esta última está normalmente formada por una o varias oraciones que informan al usuario sobre la manera en que se organiza el documento. El problema es que este tipo de información es típica de documentos científicos, pero difícilmente aparece en otros dominios.

En (Angheluta, 2002), el algoritmo propuesto permite construir una estructura jerárquica de los temas y asociar palabras clave a dichos temas, al estilo del índice de un libro. La tasa de compresión deseada para el resumen determina el nivel de la jerarquía del índice a considerar.

Existen, además, otros problemas aparte de las referencias anafóricas y el desequilibrio. Por ejemplo, aquellas oraciones que están relacionadas mediante una conjunción. En (Nanba y Okumura, 2000) se afronta utilizando una lista predefinida de 52 expresiones conjuntivas y realizando un análisis parcial de la estructura retórica de las tres oraciones que circundan a la que incluye la conjunción. La conjunción es eliminada si la oración relacionada con la conjunción no se ha incluido en el resumen. El trabajo de Nanba y Okumura también se ocupa de los sujetos elípticos. Si una oración seleccionada para el resumen omite el sujeto, las reglas de revisión lo obtienen de las oraciones más cercanas que la preceden y cuyo sujeto está presente en el texto fuente. También puede mencionarse la eliminación de expresiones entre paréntesis, problema que es tratado en (Nomoto, 2001).

No obstante, aunque las técnicas de extracción pueden producir resúmenes inconsistentes que dificulten la lectura del mismo, esta investigación se rige por el criterio de que los posibles problemas de legibilidad no afectarán a la utilidad del resumen. Esta hipótesis se ve respaldada por trabajos como (Morris, 1992), en el que se concluye que los usuarios muestran un nivel de comprensión de lectura de los resúmenes generados por extracción similar al de resúmenes confeccionados manualmente.

1.8. Técnicas de evaluación de resúmenes generados de forma automática

Desde el trabajo de Edmundson (1969) se advierte la importancia que conlleva la evaluación de los resúmenes generados automáticamente. Los algoritmos requieren ser validados mediante alguna forma que evalúe la calidad, coherencia, e información relevante ofrecida al usuario. Incluso los trabajos que

establezcan hipótesis o técnicas teóricas necesitan ser evaluados. Las técnicas de evaluación permiten, además, establecer comparaciones entre diferentes métodos, como es el caso de la presente investigación. Debido a la importancia asociada a esta tarea, seguidamente hacemos un estudio de las principales técnicas empleadas para evaluar los resúmenes automáticos, aunque es válido aclarar primeramente que, a pesar de todos los esfuerzos que se han llevado a cabo en el área, la comunidad científica no ha llegado, hasta el momento, a un acuerdo pleno sobre cuál es el método de evaluación óptimo. Según (Mani, 2001), las causas de ello deben buscarse en:

- La dificultad existente para llegar a un acuerdo sobre la idoneidad de un resumen automático. El hecho de que no coincida con un resumen confeccionado por una persona no significa que sea incorrecto. Por lo tanto, la evaluación mediante comparación con un resumen “ideal" es de difícil aplicación. Según estudios realizados por (Salton, 1997) es escaso el grado de coincidencia entre jueces humanos.

- La necesidad de utilizar personas para juzgar los resúmenes encarece el proceso de evaluación y hace que sea difícil de repetir. Debido a esto, son preferibles las técnicas que permiten resolverlo de forma automática.

Según (Sparck-Jones y Galliers, 1996) las técnicas de evaluación pueden clasificarse en *intrínsecas* y *extrínsecas.* Los métodos *intrínsecos* analizan directamente el resumen, utilizando algún criterio que permita medir su adecuación o fiabilidad. Los métodos extrínsecos juzgan la calidad del extracto en función de su utilidad para realizar alguna otra tarea.

1.8.1. Método intrínseco

Como mencionamos anteriormente los métodos intrínsecos evalúan a los resúmenes como entes individuales, generalmente comparándolos con un resumen de referencia. (Baldwin, 2000) establece que estos métodos pueden tener en cuenta tanto criterios de calidad y corrección gramatical o cohesión del texto que constituye el resumen, como criterios de cobertura informativa. Los jueces humanos son imprescindibles para este tipo de evaluación, tanto para confeccionar un resumen inicial con el cual comparar el automático, como para opinar sobre la legibilidad de los resúmenes o indicar el grado de

relevancia de la información. Según (Maña, 2003) esto conlleva toda una serie de problemáticas asociadas, puesto que para lograr obtener conclusiones relevantes, los experimentos deben llevarse a cabo sobre colecciones de cierto tamaño, lo que multiplica el coste del proceso. Además, el posible desacuerdo que se puede producir entre los jueces puede conllevar a anular los experimentos.

Criterio de calidad

La legibilidad de los resúmenes es uno de los aspectos que inciden en su calidad. En este caso se tienen en cuenta aspectos como la corrección gramatical, la cohesión (secuencia de frases que se produce de una forma natural) la organización global de las ideas, errores en el uso de mayúsculas, orden incorrecto de las palabras, falta de concordancia en el número entre sujeto y verbo, falta de componentes importantes de la oración que afecten la claridad de la misma, fragmentos no relacionados unidos en la misma oración, omisión o uso incorrecto de artículos, pronombres con antecedentes incorrectos u omitidos, entre otros (Maña, 2003).

Estos criterios garantizan una buena legibilidad del extracto, sin embargo, no son suficientes para asegurar que el resumen sea bueno. Se necesitan otros criterios que sean capaces de medir su utilidad o la relevancia de la información que contienen.

Criterio de cobertura informativa

El método más empleado para medir la relevancia de la información consiste en comparar el resumen generado automáticamente con uno confeccionado manualmente y considerado “ideal". Las métricas tradicionales de IR: *precisión* y *recall* (*cobertura*) (Hovy, 1999), han sido utilizadas con frecuencia para medir la eficacia de los resúmenes automáticos en comparación con otros confeccionados manualmente y tomados como referencia.

Otros criterios

Otro medio de evaluar un resumen comparando su contenido con el del resumen "ideal" o con el de su texto fuente es la comparación del contenido entre los dos textos usando una medida de solapamiento de vocabularios, como el *Coeficiente de Dice* o *La Medida del Coseno* (Salton y McGill, 1983).

También puede emplearse, para evaluar la calidad de un extracto, un método que considere los *n-gramas* (secuencia de *n* palabras consecutivas de un texto) del resumen en lugar de las oraciones. ROUGE-n (Lin y Hovy, 2003) es una medida basada en la ocurrencia de los *n-gramas* que evalúa la relevancia de un resumen *R* a partir de un conjunto de resúmenes de referencia *C* y se define como:

$$\frac{\sum_{t \in C} \sum_{g \in n-gramas(t)} mín\{cant(g,t), cant(g,r)\}}{\sum_{t \in C} \sum_{g \in n-gramas(t)} cant(g,t)}$$

donde *n-gramas* (*p*) es el conjunto de *n*-gramas del texto *p* y *cant* (*g*, *p*) denota la cantidad de veces que aparece el *n-grama g* en *p*.

1.8.2 Método extrínseco

Los métodos de evaluación extrínseca se basan en medir el efecto que tienen los resúmenes sobre la realización de alguna otra tarea. Generalmente exigen una activa participación de personas. Algunos experimentos evalúan la efectividad en la comprensión de la lectura (Mani, 2001), según el porcentaje de respuestas correctas que alcanza una persona en una prueba que le es realizada después de la lectura del extracto. Este método también puede ser utilizado para evaluar el contenido informativo del resumen. En otros métodos, como en (Mani et al 98), se estudia la utilidad de los resúmenes para lograr un ahorro de tiempo sin pérdida de efectividad en las decisiones de relevancia.

Según (Abreu, 2005), una medida de evaluación produce para un resumen un valor numérico que por sí solo no significa nada; pero dicho valor puede ser utilizado en conjunto con los valores obtenidos de medir

(con la misma medida de evaluación) otros resúmenes de la misma fuente para establecer un orden entre los distintos resúmenes.

Un algoritmo de construcción automática de resúmenes *A* puede ser evaluado a partir de una medida de evaluación de resúmenes *f* si se cuenta con una colección de textos fuentes *C* y con los recursos necesarios para evaluar el resumen de cada miembro de *C*. En principio el valor puede ser definido como $\sum_{t \in C} f\left(A(t), \eta_f(t)\right)/|C|$, donde $\eta_f(t)$ denota al conjunto de recursos necesarios para evaluar al resumen de *t* obtenido por *A* con la medida *f*; aunque el estadígrafo de tendencia central media (en este caso su estimador) muchas veces no ofrece una idea real de los datos.

Del mismo modo en que el conjunto de valores obtenidos al evaluar un conjunto de resúmenes de un texto fuente puede ser usado para establecer un orden entre los resúmenes, el conjunto de valores obtenidos al evaluar un conjunto de algoritmos de construcción automática de resúmenes en una colección de textos fuentes puede ser usado para establecer un orden entre los algoritmos.

Existen diferentes colecciones de documentos y resúmenes que se utilizan para realizar este tipo de evaluación. Los más importantes son: conferencias DUC, TIPSTER SUMMAC, Computation and Language y RST Discourse Treebank.

1.9 Conclusiones Parciales

En el presente capítulo se han abordado las nociones básicas sobre la construcción automática de resúmenes y se concluyó que la *extracción* sigue siendo la aproximación más fácil al problema de la generación automática de un resumen, y por ende la opción más acertada para cumplir el objetivo del presente trabajo, teniendo en cuenta las siguientes razones:

1-. No es necesario generar nuevo texto. El problema se reduce a la identificación de los elementos significativos del texto fuente, habitualmente oraciones, y a la selección de los mismos.

2-. Contiene un enfoque más atractivo, debido a su bajo coste. Puesto que no se pretende obtener una representación semántica del contenido del documento fuente ni generar texto, no se necesitan recursos adicionales con conocimientos del dominio o de carácter lingüístico.

3-. Por otra parte, este enfoque de análisis superficial resulta muy robusto. Su independencia del dominio, e incluso del género de los documentos, es muy alta, por lo que resulta fácilmente aplicable a contextos de propósito general.

Sin embargo, la extracción de elementos aislados puede provocar la aparición de inconsistencia y desequilibrio en los resúmenes. Aunque estos problemas no son desdeñables, sí pueden resultar tolerables en determinados contextos. En la presente tesis se ha decidido escoger, como algoritmo para someter a estudio, el algoritmo de Edmundson, teniendo en cuenta que se trata de uno de los algoritmos que constituye un pilar fundamental para la mayoría de los sistemas de generación automática de textos actuales.

En el capítulo, además, se presentaron las principales técnicas existentes para la evaluación de los extractos generados, sus ventajas y desventajas, inclinándonos por escoger técnicas de evaluación automáticas o que no necesiten de la intervención de jueces humanos, ya que encarece demasiado el proceso, e incluso se corre el riesgo de que se anulen los experimentos.

CAPÍTULO 2: DISEÑO E IMPLEMENTACIÓN

Tal y como había quedado plasmado en el capítulo anterior, la presente investigación va a desarrollar el estudio de uno de los algoritmos para la construcción de extractos de estrategia poco profunda y dos variantes del mismo; su funcionamiento se basa, básicamente, en que el tipo de elementos que son escogidos para conformar el resumen son oraciones. Este único elemento es suficiente, puesto que las oraciones son elementos lingüísticos que, por lo general, expresan proposiciones o ideas semánticamente completas (Abreu, 2005).

A continuación se presentan los algoritmos desarrollados así como las implementaciones de cada uno, teniendo en cuenta las distintas fases que caracterizan a dichos algoritmos. Primeramente se ofrece una descripción global y luego otra más detallada de cada módulo.

2.1. Algoritmo de Edmundson

Como ya habíamos visto anteriormente, los algoritmos basados en extracción presentan dos fases: *análisis* y *síntesis*. Primeramente, en la fase de análisis se mide la relevancia de las oraciones de acuerdo a una función de peso que otorga un valor numérico a cada oración. En la siguiente fase las oraciones con mejor puntuación son extraídas. En la *Fig. 2.1.* se muestra esta arquitectura, según (Hahn y Mani, 2000). De esta forma, el problema se resume a ordenar las oraciones del documento fuente de acuerdo a su relevancia.

Precisamente en esta filosofía se basa el algoritmo de Edmundson; este definió el área de trabajo para la mayoría de las aplicaciones en lo que a construcción de extractos se refiere hoy en día.

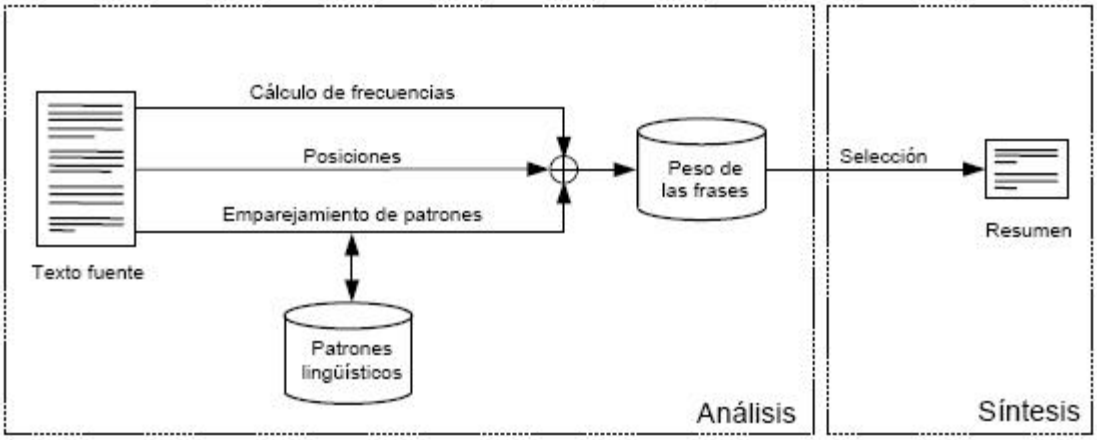

Fig. 2.1. Arquitectura de un sistema de generación de textos basado en técnicas de extracción.

Edmundson propone una función de peso a partir de la combinación lineal de diferentes métricas. De esta forma, el peso de una oración *f* se calcula usando la siguiente expresión:

$peso\ (f) = \alpha C\ (f) + \beta K\ (f) + \gamma L\ (f) + \delta T\ (f)$

donde *C (f), K (f), T (f) y L (f)* representan los valores de las métricas *frases pista*, *palabras clave, palabras título* y *localización, y* α, β, γ, δ son los pesos asociados a cada una de las métricas, que en el trabajo de Edmundson se calculan manualmente.

A continuación explicamos cada una de las métricas mencionadas.

Frases pista

El empleo de esta métrica se basa en la hipótesis de que ciertas frases en una oración, como por ejemplo “This work” o “This paper” pueden indicar el grado de relevancia de la misma, puesto que la presencia de estas frases pudiera indicar que el contenido de la oración hace referencia al tema principal del documento; por ende, esta oración debería formar parte del resumen.

Palabras clave

Cualquier escritor repite ciertas palabras a lo largo de un documento, conforme elabora sus argumentos, y esto puede explotarse para calcular el factor de importancia de cada oración. Precisamente, este es el principio en el que se basa esta métrica: las palabras que aparecen frecuentemente dentro del documento son relevantes. La puntuación final de las oraciones se obtiene sumando las frecuencias de cada una de las palabras clave que incluye.

Palabras título

El título de un documento suele estar fuertemente relacionado con su contenido. Por tanto, las palabras del título se pueden utilizar como palabras clave alternativas a las palabras de alta frecuencia. La métrica *palabras título* precisamente se basa en esta hipótesis. Edmundson es el primero en incorporar esta métrica a un sistema generador de resúmenes. La puntuación final de una oración se calcula sumando los pesos de los términos que incluye.

Localización

Esta métrica explota el hecho de que en algunos géneros, las regularidades de la estructura del discurso o la propia forma de exponer la información hace que ciertas posiciones tiendan a contener puntos importantes del documento. Edmundson favorece a las oraciones que se encuentran cerca del comienzo y del final del texto, pues se supone que estas contienen información importante para los resúmenes por pertenecer a la introducción y a las conclusiones.

2.2. Variante 1 del Algoritmo de Edmundson

Para desarrollar la *Variante 1* del Algoritmo de Edmundson la presente investigación se basó en los estudios de Hoey (1991). Este autor, como habíamos explicado en la *Sección 1.4*., fundamentó su algoritmo en la cohesión léxica. En el algoritmo de Hoey se buscan los *vínculos* y *lazos* que existen entre cada par de oraciones del documento. Se dice que existe un *vínculo* entre dos oraciones de un documento determinado cuando ambas oraciones poseen un término en común. En la *Fig. 2.2.* se muestra un ejemplo

de lo anteriormente expuesto, tomado de (Abreu, 2005). Por otra parte, existirá un *lazo* entre dos oraciones cuando la cantidad de *vínculos* entre ellas supere el promedio de *vínculos* por pares de oraciones.

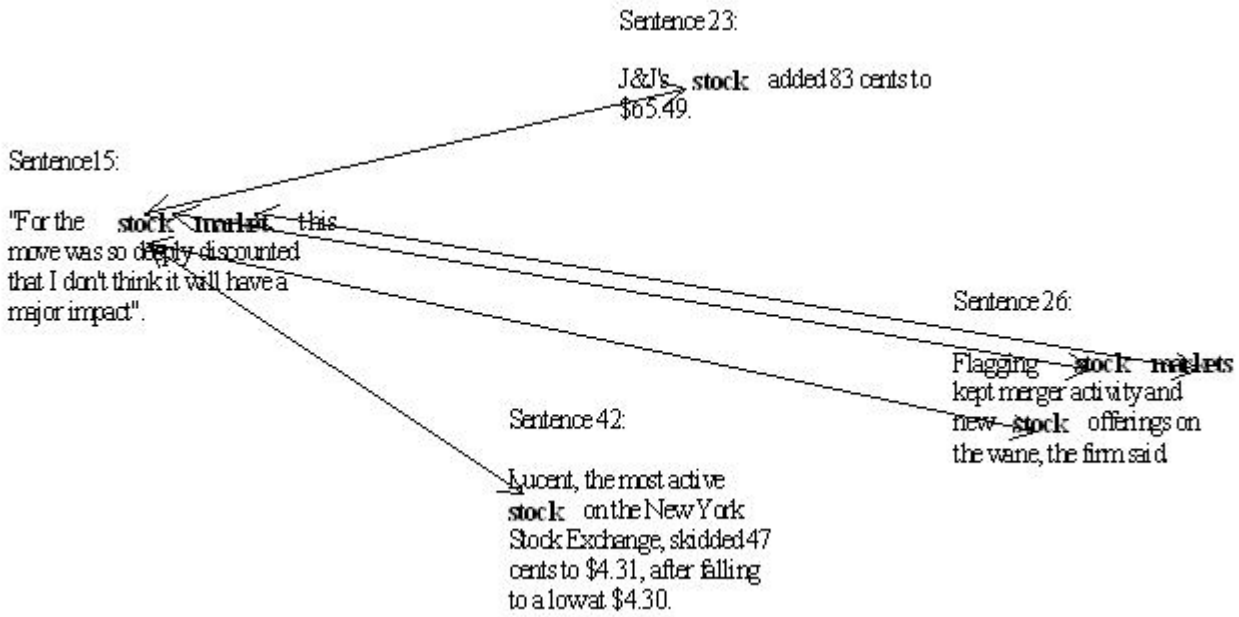

Fig. 2.2. Representación de los vínculos existentes entre cada par de oraciones del fragmento de un documento.

De acuerdo con Hoey, “las oraciones más enlazadas” en el texto son aquellas que tienen un número mayor de *lazos* con otras oraciones, y son aquellas que podríamos denominar *oraciones centrales.* Una oración es *introductoria* si se enlaza con más oraciones subsecuentes que con precedentes, y es *concluyente* si, por el contrario, está más enlazada con las oraciones precedentes que con las subsecuentes. La combinación de las oraciones *introductorias, centrales* y *concluyentes* son las seleccionadas para obtener el resumen del documento.

Como mencionamos al inicio de esta sección, la *Variante 1* se basa en la filosofía que plantea el algoritmo de Hoey, aunque no la implementa de la misma forma. La *Variante 1* utiliza una nueva métrica para la oración, que utiliza el método de la cohesión léxica, con el objetivo de analizar los resultados de la implementación de este experimento. Más adelante será explicado con mayor detalle esta métrica.

2.3. Variante 2 del Algoritmo de Edmundson

La *Variante 2* que se describe en esta sección es una propuesta de mejora del Algoritmo de Edmundson. Los autores de la presente investigación proponen un nuevo método para calcular las métricas *palabras título* y *frases pista*. Teniendo en cuenta que en la literatura especializada no existe ningún método que diferencie la importancia de cada palabra del título de un documento de las demás *palabras título*, la *Variante 2* desarrolla una métrica que le asigna un peso a las mismas, considerando que su frecuencia de aparición en el documento puede ayudar a definir el grado de relevancia que tengan las oraciones que las incluyan.

La métrica *frases pista* tampoco es definida en la literatura con un peso para que diferencie estas palabras indicativas. Las cuales, según sea su repetición en el texto, podría delimitar grados de trascendencia diferentes, estimando que en cada idioma, según el dominio del que se trate, las personas acostumbran a repetir ciertas frases más que otras, lo cual podría ser un identificador de que en estas oraciones hay más probabilidades de encontrar contenido importante, y por ende, deberían tenerse en cuenta en mayor medida que a otras oraciones que incluyan *frases pista* de menor relevancia para el resumen final. Más adelante se ofrecen mayores detalles sobre esta *Variante.*

También, la *Variante 2* incluye una mejora propuesta por (Teufel y Monees, 1997), donde se cambia la forma de calcular el peso asociado a la elección de las *palabras clave.* Estos investigadores proponen que no sólo se determine la *frecuencia del término* (*tf*), sino también la inversa de la *frecuencia del documento* (*idf*)*,* dada por la siguiente fórmula:

idf = *Log N*/*n* (*t*)

donde *N* es el número total de oraciones del documento y *n* (*t*) el número de oraciones en las que aparece el término *t.*

Más adelante se ofrecen más detalles de esta *Variante*.

2.4. Implementación de los Algoritmos

A continuación se pasa a describir, teniendo en cuenta las fases por las que se compone el proceso de extracción, la implementación de los algoritmos seleccionados en la presente investigación.

2.4.1 Fase de Análisis

El primer módulo del *Análisis* es común para todos los algoritmos implementados. En él se construye la representación interna del documento fuente. En el segundo módulo se aplican las métricas para la ponderación de las oraciones del texto. Podríamos ver este módulo como una sub-fase donde se le asocia a cada documento fuente una estructura de datos que lo representa. En la *Fig. 2.3* puede observarse la representación del proceso dividido en sus respectivos módulos.

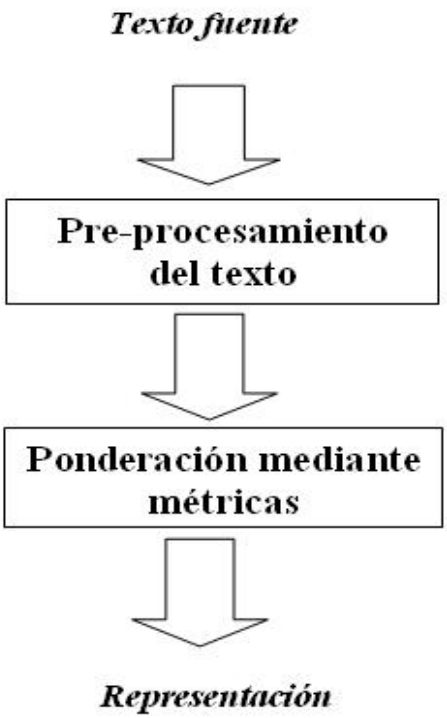

Fig. 2.3. Representación de la fase de análisis de un resumidor basado en técnicas de extracción.

2.4.1.1. Pre-procesamiento del texto

A continuación se analizará en detalle el primer módulo del A*nálisis*. Para mejor comprensión del proceso, se divide, a su vez, en los módulos por los que está compuesta la herramienta implementada.

Módulo de eliminación de elementos superfluos

Como parte del pre-procesamiento del texto, primeramente se realiza la eliminación de los signos de puntuación del documento, caracteres especiales como asteriscos, backslash, etc., así como las stop words (palabras vacías) que contenga el documento, puesto que se concibe que estas son palabras sin significado cuya función es meramente la de enlazar otras y que, además, suelen aparecer frecuentemente; si no son eliminadas se puede afectar el proceso de selección de las *palabras clave*. En esta categoría léxica entran las siguientes palabras: determinantes, conjunciones, pronombres, preposiciones, adverbios, verbos auxiliares y verbos modales, de modo que solamente queden las palabras significativas del texto. En el *Anexo A* puede consultarse la lista de las palabras vacías eliminadas.

Módulo de transformación de los elementos restantes

Seguidamente se procede a realizar un proceso de transformación de las palabras significativas. Este módulo usa una técnica importada de IR (Automated Information Retrieval), que consiste en la eliminación de los sufijos de cada palabra, de forma tal que se conserve sólo el lema de la misma, puesto que este permite considerar como un único concepto todas las formas flexivas de la palabra.

El módulo, además, mediante una serie de reglas establecidas por (Porter, 1980) le adiciona al lema de la palabra la terminación correspondiente, obteniendo la *raíz* del término. A continuación se puede apreciar un ejemplo utilizado para eliminar el plural de diversas palabras:

Si el término termina en "ies" pero no en "eies" o en "aies"

Entonces "ies" -> "y"
Si el término termina en "es" pero no en "aes", "ees", o en "oes"
Entonces "es" -> "e"
Si el término termina en "s", pero no en "us" o en "ss"
Entonces "s" -> NULL

Este proceso es necesario, de otra forma los módulos de ponderación podrían estar comparando dos palabras sintácticamente distintas, sin embargo cuya semántica sea la misma. En el *Anexo B* pueden consultarse las condiciones y reglas del algoritmo Porter.

2.4.1.2. Ponderación mediante métricas

Este módulo del *Análisis* recibe como entrada el texto fuente ya preprocesado; o sea, la representación interna del documento de la cual se hablaba antes, y es el encargado de aplicar los métodos de puntuación de acuerdo con las métricas establecidas para cada algoritmo. Entrega como salida una tabla con la puntuación de cada métrica.

Ponderación para el Algoritmo de Edmundson

Como se vio en la sección 2.1, el Algoritmo de Edmundson le asigna a cada oración un

$peso\ (f) = \alpha C\ (f) + \beta K\ (f) + \gamma L\ (f) + \delta T\ (f)$

donde *C (f), K (f), T (f) y L (f)* representan los valores de las métricas *frases pista*, *palabras clave, palabras título* y *localización, y* α, β, γ, δ son los pesos asociados a cada una de las métricas. En (Abreu, 2005) es utilizado el mismo corpus que se emplea en este trabajo, y en él se determina que los siguientes valores son válidos a los efectos de este tipo de investigación: $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.25$ y $\delta = 0.5$

Frases pista

Las frases pista son aquellas que indican gran probabilidad de resumen. Según el dominio sobre el cual se esté trabajando, estas frases pueden variar. Sin embargo, debido a las características del corpus utilizado en esta investigación (DUC 2001), fueron seleccionadas las siguientes:

this article, the article, this investigation, present investigation, this paper, this study, this work, present work, this letter, in conclusion, is concluded, conclude that, we conclude, in summary, the results, our results, results show, results indicate, results are, in a nutshell, this essay

El algoritmo busca estas frases dentro del texto del documento fuente, otorgando una puntuación a las oraciones que las contienen. En cada oración *f* que aparezca alguna de estas frases, *C (f)* tomará valor de 1, y en las que no, valor 0.

Palabras clave

Según la teoría de Luhn estudiada en el *Capítulo 1*, a cada documento se le puede asociar un conjunto de palabras representativas de su contenido. El módulo de puntuación crea una tabla con todas las palabras del documento fuente y les asigna una puntuación proporcional a su frecuencia de aparición. De esta tabla inicial se eliminan los de baja frecuencia con el objetivo de obviar palabras irrelevantes y también errores ortográficos. El umbral utilizado con este objetivo es el número medio de repeticiones de los términos en el texto.

Luego se le asigna una puntuación a cada oración. El cálculo se realiza comparando los términos de dicha oración con todas las palabras de la tabla creada anteriormente. Se cuenta la cantidad de palabras de la oración coincidentes con las palabras de la tabla y el resultado final se divide por el número total de palabras de la tabla, garantizando que *K (f)* quede normalizado por uno, facilitando así la suma posterior de los resultados obtenidos por las diferentes métricas.

Palabras título

El módulo pondera las oraciones del documento conforme a la aparición en este de las palabras del título (téngase en cuenta que ya en el módulo de pre-procesado se han eliminado las *palabras vacías* del título, así como los sufijos de cada una de sus palabras). Esta métrica queda definida como el cociente entre la cantidad de *palabras título* en la oración y la cantidad de palabras del título, de esta forma el valor de *T (f)* queda normalizado a uno.

Localización

Para la aplicación de esta métrica se tienen en cuenta las oraciones del inicio y del final del documento, pues se supone que éstas contienen información relevante para el resumen por pertenecer, hipotéticamente, a la introducción y a las conclusiones. Para la aplicación de este criterio se han valorado las características del corpus de documentos empleados para las pruebas (DUC 2001).

De esta forma, si la oración se encuentra en alguna de estas partes del documento *L (f)* tomará valor uno, y será cero en caso contrario.

Ponderación para la Variante 1 del Algoritmo de Edmundson

Como se había dicho anteriormente, esta *Variante* se basaba en la estrategia que aplica Michael Hoey en su algoritmo. Esa misma filosofía es aplicada al algoritmo de Edmundson con el objetivo de adicionar una nueva métrica fundamentada en la cohesión léxica. La fórmula quedaría expresada de la siguiente forma:

$$peso\ (f) = \alpha C\ (f) + \beta K\ (f) + \gamma L\ (f) + \delta T\ (f) + \partial V\ (f)$$

donde $\partial = 0.5$ *y V (f)* se calcula teniendo en cuenta la cantidad de *lazos* que se establecen entre las oraciones del documento; pero en este caso los *vínculos* están determinados por las *palabras clave* del texto fuente. De tal forma que si una oración tiene un *lazo* con otra se le asigna un peso de 0.2. Cada vez que se encuentre un nuevo *lazo* para la misma oración, el peso aumentará en 0,1. Finalmente, el

resultado total será normalizado a uno, para facilitar la combinación posterior de los resultados obtenidos por las restantes métricas. De este modo se intenta aumentar el umbral de cohesión del resumen.

Ponderación para la Variante 2 del Algoritmo de Edmundson

En el caso de la *Variante 2,* los autores proponen una mejora al algoritmo de Edmundson; los principios en los que se basan ambas teorías están explicados en la *Sección 2.3*, y consisten en los siguientes cambios introducidos a las métricas *frases pista* y *palabras título.*

Frases pista

Para realizar la mejora de esta métrica primeramente se llevó a cabo un estudio en una colección de 60 documentos del corpus utilizado para la presente investigación, con el objetivo de determinar la razón de repetición de cada una de las *frases pista* seleccionadas anteriormente. A cada *frase pista* se le asignó un peso, de acuerdo con la siguiente fórmula: $p\ (fp) = n/N$ donde *n es* el número de veces que se repite la frase en la colección de documentos, y *N* es el número de documentos analizados. En el *Anexo C* puede consultarse la lista de *frases* y sus respectivos pesos asociados.

Luego, si en alguna oración *f* se encuentra alguna de estas frases, *C (f)* tomará el valor de la sumatoria de los pesos de las *frases pista* que contenga la oración *f*, teniendo en cuenta que ya sus pesos están normalizados.

Palabras título

En este caso el módulo le asigna un peso a cada *palabra título*, el cual será el cociente entre la frecuencia de aparición de la *palabra título* en el texto fuente entre la cantidad de palabras del documento; de esta forma se garantiza que el peso esté normalizado a uno. Luego es comprobada para cada oración *f* si posee alguna *palabra título*, en caso de que la respuesta sea negativa la métrica toma valor cero; de otra forma, *T (f)* será igual a la sumatoria de todos los pesos de las *palabras título* encontradas en la oración.

2.4.2 Fase de síntesis

Esta fase también es común para cada uno de las variantes implementadas. Está conformada por un módulo de post-procesado que es el encargado de, tal como se había analizado anteriormente, combinar las puntuaciones de las métricas de las oraciones, debidamente normalizadas a uno, mediante la fórmula de suma lineal *peso(f)* y así obtener una ponderación total para cada una de ellas.

Este módulo escoge aquellas oraciones que han obtenido mayores puntuaciones, teniendo en cuenta la tasa de compresión deseada para el resumen, que en el caso de la presente investigación ha sido ubicada en un 25%, y genera el resumen final que es entregado como salida del sistema.

Esta fase podría ser también la encargada de comprobar la presencia de ciertas expresiones o marcadores discursivos al comienzo de las oraciones, con el objetivo de editarlas si fuera necesario, aunque para los objetivos de este estudio, y como ya mencionamos en la *Sección 1.6*, no necesariamente debe ser obligatorio, pues el resumen generado mediante extracción es lo suficientemente comprensible por humanos.

La resolución de anáforas es un campo bastante complejo que requiere todavía amplia investigación. Permitiría aumentar la precisión de los métodos basados en técnicas de extracción, así como mejorar la coherencia del extracto.

2.5 Conclusiones parciales

En este capítulo han sido presentados los algoritmos desarrollados en la investigación. Primeramente el algoritmo que se seleccionó luego del estudio del estado del arte y las diferentes técnicas realizado en el *Capítulo 1*. Este algoritmo se implementó tal y como aparece en la literatura.

Seguidamente se describieron también dos variantes al algoritmo anterior: una combinando la teoría de otro investigador reconocido en el campo, con el objetivo de evaluar los resultados de dicho experimento, y una tercera variante, donde se proponen posibles mejoras personalizadas de los autores al primer algoritmo.

Se describieron las implementaciones llevadas a cabo a los tres algoritmos, teniendo en cuenta las fases por las que transita el proceso de sumarización de textos basado en técnicas de extracción, así como la descripción de cada uno de los módulos del sistema.

CAPÍTULO 3: EVALUACIÓN Y ESTUDIO DE LOS RESULTADOS

La evaluación es, como ya se ha comentado, un aspecto fundamental de la investigación en el campo de la generación de resúmenes. A continuación se pasa a describir las evaluaciones que se llevaron a cabo sobre el sistema y los resúmenes obtenidos del mismo.

3.1 Preliminares

En el capítulo anterior se llevó a cabo la implementación de los algoritmos, para lo cual fue desarrollada una herramienta en C#, aprovechando las funcionalidades de la plataforma .NET, con el objetivo de obtener los resúmenes de los algoritmos deseados. La Fig. 3.1 muestra la interfaz de la herramienta implementada.

A continuación se ilustra su funcionamiento con un ejemplo. En el *Anexo D* se muestra un *Documento de Prueba*, tomado al azar; y en (1), (2) y (3) se muestran los respectivos resúmenes producidos por el *Algoritmo de Edmundson*, la *Variante 1* y la *Variante 2*. Este experimento es sólo ilustrativo, con él se busca observar cómo se desenvuelven los algoritmos implementados en un dominio desconocido, puesto que se trata de un documento ajeno al corpus utilizado para esta investigación. Además, en (4) es presentado el extracto generado por el programa Microsoft Office Word 2003, y en (5) el resumen producido por la herramienta TestAnalyst 2.3 (procesador de lenguaje natural). Todos los resúmenes han sido obtenidos con una tasa de compresión del 25%.

(1) *A monster terrifies the sea*

In essence, over a period of time several ships had encountered "an enormous thing" at sea, a long spindle-shaped object, sometimes giving off a phosphorescent glow, infinitely bigger and faster than any whale. In lighthearted countries, people joked about this phenomenon, but such serious, practical countries as England, America, and Germany were deeply concerned. In every big city the monster was the latest rage; they sang about it in the coffee houses, they ridiculed it in the newspapers, they dramatized it in the theaters. In those newspapers short of copy, you saw the reappearance of every gigantic

imaginary creature, from "Moby Dick," that dreadful white whale from the High Arctic regions, to the stupendous kraken whose tentacles could entwine a 500 ton craft and drag it into the ocean depths. They even reprinted reports from ancient times: the views of Aristotle and Pliny accepting the existence of such monsters, then the Norwegian stories of Bishop Pontoppidan, the narratives of Paul Egede, and finally the reports of Captain Harrington - whose good faith is above suspicion - in which he claims he saw, while aboard the Castilian in 1857, one of those enormous serpents that, until then, had frequented only the seas of France's old extremist newspaper, The Constitutionalist.

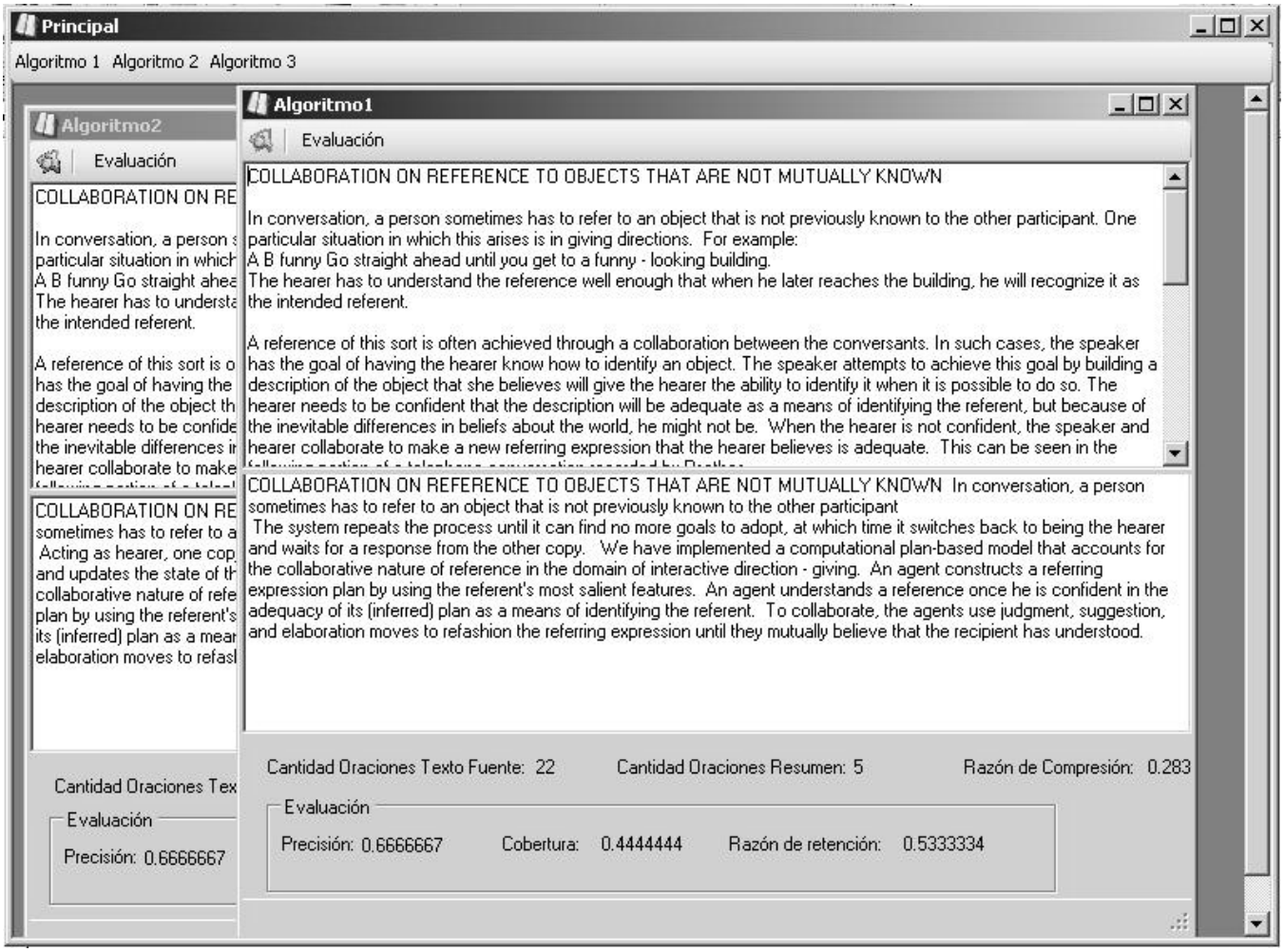

Fig. 3.1 Interfaz de la herramienta de sumarización de textos

(2) *A monster terrifies the sea*

In essence, over a period of time several ships had encountered "an enormous thing" at sea, a long spindle-shaped object, sometimes giving off a phosphorescent glow, infinitely bigger and faster than any whale. One after another, reports arrived that would profoundly affect public opinion: new observations taken by the transatlantic liner Pereire, the Inman line's Etna running afoul of the monster, an official report drawn up by officers on the French frigate Normandy, dead - earnest reckonings obtained by the general staff of Commodore Fitz James aboard the Lord Clyde. In every big city the monster was the latest rage; they sang about it in the coffee houses, they ridiculed it in the newspapers, they dramatized it in the theaters. In those newspapers short of copy, you saw the reappearance of every gigantic imaginary creature, from "Moby Dick," that dreadful white whale from the High Arctic regions, to the stupendous kraken whose tentacles could entwine a 500 ton craft and drag it into the ocean depths. They even reprinted reports from ancient times: the views of Aristotle and Pliny accepting the existence of such monsters, then the Norwegian stories of Bishop Pontoppidan, the narratives of Paul Egede, and finally the reports of Captain Harrington - whose good faith is above suspicion - in which he claims he saw, while aboard the Castilian in 1857, one of those enormous serpents that, until then, had frequented only the seas of France's old extremist newspaper, The Constitutionalist.

(3) *A monster terrifies the sea.*

In essence, over a period of time several ships had encountered "an enormous thing" at sea, a long spindle-shaped object, sometimes giving off a phosphorescent glow, infinitely bigger and faster than any whale. In lighthearted countries, people joked about this phenomenon, but such serious, practical countries as England, America, and Germany were deeply concerned. Similar events were likewise observed in Pacific seas, on July 23 of the same year, by the Christopher Columbus from the West India & Pacific Steam Navigation Co. In every big city the monster was the latest rage; they sang about it in the coffee houses, they ridiculed it in the newspapers, they dramatized it in the theaters. They even reprinted reports from ancient times: the views of Aristotle and Pliny accepting the existence of such monsters, then the Norwegian stories of Bishop Pontoppidan, the narratives of Paul Egede, and finally the reports of Captain Harrington - whose good faith is above suspicion - in which he claims he saw, while aboard the

Castilian in 1857, one of those enormous serpents that, until then, had frequented only the seas of France's old extremist newspaper, The Constitutionalist.

(4) *A monster terrifies the sea.*

If it was a cetacean, it exceeded in bulk any whale previously classified by science. No naturalist, neither Cuvier nor Lacépède, neither Professor Dumeril nor Professor de Quatrefages, would have accepted the existence of such a monster sight unseen - specifically, unseen by their own scientific eyes. In conclusion, on July 20, 1866, the steamer Governor Higginson, from the Calcutta & Burnach Steam Navigation Co, encountered this moving mass five miles off the eastern shores of Australia. Captain Baker at first thought he was in the presence of an unknown reef; he was even about to fix its exact position when two waterspouts shot out of this inexplicable object and sprang hissing into the air some 150 feet. Now then, the biggest whales, those rorqual whales that frequent the waterways of the Aleutian Islands, have never exceeded a length of 56 meters - if they reach even that. One after another, reports arrived that would profoundly affect public opinion: new observations taken by the transatlantic liner Pereire, the Inman line's Etna running afoul of the monster, an official report drawn up by officers on the French frigate Normandy, dead - earnest reckonings obtained by the general staff of Commodore Fitz James aboard the Lord Clyde. In lighthearted countries, people joked about this phenomenon, but such serious, practical countries as England, America, and Germany were deeply concerned.

(5) *A monster terrifies the sea.*

Striking an average of observations taken at different times - rejecting those timid estimates that gave the object a length of 200 feet, and ignoring those exaggerated views that saw it as a mile wide and three long - you could still assert that this phenomenal creature greatly exceeded the dimensions of anything then known to ichthyologists, if it existed at all. In conclusion, on July 20, 1866, the steamer Governor Higginson, from the Calcutta & Burnach Steam Navigation Co, encountered this moving mass five miles off the eastern shores of Australia. So, unless this reef was subject to the intermittent eruptions of a geyser,

the Governor Higginson had fair and honest dealings with some aquatic mammal, until then unknown, that could spurt from its blowholes waterspouts mixed with air and steam. Similar events were likewise observed in Pacific seas, on July 23 of the same year, by the Christopher Columbus from the West India & Pacific Steam Navigation Co. Consequently, this extraordinary cetacean could transfer itself from one locality to another with startling swiftness, since within an interval of just three days, the Governor Higginson and the Christopher Columbus had observed it at two positions on the charts separated by a distance of more than 700 nautical leagues. From their simultaneous observations, they were able to estimate the mammal's minimum length at more than 350 English feet. In those newspapers short of copy, you saw the reappearance of every gigantic imaginary creature, from "Moby Dick," that dreadful white whale from the High Arctic regions, to the stupendous kraken whose tentacles could entwine a 500 ton craft and drag it into the ocean depths.

En la *Tabla 1* se muestran los resultados en términos de reducción:

Factores	***Texto Fuente***	***(1)***	***(2)***	***(3)***	***(4)***	***(5)***
N. oraciones	22	5	5	5	7	6
N. palabras	865	213	249	195	218	279

Tabla 1. Resultados de los extractos.

Obsérvese que los tres resúmenes desarrollados en esta investigación conservan algo de sentido, y contienen material relevante, aún cuando el resumidor no posee conocimiento experto sobre el dominio.

Se puede apreciar que el primer párrafo del texto fuente no contiene información mayormente relevante sobre el tema. Los tres algoritmos, al realizar la ponderación de métricas, eliminan estas ideas vacías en sus respectivos extractos, favoreciendo mayormente la oración "*In essence*", la cual es indicadora de resumen de contenido.

En términos de cobertura informativa, según Cooper (1971): “un documento es relevante a una necesidad de información, si y solamente si, contiene por lo menos una oración que sea relevante a esa necesidad”. En este sentido, los tres resúmenes muestran resultados satisfactorios, incluso puede afirmarse que la *Variante 2* implementada con las mejoras propuestas por los autores define –en este caso- una mejor adecuación con respecto al contenido del texto fuente y tiene un mayor carácter informativo que las otras dos.

Al realizar una comparación con las herramientas de sumarización del *Microsoft Office Word* (4) y del *TextAnalyst* (5), se puede apreciar que los algoritmos de este trabajo contienen una tasa de compresión mayor. En la *Tabla 2* se observa el grado de coincidencia de los resúmenes (1), (2) y (3) con los generados por las otras dos herramientas.

Extractos	(1)	(2)	(3)
(4)	1	1	1
(5)	1	1	1

Tabla 2 Relación de coincidencia

La tabla muestra que cada uno de los resúmenes del presente trabajo tiene al menos una oración coincidente con los otros dos extractos. Lo cual no es poco, si se analiza que la probabilidad de que en cada uno de los tres casos en que se escogieron 5 oraciones de un total de 22, dos de ellas coinciden con las escogidas por los dos extractos tomados como referencia.

El objetivo de esta sección ha sido mostrar, mediante un ejemplo, el funcionamiento de los algoritmos implementados, utilizando para ello como referencia dos extractos generados por herramientas especializadas.

Sin embargo, es de señalar el hecho de que hay muy pocas variaciones entre uno y otro. Ello se debe, como ya se explicó anteriormente, a que el documento es ajeno al corpus utilizado en la investigación. Algunas de las métricas (como la que pondera las *frases pista*) no pueden desarrollarse a plenitud, o se encuentran fuera de contexto. No obstante, los resultados obtenidos en este experimento permiten demostrar la robustez del método escogido (*método de extracción*).

No obstante, estos criterios comparativos no son definitorios. Seguidamente se describirá el proceso de evaluación de los resúmenes generados llevado a cabo como colofón de la investigación.

3.2. Proceso de Evaluación

En esta sección se presentan los resultados obtenidos de evaluar los algoritmos implementados. Para el experimento se seleccionó un dominio de conocimiento particular, tomado de la colección de documentos de DUC 2001 (*Document Understanding Conference*), que cuentan con una serie de textos científicos, artículos periodísticos y extractos "ideales" agrupados, exclusivamente, con el objetivo de ser utilizados para este tipo de investigaciones. En el *Anexo F* puede consultarse uno de estos documentos.

Para los efectos de nuestra evaluación fue realizada la selección de 30 documentos del corpus, preferentemente aquellos cuyo contenido no se fundamentara en análisis numéricos, con la finalidad de garantizar la mayor cantidad de información textual posible. Para ello se revisó el texto completo verificando la cantidad de gráficos, tablas, etc.

Debido a que la apreciación de un resumen por usuarios varía significativamente según las expectativas personales, y el uso de jueces en el proceso no sólo encarece el mismo, sino que puede anularlo, decidimos desechar la opción de realizar evaluaciones extrínsecas. Para llevar a cabo la evaluación directa de los resúmenes, escogimos el método descrito en (Hovy, 1999), el cual se detalla a continuación.

3.2.1 Método de evaluación

Según Hovy, la complejidad de la evaluación de resúmenes estriba en el hecho de que es muy difícil especificar qué es lo que realmente se necesita medir y por qué, si no se tiene una clara formulación de lo que el extracto está tratando de capturar.

En general, para considerarlo como tal, un extracto debe cumplir dos requerimientos: debe ser más corto que el texto fuente, y debe contener la información más importante del original. Para calcular si un resumen *S* cumple dichos requerimientos con respecto a un texto *T,* debe calcularse:

Razón de compresión: RC = (longitud S) / (longitud T)

Razón de retención: RR = (info en S) / (info en T)

Medida de la longitud: Medir la *longitud* es sencillo: sólo se debe contar el número de palabras.

Medida del contenido de la información: Idealmente, lo que debería medirse no es el contenido de la información, sino sólo el contenido de la información *interesante*. Sin embargo es muy difícil definir qué puede resultar *interesante* a los efectos de un texto. Las medidas del *contenido de la información*, en cambio, puede ser determinado de diferentes formas. La siguiente es una de ellas: medir los valores de *precisión* y *cobertura* del resumen creado por el sistema contra los del extracto "ideal".

Precisión = correct / (correct + wrong)

Cobertura = correct / (correct + missed)

Donde *correct* es el número de oraciones extraídas de forma automática y de forma manual, w*rong* es el número de oraciones extraídas de forma automática, pero no de forma manual, y *missed* es el número de oraciones extraídas de forma manual pero no de forma automática.

De esta forma, queda claro que *Precisión* refleja cuántas oraciones extraídas por el sistema fueron acertadas, y *Cobertura* cuántas oraciones acertadas el sistema dejó pasar.

Una vez calculados los factores de P*recisión* y C*obertura, la Razón de Retención* se define por la siguiente fórmula:

RR = 2PC / (P + C)

3.2.2 Resultados alcanzados

	Algoritmo de Edmundson				***Variante 1***				***Variante 2***			
Doc.	***RC***	*P*	*C*	***RR***	***RC***	*P*	*C*	***RR***	***RC***	*P*	*C*	*RR*
1	0.28	0.38	0.53	0.44	0.32	0.54	0.60	0.56	0.23	0.66	0.61	**0.63**
2	0.22	0.45	0.62	0.52	0.40	0.66	0.42	0.51	0.34	0.20	0.12	**0.03**
3	0.26	0.44	0.57	0.49	0.32	0.33	0.25	0.28	0.56	0.53	0.77	**0.62**
4	0.25	0.30	0.75	0.42	0.31	0.20	0.40	0.26	0.43	0.84	0.73	**0.78**
5	0.27	0.85	0.75	0.79	0.28	0.57	0.36	0.44	0.54	0.55	0.38	**0.44**
6	0.31	0.44	0.40	0.41	0.37	0.57	0.72	0.63	0.40	0.66	0.60	**0.62**
7	0.33	0.78	0.83	0.80	0.34	0.46	0.63	0.53	0.31	0.38	0.41	**0.39**
8	0.27	0.36	0.66	0.46	0.35	0.63	0.70	0.66	0.41	0.27	0.75	**0.39**
9	0.25	0.81	0.64	0.71	0.42	0.15	0.40	0.21	0.53	0.76	0.86	**0.80**
10	0.21	0.42	0.37	0.31	0.34	0.54	0.60	0.56	0.22	0.52	0.54	**0.52**
11	0.46	0.45	0.37	0.54	0.44	0.57	0.94	0.23	0.33	0.52	0.82	**0.23**
12	0.43	0.23	0.26	0.35	0.03	0.75	0.24	0.13	0.02	0.40	0.62	**0.13**
13	0.61	0.45	0.74	0.37	0.58	0.75	0.97	0.82	0.75	0.63	0.44	**0.23**
14	0.27	0.67	0.57	0.21	0.41	0.40	0.14	0.73	0.75	0.72	0.35	**0.13**
15	0.42	0.76	0.54	0.43	0.87	0.83	0.08	0.43	0.64	0.36	0.53	**0.82**
16	0.54	0.62	0.62	0.37	0.63	0.66	0.64	0.53	0.36	0.70	0.66	**0.73**
17	0.65	0.27	0.55	0.52	0.18	0.64	0.17	0.24	0.16	0.04	0.11	**0.43**
18	0.36	0.49	0.71	0.12	0.24	0.37	0.13	0.43	0.42	0.61	0.65	**0.53**
19	0.54	0.32	0.69	0.34	0.32	0.26	0.35	0.03	0.28	0.24	0.13	**0.24**
20	0.44	0.91	0.33	0.16	0.52	0.43	0.37	0.58	0.56	0.23	0.21	**0.42**
21	0.65	0.34	0.54	0.72	0.25	0.72	0.14	0.14	0.04	0.45	0.52	**0.46**
22	0.4	0.43	0.36	0.24	0.67	0.54	0.35	0.24	0.38	0.08	0.43	**0.30**
23	0.28	0.56	0.41	0.45	0.26	0.26	0.37	0.63	0.61	0.37	0.35	**0.85**
24	0.73	0.17	0.12	0.57	0.12	0.45	0.65	0.81	0.46	0.72	0.24	**0.14**
25	0.28	0.23	0.46	0.63	0.94	0.17	0.16	0.42	0.74	0.31	0.34	**0.77**
26	0.76	0.15	0.67	0.52	0.18	0.46	0.17	0.24	0.15	0.04	0.12	**0.35**
27	0.2	0.31	0.34	0.12	0.24	0.73	0.13	0.43	0.45	0.06	0.65	**0.22**
28	0.86	0.12	0.23	0.64	0.54	0.72	0.45	0.41	0.75	0.49	0.32	**0.33**
29	0.53	0.63	0.12	0.34	0.32	0.62	0.53	0.30	0.50	0.42	0.31	**0.20**
30	0.62	0.22	0.34	0.61	0.45	0.74	0.37	0.58	0.75	0.97	0.82	**0.75**
Promedio	**0.42**	**0.45**	**0.50**	**0.45**	**0.38**	**0.52**	**0.41**	**0.43**	**0.44**	**0.46**	**0.48**	0.45

Tabla 3 Resultados de la evaluación de los extractos.

En la *Tabla 3* se muestran los resultados obtenidos para cada uno de los extractos, aplicando el método de Hovy.

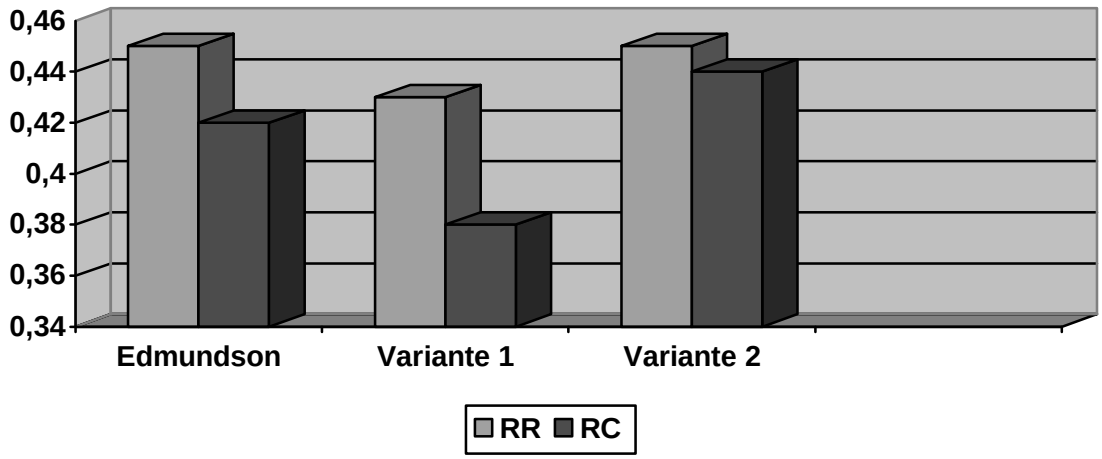

Fig.3.2 Comparación de los resultados obtenidos en términos de RR y RC.

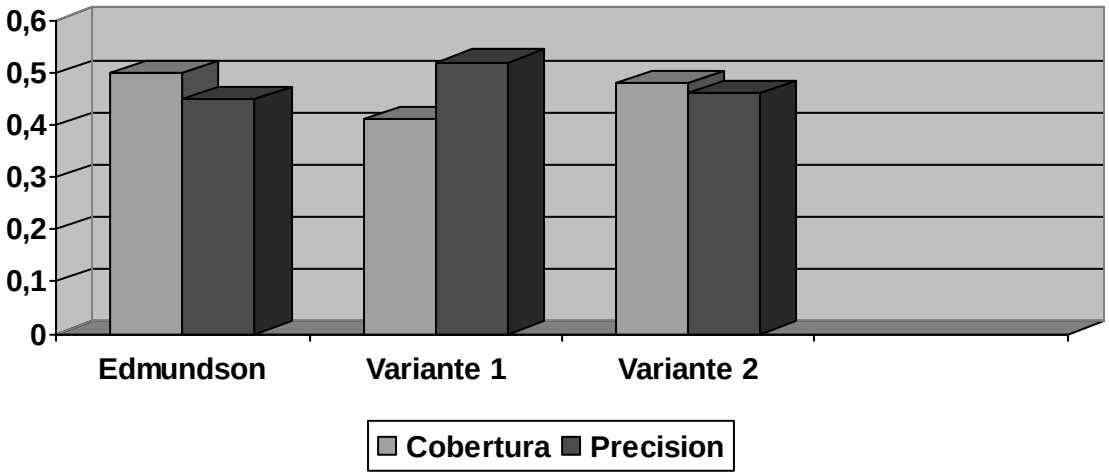

Fig. 3.3 Comparación en términos de C y P.

En la *Fig. 3.2* se ilustran los resultados obtenidos en cuanto a *Razón de Compresión* (*RC)* y *Razón de Retención* (*RR)*, después de evaluar los algoritmos mediante el método estadístico de Hovy. Se puede

apreciar que la *Variante 1* obtuvo el mejor rendimiento en cuanto a RC, es decir, este algoritmo fue el que mayor compresión de los documentos alcanzó. La *Variante 2* implementada por los autores de esta investigación no alcanzó las mejores puntuaciones en este factor, y en RR no logró mejorar el puntaje del algoritmo original.

En la *Fig. 3.3* se muestra la comparación realizada teniendo en cuenta los términos de *Cobertura* (*C*) y *Precisión (P*). Una vez más, la *Variante 1* obtuvo la mejor *P* y también *C* (recuérdese que la *cobertura* representa el número de oraciones relevantes que el sistema no incluyó en su extracto).

También se realizó una evaluación mediante métodos de inferencia basados en pruebas de hipótesis a los parámetros de evaluación de resúmenes (*Cobertura, Precisión, Razón de Compresión y Razón de Retención*), la cual puede consultarse en el *Anexo G*; luego de lo cual se determinó que las variantes propuestas por los autores mejoran al Algoritmo Básico en el parámetro de *cobertura y* la *razón de compresión*, con nivel de significación del 10%, no siendo así para la *precisión* y la *razón de retención*, cosa que podría cambiar para un nivel de confianza mayor que el 90%.

3.3 Conclusiones parciales

En este capítulo se ha analizado la técnica seleccionada para evaluar a los tres algoritmos implementados, así como las comparaciones realizadas teniendo en cuenta factores estadísticos y de inferencia con un nivel de significación del 10% que ayudan a determinar la calidad de los resúmenes generados. Después de llevadas a cabo las evaluaciones, se ha llegado a la conclusión de que la *Variante 1* implementada por los autores fue la que mejor resultados obtuvo. La *Variante 2* quedó muy por debajo de las otras dos, generando resúmenes de gran tamaño y escasa relevancia.

CONCLUSIONES GENERALES

Una vez analizados los resultados puede afirmarse que los objetivos trazados al inicio de la investigación han sido cumplidos:

- Se desarrolló un estudio de los algoritmos existentes en la literatura y se seleccionó uno de ellos.

- Se propusieron dos variantes a dicho algoritmo, una de ellas una propuesta nueva de los autores de la presente investigación.

- Se implementaron las tres variantes del algoritmo y se demostró que la *Variante 1* mejora al algoritmo original.

RECOMENDACIONES

- Diseñar una herramienta de pre-procesado que incluya etiquetamiento léxico-morfológico para no restringirse sólo a la eliminación de los sufijos del idioma.

- Continuar el estudio de técnicas de post-procesado, como la eliminación de anáforas, para algoritmos de extracción de frases, con el objetivo de lograr resúmenes de mayor coherencia y calidad.

- Realizar un estudio más profundo de las bases de conocimiento léxico para ampliar la búsqueda con sinónimos de las palabras.

- Estudiar la posibilidad de aplicar los algoritmos para el idioma español.

BIBLIOGRAFÍA

Abracos, J. y Lopes, G. P. 1997. Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles. I. Mani y M. T. Maybury, eds., Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics. Madrid, Spain.

Abreu, I. 2005. Algoritmos para la construcción automática de resúmenes de documentos de texto. Santiago de Cuba, Universidad de Oriente, 2005.

Alfonseca, E. y Rodríguez, P. 2003. Description of the UAM system for generating very short summaries at DUC 2003. Procceding of the Third Document Understanding Conference (DUC '03), Edmonton, Canada, May 31 – June 01, 2003.

Alfonseca, E. y Rodríguez, P. y Moreno-Sandoval, A. 2004. Description of the UAM system for generating very short summaries at DUC 2004. Procceding of the Fourth Document Understanding Conference (DUC '04), Boston, Massachusetts, USA, May 6-7, 2004.

Angheluta, R., Busser, R. D. y Moens, M.-F. 2002. The Use of Topic Segmentation for Automatic Summarization. Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D. R., Jones, K. S., Sundheim, B., Teufel, S., Weischedel, R. y White, M. 2000. An Evaluation Road Map for Summarization Research. Informe técnico, DARPA's TIDES (Translingual Information Detection, Extraction, and Summarization) program.

Borko, H. y Bernier, C. 1975. Abstracting Concepts and Methods. Academic Press, New York.

Brandow, R., Mitze, K. y Rau, L. F. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. Information Processing and Management, 31(5):675-685.

Burgos, R., J.A. Chicharro y M. Bobenrieth. 1994. Metodología de investigación y escritura científica en la clínica. Escuela Andaluza de Salud Pública. Granada.

Chuang, W. T. y Yang, J. 2000. Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages. 152-160.

Conroy, J. y O'Leary, D. P. 2001. Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition. Technical Report, Dept. Comp. Sci., CS-TR-4221, Univ. Maryland.

Contreras, H. y Dávila, J. 2001. Procesamiento del lenguaje natural basado en una "gramática de estilos" para el idioma español. CLEI'2001. Mérida, Venezuela.

Cooper, W.S. 1971. A definition of relevance for information retrieval. Information Storage and Retrieval, 7, 19-37.

Dunlavy, D. M. 2003. Q C S: An Information Retrieval System for Improving Efficiency in Scientific Literature Searches. QCS IR System Project Proposal. Final Report for version 1.0. May, 2003.

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

Fuentes, M. y H. Rodríguez 2002. Using cohesive properties of text for Automatic Summarization. JOTRI 2002- Workshop on Processing and Information Retrieval.

Fukumoto, F. y Suzuki, Y. 2000. Extracting Key Paragraphs Based on Topic and Event Detection - Towards Multi-Document Summarization. *Proceedings of the Workshop on Automatic Summarization at*

the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics.

Goldstein, J., M. Kantrowitz, V. Mittal y J. Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Carnegie Mellon University.

Hahn, U. y Mani, I. 2000. The Challenges of Automatic Summarization. Computer, 33(11):29-36.

Halliday, M. y Hasan, R. 1996. Cohesion in English. Longmans, London.

Hoey, M.. *Patterns of Lexis in Text*. Oxford: Oxford University Press, 1991.

Hovy, E. y Lin, C. Y. 1999. Automated Text Summarization in SUMMARIST. I. Mani y M. T. Maybury, eds., *Advances in Automatic Text Summarization*, pages. 81-94. The MIT Press.

Johnson, F. C., Paice, C. D., Black, W. J. y Neal, A. P. 1993. The Application of Linguistic Processing to Automatic Abstract Generation. Journal of Document and Text Management, 1(3):215-239.

Kupiec, J., Pedersen, J. O. y Chen, F. 1995. A Trainable Document Summarizer. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Lin, C.-Y. and Hovy, E.: *Automatic evaluation of summaries using n-gram co-occurrence statistics*. *Proceedings of HLTNAACL*, 2003.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*

Mani, I. y Bloedorn, E. 1998. Machine Learning of Generic and User-Focused Summarization. Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98), pages. 821-826. Menlon Park, California.

Mani, I. y Bloedorn, E. 1998. Machine Learning of Generic and User-Focused Summarization. Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98), pages. 821-826. Menlon Park, California.

Mani, I. 2001. *Automatic Summarization*. John Benjamin's Publishing Company, Amsterdam-Philadephia.

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T. y Sundheim, B. 2002. SUMMAC: A Text Summarization Evaluation. *Natural Language Engineering.*

Mann, W. y Thompson, S. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. Text, 8(3):243-281.

Maña, M. J., de Buenaga, M. y Gómez, J. M. 1999. Using and evaluating user directed summaries to improve information access. *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*.

Maña, M.J., 2003. Generación automática de resúmenes de texto para el acceso a la información. Tesis doctoral. Departamento de Informática. Universidad de Vigo. Septiembre 2003.

Maña, M. J., de Buenaga, M. y Gómez, J. M. 2003. Multi-document summarization: An added value to clustering in interactive retrieval. *ACM Transactions on Information Systems (TOIS)*.

Maybury T. M. and Mani, I. 2001, Automatic Summarization Tutorial Notes for the American/European Conference on Computational Linguistics (ACL/EACL '01).

Mitra, M., Singhal, A. y Buckley, C. 1997. Automatic Text Summarization by Paragraph Extraction. Proceedings of the Workshop on Intelligent Scalable Text Summarization, pages. 39-46. Association for Computational Linguistics, Madrid, Spain.

Morris, A., Kasper, G. y Adams, D. 1992. The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. Information Systems Research, 3(1):17-35.

Nanba, H. y Okumura, M. 2000. Producing More Readable Extracts by Revising Them. En Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), pages. 1071-1075.

Nomoto, T. y Matsumoto, Y. 2001. A New Approach to Unsupervised Text Summarization. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New Orleans, LA.

Paice, C. D. 1990. Constructing Literature Abstracts by Computer: Techniques and Prospects. Information Processing and Management, 26(1):171-186.

Paice, C. y Jones, P. A. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. R. Korfhage, E. Rasmussen y P. Willett, eds., *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." Program, 14(3), 130-37.

Radev, D. R. y Teufel, S., eds. 2003. *Proceedings of the 3rd Document Understanding Conference*. Edmonton, Canada.

Rush, J. E., Zamora, A. y Salvador, R. 1971. Automatic Abstracting and Indexing. II, Production of Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. Journal of the American Society for Information Science, 22(4):260-274.

Saggion, H. y Lapalme, G. 2002. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics.*

Salton, G. y McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Salton, G., Allan, J., Buckley, C. y Singhal, A. 1994. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. Science, 264(3):1421-1426.

Salton, G., Allan, J. y Singhal, A. 1996. Automatic Text Decomposition and Structuring. Information Processing & Management, 32(2):127-138.

Salton, G., Singhal, A., Mitra, M. y Buckley, C. 1997. Automatic Text Structuring and Summarization. Information Processing & Management, 33(2):193-207.

Sparck-Jones, K. 1993. What Might Be In A Summary. Information Retrieval 93: Von der Modellierung zur Anwendung.

Sparck-Jones, K. y Galliers, J. R. 1996. Evaluating Natural Language Processing Systems: An Analysis and Review, tomo 1083 de Lecture Notes in Artificial Intelligence. Springer, New York.

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

Teufel, S. y Moens, M. 1997. Sentence Extraction as a Classification Task. En Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics. Madrid, Spain.

Vanderwende, L., Banko, M., y Menezes, A. 2004. Event-centric summary generation. Proceedings of the Fourth Document Understanding Conference (DUC '04), Boston, Massachusetts, May 6-7, 2004

Zechner, K. 1997. A literature survey on information extraction and text summarization. Carnegie Mellon University.