

# Resúmenes automáticos: Enfoque extractivo y evaluación.

Víctor Márquez Gil, [victor.marquez@estudiante.uam.es](mailto:victor.marquez@estudiante.uam.es)  
Escuela Politécnica Superior, Universidad Autónoma de Madrid.

## Índice

<b>1. Introducción</b>	<b>2</b>
<b>2. Tipos de resúmenes</b>	<b>5</b>
<b>3. Enfoque de extracción</b>	<b>7</b>
3.1. Orígenes	7
3.2. Métodos estadísticos	7
3.3. Métodos de aprendizaje supervisado (o basados en corpus)	8
3.4. Métodos de aprendizaje no supervisado	9
3.5. Extracción de hechos	10
3.6. Ventajas e inconvenientes	11
3.7. Revisión	11
<b>4. Evaluación</b>	<b>13</b>
4.1. Orígenes	13
4.2. Clasificación de los métodos de evaluación	13
4.3. Programas y concursos	15
<b>5. Trabajo futuro</b>	<b>17</b>
<b>6. Conclusiones</b>	<b>18</b>
<b>7. Referencias bibliográficas</b>	<b>19</b>

---

**Resumen.** En este trabajo se presentan el estado actual de la cuestión y la evolución de los sistemas de resúmenes automático. Se hace especial hincapié en los métodos de generación extractivos y en la evaluación de la producción final de los mismos, desde los orígenes hasta el presente. También se presentan distintas clasificaciones de resúmenes y los factores a tener en cuenta a la hora de diseñar e implementar un sistema de estas características. Finalmente se plantea el trabajo futuro a seguir en este área de investigación, sugiriendo la incorporación de bases de conocimiento lingüístico y conceptos de estructura del discurso.

---

## 1. Introducción

El objetivo de este trabajo es dar a conocer el estado del arte de los *métodos extractivos* y de la *evaluación* de los resúmenes automáticos dónde tanto el documento fuente como el resumen están en formato de texto.

Resumir automáticamente implica que un artefacto software tome una fuente de información, extraiga el contenido más relevante y se lo presente al destinatario (ya sea un usuario u otro artefacto software) de manera *condensada* de tal forma que satisfaga las necesidades del usuario o de la aplicación (Mani, 2001). Para el objetivo que nos atañe, se define la acción de resumir como una transformación de *reducción* del texto fuente al texto resumido por medio de selección de lo que es importante en la fuente (Spark Jones, 2007a).

Es difícil imaginar la vida diaria sin algún tipo de resumen. Los titulares de noticias, el trailer de una película, los resúmenes de la contraportada de los libros o al principio de los artículos científico-técnicos, incluso la programación de la televisión en el teletexto, el mapa de una ciudad, un catálogo de productos o el resultado de un encuentro deportivo son resúmenes. Como se puede apreciar, el medio de la fuente a resumir puede ser muy variado, así como el del resumen en sí. Esta revisión del estado del arte se centrará en técnicas cuyas fuentes a resumir y sus versiones condensadas sean textuales.

Vivimos en la Sociedad de la Información y el Conocimiento: cada vez hay más información accesible en Internet. La red de redes crece de manera vertiginosa en todo tipo de contenidos. Esta explosión de información conlleva un problema: la sobrecarga. No hay tiempo para leerlo todo, sin embargo es necesario tomar decisiones críticas basadas en la información disponible. En este contexto surge la necesidad de desarrollar sistemas que resuman automáticamente los contenidos y por consiguiente de fomentar la inversión en investigación en este dominio del Procesamiento del Lenguaje Natural (PLN).

Actualmente, la investigación en este campo del PLN es muy activa gracias a programas y concursos como SUMMAC, NTCIR, DUC (Over *et al.*, 2007) o TAC (Louis, 2008) y está emparentada con la de otros campos (Mani, 2001) como:

- *Compresión de texto*: el objetivo también es crear una versión condensada de documento fuente pero con el fin de ser almacenado y transmitido de manera eficiente y no para el consumo humano.
- *Indexación*: el objetivo es la identificación de términos relevantes de un documento, normalmente para facilitar la recuperación de información. Se puede concebir la indexación como un caso particular de resumen automático, pero al ser su propósito el de la recuperación de información y no el de resumir lo dejamos como campo aparte.
- *Minería de textos*: se trata de un proceso cuyo objetivo es la detección de información nueva o anómala en grandes repositorios de textos. Su relación con los resúmenes automáticos es que la salida es una versión reducida de la entrada, la diferencia es que la minería de textos no se centra en condensar el contenido de la fuente sino en caracterizar singularidades de los datos.

Los sistemas que generan resúmenes automáticamente pueden *clasificarse* en dos grandes grupos según la *estrategia* de condensación: los que construyen resúmenes por *extracción* (o sistemas *extractivos*) y los que lo hacen por *abstracción* (o sistemas *no extractivos*). También pueden considerarse los sistemas *híbridos* que aunan ambas técnicas. A continuación se definen dichas estrategias:

- *Estrategia extractiva*: el resumen generado consta únicamente de material copiado literalmente del documento fuente.

- *Estrategia no extractiva o abstractiva*: al menos parte del material presente en el resumen no se encuentra en el documento fuente.

La *arquitectura* abstracta de los sistemas que producen resúmenes automáticos es siempre la misma (Fig. 1) y consta de tres procesos (Hahn & Mani, 2000; Spark Jones, 2007a): *interpretación o análisis*, *transformación* y *generación o síntesis*. Conviene comentar que esta es una estructura lógica de alto nivel, por lo que los módulos y procesos de implementaciones concretas no tienen por qué responder a este esquema. La descripción de cada uno de los procesos es la siguiente:

- *Interpretación o análisis*: se analiza el documento fuente y se construye una representación interna del mismo.

- *Transformación*: se transforma la representación interna del documento en una representación interna del resumen. Esta fase es sobre todo aplicable a sistemas abstractivos que se basan en técnicas de PLN para generar resúmenes.

- *Generación o síntesis*: se toma la representación interna del resumen y se construye el mismo en lenguaje natural.

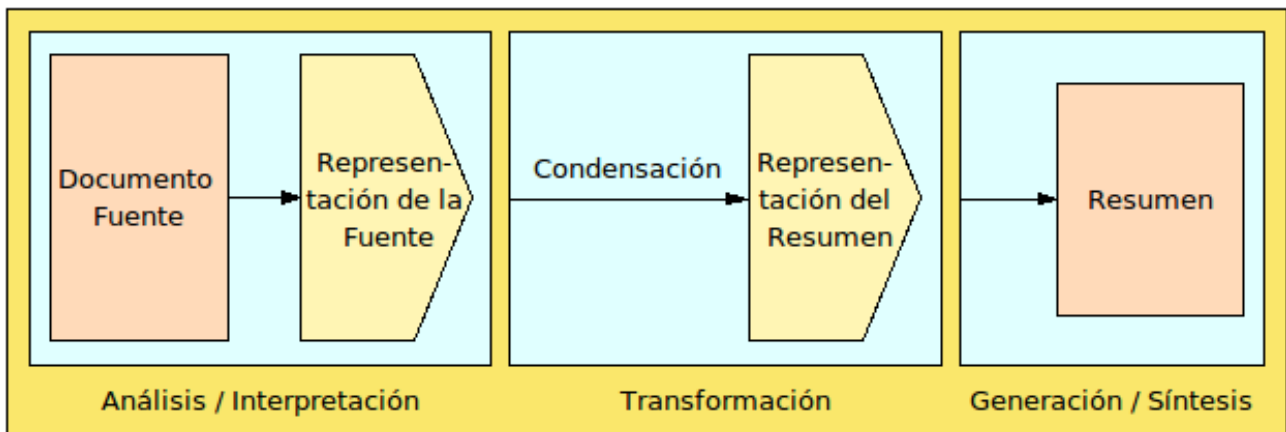


Figura 1: Arquitectura de alto nivel de los sistemas de resúmenes automáticos. [Adaptado de Mani (2000) y Spark Jones (2007a)]

Para el caso particular de los sistemas basados en *extracción*, los procesos de interpretación y transformación se fusionan como se muestra en la Figura 2.

La elección de centrarse en los resúmenes automáticos basados en técnicas *extractivas* no es trivial. El coste computacional es menor, su implementación es más fácil, y suelen dar mejores resultados que los basados en técnicas abstractivas que requieren recursos de conocimiento lingüísticos (Mani, 2001). Los acercamientos que persiguen objetivos más genéricos suelen basarse en métodos de extracción puramente estadísticos y ofrecen normalmente resultados aceptables inde-

pendientemente del idioma, del género y del propósito final (Spark Jones 2007). Además, el mayor empuje que ha recibido la investigación en el campo de los resúmenes automáticos ha sido por parte de estos métodos, junto con métodos híbridos que combinan técnicas estadísticas y simbólicas. Esto contrasta con el interés mostrado por parte de la comunidad de la lingüística computacional hacia herramientas de *representación del significado* del texto.

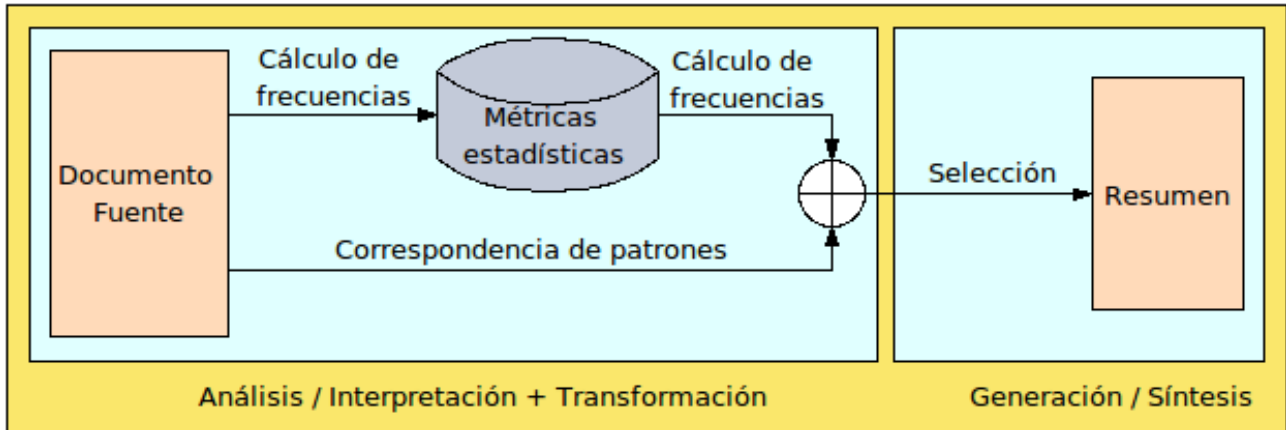


Figura 2: Arquitectura de alto nivel de los sistemas de resúmenes automáticos extractivos. [Adaptado de Mani (2000)]

En la sección 2 exploraremos las diferentes *clasificaciones* de resúmenes más aceptadas en la literatura. En la sección 3 describiremos los *orígenes* y el *progreso* de los sistemas extractivos así como los distintos *métodos* usados en ellos y sus *ventajas* e *inconvenientes*. En la sección 4 inspeccionaremos los *comienzos*, *evolución* y diferentes *clasificaciones* de la *evaluación* de resúmenes generados automáticamente. Finalizaremos con el *trabajo futuro* de esta rama de la investigación del PLN y las *conclusiones* que se pueden arrojar del trabajo realizado en las secciones 5 y 6 respectivamente.

## 2. Tipos de resúmenes

Como hemos visto, los resúmenes pueden diferenciarse por la *estrategia* empleada al generarlos: por *extracción* o por *abstracción*. Otra forma de clasificarlos es haciendo la distinción entre resúmenes *indicativos*, *informativos* y *críticos* (Hahn & Mani, 2000):

- *Indicativos*: estos resúmenes siguen el acercamiento clásico de la recuperación de la información, suministran el contenido suficiente para que los usuarios puedan hacerse una idea de los conceptos clave que se encuentran en el documento fuente para que decidan si leer el contenido completo más en detalle.

- *Informativos*: en este caso los resúmenes sirven como substitutos del documento fuente, el proceso de crear estos resúmenes consiste principalmente en recopilar información relevante de manera estructurada.

- *Críticos*: aparte de ser informativos, incorporan la opinión del escritor del resumen como valor añadido, aportando su experiencia la cual no está reflejada en el documento fuente.

También se pueden clasificar los resúmenes en términos del *propósito* final de los mismos. Los resúmenes pueden ser *genéricos* u *orientados al usuario* (Mani, 2001).

- *Genéricos*: estos resúmenes no están hechos a medida de ninguna audiencia o propósito en particular. La historia de la investigación sobre resúmenes automáticos se ha concentrado principalmente en la producción de este tipo de resúmenes (Over *et al.*, 2007). La idea de generar resúmenes automáticos de un único documento fue el primer impulso de la investigación en este campo.

- *Orientados al usuario*: estos resúmenes se adaptan a las necesidades de un usuario o grupo de usuarios para una tarea concreta (Mani, 2001). Esto significa que el sistema tiene en cuenta de alguna manera una representación de los intereses de los usuarios mediante alguna técnica de modelado de usuario o bien con un simple campo de formulario para ejecutar una consulta. Un caso particular de estos sistemas son los que el usuario introduce una pregunta y el sistema devuelve un resumen que intenta responderla.

Otra posibilidad para diferenciar los resúmenes es según el *tipo de entrada*. Desde los orígenes de la investigación en este campo el foco ha sido crear resúmenes de *un solo documento* fuente. Pero a partir del final de la última década del siglo XX, gracias al empuje de los programas de evaluación y como respuesta a la demanda de querer saber de un vistazo cual es el tema de una colección de documentos, nació la idea de los sistemas de resumen automático de *múltiples documentos*. El objetivo de estos últimos es que partiendo de una serie de documentos relacionados, se obtenga un resumen que contiene el contenido más relevante eliminando la redundancia que exista entre los documentos fuente (Mani, 2001). Los primeros trabajos sobre sistemas de resúmenes automáticos multi-documento son de Salton *et al.* (1997) y Mani & Bloedorn (1999).

Más recientemente todavía, se ha empezado a manejar el concepto de resúmenes de fuentes *multimedia*. Se trata no sólo de resumir textos sino también vídeos (Mani, 2001), imágenes (Simakov, 2008), grabaciones de audio de reuniones (Mani *et al.*, 2000) o diagramas (Frutelle, 1999) y combinaciones de estos elementos como por ejemplo el trabajo de Merlino & Maybury (1999) con noticias de informativos de televisión.

### 3. Enfoque de extracción

#### 3.1. Orígenes

Probablemente el primer trabajo sobre resúmenes automáticos sea el de [Luhn \(1958\)](#). En él se describe una simple técnica para generar resúmenes *genéricos extractivos* de un sólo documento fuente. Esta consiste en usar frecuencias de términos para dar pesos a las frases que posteriormente se seleccionan para aparecer en el resumen. El proceso consiste en lo siguiente: primero se *filtran palabras* como artículos, pronombres y preposiciones; a continuación *se normalizan los términos* para reducirlos a su lexema; posteriormente se *agregan* las palabras con la misma raíz y se calculan las *frecuencias* de los términos agregados descartando los que tengan menores frecuencias; finalmente, a partir de los resultados del paso anterior se le da un peso a cada oración y las frases con mayor peso se *seleccionan* para el resumen final.

Otro avance importante de los orígenes de los resúmenes automáticos fue el trabajo de [Edmundson \(1969\)](#). En él se añaden tres nuevas *características* a la hora de resumir automáticamente aparte de la frecuencia de términos:

- *Expresiones clave* que aumentan o disminuyen la puntuación de la frase en la que se encuentran, algunos ejemplos son: “significante”, “imposible” o “difícilmente”.
- Palabras que si aparecen en el *título* o en los *subtítulos* aumentan la puntuación de la frase.
- La *posición* de la frase dentro del documento o párrafo y su cercanía con los subtítulos también modifica la valoración de la oración a la hora de tenerla en cuenta para el resumen.

Al estudiar los resúmenes generados, [Edmundson \(1969\)](#) descubrió que estas tres características, cada una por separado, daba mejores resultados que el acercamiento de [Luhn \(1958\)](#). También se encontró con que la mejor combinación de características era la de expresiones clave, títulos y posición. Además, la característica aislada que proporcionaba mejores resúmenes era la de posición, y la que proporcionaba los peores era la de frecuencia de términos.

#### 3.2. Métodos estadísticos

Los métodos estadísticos simples suelen seguir el acercamiento de [Luhn \(1958\)](#). Que, como hemos visto, consiste en puntuar las oraciones según el peso de los términos que aparecen en ella. Después de filtrar y la normalizar los términos, se extraen las oraciones en el mismo orden que aparecen en la fuente. Por último, se seleccionan las  $n$  oraciones mejor puntuadas según la tasa de compresión que se quiera alcanzar. También hemos visto cómo se pueden añadir características como las de palabras clave-título-posición a las frecuencias de términos con el acercamiento de [Edmundson \(1969\)](#).

Una de las mayores líneas de investigación en este campo ha sido la elaboración de nuevas características de las oraciones. También se ha explorado la puntuación de unidades más

elementales que la oración: unidades léxicas como sintagmas, n-gramas u otras ventanas de texto (Spark Jones, 2007a).

El avance de las técnicas, métodos y herramientas de PLN han motivado que la extracción pueda tener una base lingüística utilizando relaciones de redes semánticas como WordNet para, por ejemplo, agrupar conceptos por sinonimia (Bellare *et al.*); o tesauros para generalizar conceptos (McCargar, 2004).

### 3.3. Métodos de aprendizaje supervisado (o basados en corpus)

Hemos visto cómo ciertas *características* ayudan a descubrir lo que es *relevante* en un texto fuente y por tanto debe extraerse para el resumen. El problema consiste ahora en determinar la contribución de cada una de las características para generar resúmenes lo más cercanos a los modelos producidos por humanos. La solución a este problema no es trivial y, de hecho, es sumamente *dependiente del género* del documento fuente. Por ejemplo, la característica de posición en noticias de prensa escrita hará que se extraiga el titular y la entrada o copete de la noticia, ya que en estas partes se resume la misma; sin embargo, en un artículo científico-técnico, esta característica debería dar más peso a oraciones pertenecientes al resumen, la introducción, la conclusión y las primeras frases tras los subtítulos.

La determinación de la importancia de las características puede hacerse mediante el uso de un *corpus* de textos del mismo género, dónde están emparejados el documento fuente y el resumen realizado por un humano. El uso del corpus también le permite al sistema aprender automáticamente nuevas reglas útiles para la generación de resúmenes automáticos (Many & Maybury, 1999).

Uno de los primeros trabajos que introdujeron el uso de corpus para el entrenamiento de sistemas de resúmenes automáticos fue el de Kupiec *et al.* (1995). En él se usan los *resúmenes modelo* para etiquetar vectores de las oraciones de los documentos fuente como ejemplares positivos o negativos indicando si son candidatos para aparecer o no en el resumen. El corpus consta de 188 pares documento fuente/resumen pertenecientes a 21 colecciones de documentos científicos. El proceso es el que sigue: un algoritmo de clasificación Bayesiano toma cada oración del conjunto de test y calcula una *probabilidad* de estar incluida en el resumen basándose en la frecuencia de características en los vectores del documento fuente y las *etiquetas* de los vectores (1 si debería ser incluida en el resumen, 0 en cualquier otro caso); finalmente se extraen las *n* oraciones más probables de aparecer en el resumen, dependiendo de la tasa de reducción. En la [Figura 3](#) se ilustra la estructura del sistema. Las características usadas en este trabajo son: la longitud de las oraciones, presencia de expresiones clave, posición de las frases dentro de los párrafos del documento fuente, y presencia de nombres propios.

El trabajo de Kupiec *et al.* (1995) inspiró a Myaeng & Jang (1999). Estos últimos, en su variante aplicada a artículos técnicos en Coreano, consideran el material de la Introducción y la Conclusión y *etiquetan* las oraciones manualmente si representan los antecedentes, algún tema principal, la descripción de la estructura del documento o la descripción del trabajo futuro. También se etiquetan las frases candidatas a aparecer en un resumen realizado por un humano. Su método de entrenamiento utiliza primero un clasificador Bayesiano para determinar si la oración considerada pertenece a algún tema principal, y a continuación combina los indicios de múltiples clasificadores de características Bayesianos para determinar si la oración se añade al resumen. Finalmente se

aplica un filtro para *eliminar frases redundantes*. Los autores descubrieron que, con sus datos, usar una combinación de palabras clave, posición, y presencia en la oración de palabras del título, daba los mejores resultados.

Aone *et al.* (1999) utilizan un acercamiento similar pero usando conceptos como “familias de sinónimos” y en Hovy & Lin (1999) se describen diferentes técnicas para generar resúmenes automáticos basadas en corpus.

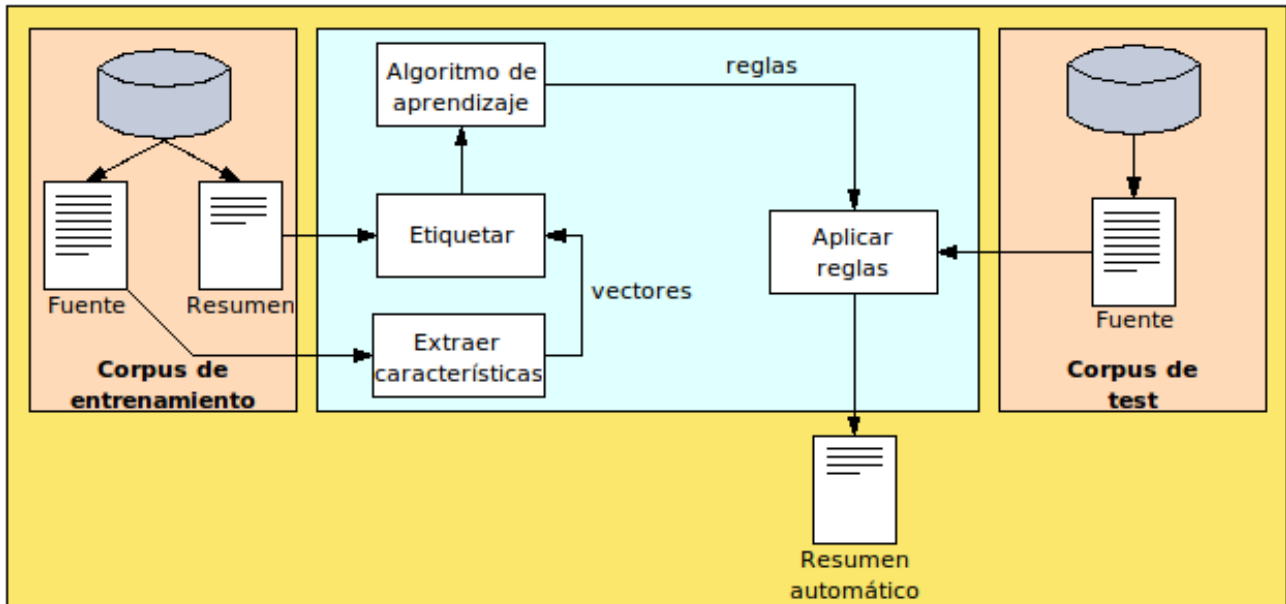


Figura 3: Metodología del sistema de Kupiec *et al.* (1995).  
[Adaptado de Mani & Maybury (1999)]

### 3.4. Métodos de aprendizaje no supervisado

Alfonseca & Rodríguez (2003) proponen un procedimiento de generación de resúmenes automáticos basado en algoritmos genéticos. El genotipo de un resumen es la lista de oraciones que aparecerán en él. Definen una función de ajuste basada en características de los resúmenes informativos y no en características de las oraciones como hemos visto en ejemplos anteriores. Dichas características son:

- *Longitud del resumen*: los resúmenes que contienen oraciones largas son mejores que los que tienen oraciones cortas.

- *Posición*: los resúmenes que contienen oraciones del principio de los párrafos del documento fuente son mejores que los que contienen oraciones de cualquier otra parte de los párrafos.

- *Orden de las oraciones*: los resúmenes que tienen sus oraciones en el mismo orden que en el documento fuente son mejores que los que tienen las oraciones en otro orden.

- Los resúmenes que contienen *frases de todos los párrafos* del texto fuente son mejores que los generados en la situación contraria.

- Los resúmenes que contienen *oraciones relacionadas con el perfil del usuario* son mejores que los que no las contienen.
- Si el usuario especifica una consulta, los resúmenes que *contienen palabras de la consulta* son mejores que los que no las contienen.
- Los resúmenes que contienen *oraciones completas*, es decir, con sujeto y verbo, son mejores que los que no contienen alguno de estos elementos.
- Las *preguntas* son *poco informativas*, por lo que se penaliza su presencia en el resumen.

El proceso seguido es el siguiente: se parte de una población inicial aleatoria de resúmenes (o individuos); después, en cada generación, los dos individuos menos adaptados mueren y los dos mejor adaptados tienen descendencia; los individuos se alteran por mutación, cambiando una oración por otra, y por crossover, dónde dos individuos intercambian una porción aleatoria de sus genotipos. Como se puede observar en la [Tabla 1](#), el valor de la función de ajuste va aumentando en el resumen con mejor puntuación de diferentes generaciones.

Generación	Oraciones	Ajuste
0	1 6 8 13 20 28 29 33 35 41 42 44 46	46.501553
5	1 6 13 20 22 28 29 33 35 41 42 44 46	47.385387
10	1 3 4 6 13 22 29 33 35 41 42 44 46	49.599186
20	1 3 4 6 13 26 29 33 35 39 41 42 44	51.74695
50	3 4 19 24 25 26 29 39 40 41 42 43 44	54.43973

**Tabla 1: Resumen con la mejor puntuación en diferentes generaciones.**  
[Adaptado de [Alfonseca & Rodríguez \(2003\)](#)]

En el trabajo de [Silla et al. \(2004\)](#) se plantea la generación automática de resúmenes como un *problema de clasificación*: el sistema extrae las oraciones individuales del documento fuente, a cada oración se le asocia un vector de atributos cuyos valores se derivan del contenido de la misma y finalmente se clasifica la oración en dos clases dependiendo de si entrará en el resumen o no.

### 3.5. Extracción de hechos

[Spark Jones \(1998\)](#) define la *extracción de hechos* en contraposición a la *extracción de texto*. Con la extracción de texto, “lo que ves es lo que obtienes”, es decir, parte de lo que se ve en el documento fuente se transfiere al resumen generado. El enfoque de la extracción es diferente ya que “lo que sabes es lo que obtienes”, es decir, se decide *a priori* que tipo de contenido se va a buscar en la fuente.

La extracción de texto es un enfoque abierto en el que se deja emerger el contenido relevante de la fuente. Mientras que la extracción de hechos es un enfoque cerrado en el que el texto fuente no proporciona más que alguna instanciación de requisitos de contenidos genéricos previamente establecidos, por lo que solamente permite una único punto de vista de lo que es relevante del documento original.

La *extracción de hechos* consiste en un procesado del texto fuente en busca de conceptos preestablecidos para rellenar algún tipo de plantilla con mayor o menos modificación de la expresión original. Un ejemplo de sistema que utiliza este método es el de [Young & Hayes \(1985\)](#)

que trabaja con telegramas bancarios.

### 3.6. Ventajas e inconvenientes

Las *ventajas* del enfoque extractivo son:

- Su implementación es *sencilla*.
- Su *bajo coste* en esfuerzo humano, económico, computacional y temporal (Mani & Maybury, 1999).
- Es *consistente* y evita la subjetividad de los redactores de resúmenes humanos (Luhn, 1958; Rath *et al.*, 1961).
- Suelen dar *mejores resultados* que las técnicas abstractivas (Mani, 2001).

Y sus *inconvenientes* son los siguientes (Mani, 2001):

- La falta de *coherencia* del resumen generado.
- La *redundancia* del contenido por tratar las oraciones de manera independiente.

### 3.7. Revisión

Cuando se extraen oraciones del documento fuente, seleccionar oraciones independientes fuera de contexto aparecen los *problemas* que acabamos de mencionar: la *incoherencia* y la *redundancia*. Existen técnicas para mitigar el impacto de estos inconvenientes. A continuación veremos a fondo en que consisten estos problemas y como paliarlos.

La *incoherencia* en los resúmenes se da por algunos de los siguientes motivos (Mani, 2001):

- *Anáforas* no resueltas: se extrae, por ejemplo, una oración que contiene un pronombre que hace referencia a un antecedente que se encuentra en otra oración no seleccionada para ser extraída.
- *Lagunas*: normalmente los textos están escritos de tal manera que las ideas se conectan entre sí, en un texto en el que se extraen de manera independiente las oraciones pueden perderse algunas de estas conexiones.
- *Entornos estructurados*: elementos de los textos como listas o tablas crean problemas a la hora de resumir automáticamente, si se extrajese de un texto una oración como “Las tres partes del sistema son:” y no se extraen los tres elementos que componen la lista, o si se extrae alguna de las partes y no su descripción se pierde el contexto y se genera el problema de la incoherencia.

La solución consiste en el *suavizado superficial de coherencia* que consiste en identificar cuando una *anáfora* queda sin antecedente requiere encontrar en el texto expresiones que se refieran a dicho antecedente, esto no es fácil ya que se necesita conocimiento lingüístico y conocimiento específico de dominio. La solución más fácil es excluir todas las oraciones que contengan anáforas (Brandow *et al.*, 1995). Otra estrategia más sofisticada sería incluir una ventana de oraciones anteriores esperando que el antecedente esté en ellas (Mani, 2001). Otro sistemas más avanzados intentan localizar el antecedente como en el trabajo de Paice (1990). Para tratar las *lagunas* se han utilizado métodos muy simples, un ejemplo se encuentra en el trabajo de Brandow *et al.*, (1995), en

el que incluye en el resumen frases no seleccionadas para ser extraídas entre dos que si se han seleccionado; o si se selecciona e incluye la n-ésima oración de un párrafo, incluir también la primera. En el caso de los *entornos estructurados* es muy difícil analizar la estructura del entorno, por lo que la solución más fácil es identificar dicha estructura y excluirla; dado el auge del XML, otra idea sería utilizar los meta-datos para intentar resumir la estructura.

Para mermar el efecto negativo de la redundancia, se puede aplicar un método conocido como Relevancia Máxima Marginal (MMR: Maximal Marginal Relevance), su uso requiere una representación más rica de la fuente que registre las palabras de las oraciones. [Carbonell & Goldstein \(1998\)](#) aplican este método: sólo se añaden las oraciones a la selección si difieren de las extraídas previamente.

## 4. Evaluación

La evaluación es una parte esencial de una disciplina práctica como la de la generación de resúmenes automáticos. De hecho la evaluación es parte de lo que se acuña como método científico; la habilidad de diseñar experimentos y evaluar los resultados obtenidos, puede ayudar a construir un argumento científico a favor o en contra de una teoría o un método. Se puede considerar la evaluación desde el punto de vista de la desarrollo de teorías; así, la evaluación proporciona una prueba para confirmar o refutar una hipótesis o un conjunto de ellas. De hecho, la evaluación puede dar lugar a nuevas hipótesis, por lo que proporciona una estrategia de investigación y un marco teórico para varias etapas del desarrollo (Mani, 2001).

La generación de resúmenes es todavía un campo de investigación práctico, no existe todavía un marco teórico con el que trabajar. Por eso, se hace necesario que diferentes métodos puedan ser comparados para que sus ventajas y desventajas particulares puedan ser mejor comprendidas.

### 4.1. Orígenes

En los albores de la investigación en el campo de los resúmenes automáticos ya se idearon tanto métodos informales de evaluación (Pollock & Zamora, 1975), estudios más organizados (Edmundson, 1969), comparativas contra otros sistemas y contra líneas base (Brandow *et al.*, 1995). A finales de los 90 empezaron a surgir programas como SUMMAC (Mani *et al.*, 1999) o DUC (Baldwin *et al.*, 2000; Over *et al.*, 2007) para juzgar sistemas que generan resúmenes automáticos.

### 4.2. Clasificación de los métodos de evaluación

En el terreno de la *evaluación* se suele hacer la distinción entre evaluación *intrínseca* y evaluación *extrínseca* (Mani, 2001):

- *Intrínseca*: evaluación cuyo énfasis se centra directamente en la calidad del resumen creado.
- *Extrínseca*: evaluación cuyo énfasis se centra en cuan bien ayuda el resumen a realizar una tarea con un propósito específico.

Los atributos más habitualmente evaluados de manera *intrínseca* son la *calidad* de la salida generada y lo *informativo* que es el resumen generado. Estos juicios son subjetivos ya que los realizan humanos, y los jueces pueden disentir. Si la diferencia entre las posturas de dichos jueces es demasiado grande, la evaluación puede llegar a no ser útil entre tanto desacuerdo.

Para medir la calidad Minel *et al.* (1997) los sujetos debían valorar la *legibilidad* de los resúmenes generados, y puntuarlos basándose en la presencia de *anáforas* sin resolver, falta de *conservación* de la integridad de los *entornos estructurados* como listas o tablas, presencia de afirmaciones tautológicas como “Predecir el futuro es difícil”, etc.

Para medir cuan *informativo* es un resumen generado automáticamente se puede comparar

contra modelos escritos por humanos como hace Edmundson (1969) en su trabajo. También se puede valorar la fidelidad al documento fuente como es el caso de Brandow *et al.* (1995).

La idea de evaluación *extrínseca* de los resúmenes creados automáticamente es determinar el efecto de la realización del resumen en alguna otra *tarea*. Los métodos de este tipo de evaluación son los que siguen:

- Valoración de la *relevancia*: han habido muchos acercamientos extrínsecos de evaluación para la valoración de la relevancia, esto suele consistir en que a un sujeto se le presenta un tema y un documento. A continuación, se les pide determinar la relevancia de un tema con respecto al documento. Finalmente, se estudia la influencia de la síntesis de resúmenes automáticos sobre la precisión de valoración de la relevancia y el tiempo empleado para realizar la tarea. Un ejemplo es el trabajo de Tombros & Sanderson (1998), en el que se les pedía a los sujetos que encontrasen tantos documentos relevantes como les fuese posible en un tiempo de cinco minutos, los resultados indicaron que los resúmenes *orientados al usuario* proporcionaban mejores resultados en la precisión y tiempo de la valoración de la relevancia.

- *Comprensión* de la lectura: en tareas destinadas a valorar la comprensión de la lectura, el sujeto humano lee bien el documento o el resumen generado; a continuación, debe contestar una serie de preguntas multi-respuesta; los resultados del test son almacenados por el sistema como un porcentaje de preguntas contestadas correctamente. De esta manera se puede valorar objetivamente la comprensión y comparar los resultados obtenidos con el resumen y el documento original. El razonamiento es que si con la lectura del resumen se obtienen resultados similares que con la del documento fuente, el valor informativo del resumen es alto. Un ejemplo es el trabajo de Morris *et al.* (1992) donde se evalúa el impacto de la acción de resumir en una tarea de respuestas a preguntas.

- Otras valoraciones extrínsecas son las de *estrategias de presentación*, especialmente útiles para la evaluación de sistemas de resúmenes *multimedia*; y la *evaluación de sistemas maduros*, que se hace útil cuando el sistema de generación de resúmenes automáticos es lo suficientemente maduro como para tener usuarios finales (Mani *et al.*, 2001).

Por otra parte, Spark Jones (2007a) establece un gradiente clasificando los métodos de evaluación enfocados al propósito del resumen que va de *semi-orientado al propósito*, pasando por *quasi-orientado al propósito* y *pseudo-orientado al propósito* hasta llegar al *totalmente orientado al propósito* (Fig. 4). El orden de esta clasificación, tal y como ha sido anteriormente expuesto corresponde con un gradiente de intrínseco a extrínseco. La autora afirma que es imposible evaluar resúmenes si no se conoce para que son. También distingue dos tipos de evaluaciones: por comparación con el texto fuente y por comparación con modelos de resúmenes escritos por humanos. Finalmente propone que la dirección a tomar en la evaluación de resúmenes debe orientarse a la toma de decisiones, es decir, cuanto más ayude el resumen a tomar una decisión correcta, mejor resumen será.

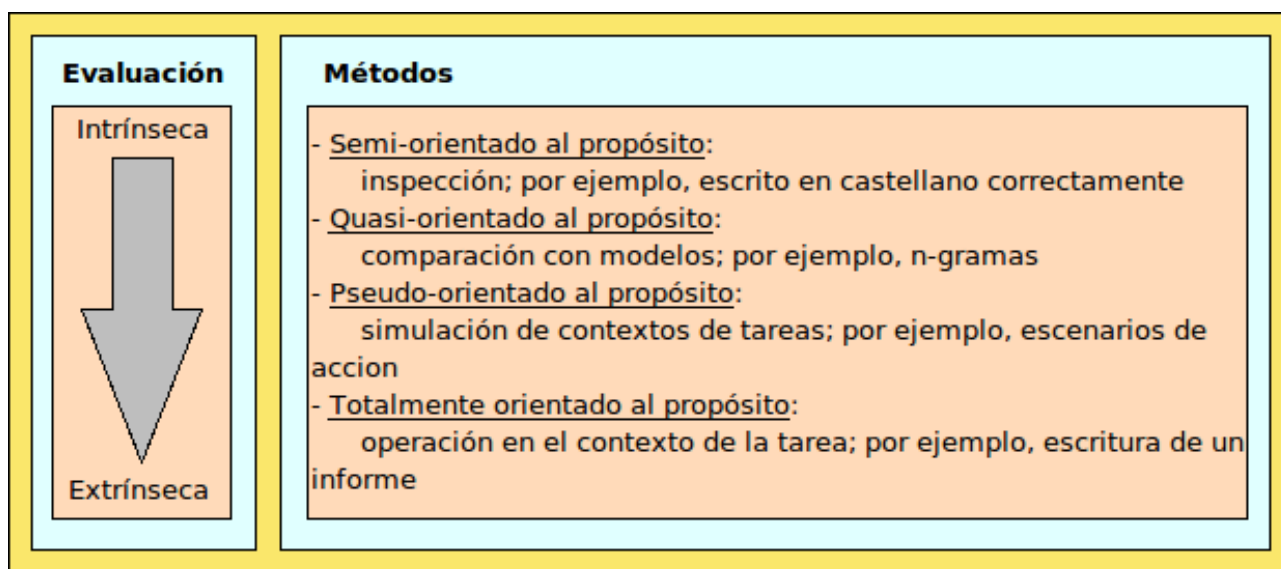


Figura 4: Gradiente de la evaluación relacionada con el contexto de la tarea.  
[Adaptado de Spark Jones (2007b)]

### 4.3. Programas y concursos

En los últimos ha habido un gran interés en la evaluación de prototipos de sistemas de síntesis de resúmenes automáticos (Over *et al.*, 2007). En esta sección describiremos brevemente los programas de evaluación más importantes de las últimas dos décadas. Estos programas son:

- SUMMAC (Summarization Evaluation Conference): fue una evaluación a gran escala de sistemas de resúmenes automáticos de texto que tuvo lugar en 1998 como parte del programa TIPSTER de la Administración de Proyectos Avanzados de Investigación de Defensa (Defense Advanced Research Projects Administration [DARPA]) (Mani *et al.*, 1999). Participaron 16 sistemas teniendo en cuenta la evaluación *extrínseca* de dos tareas del mundo real. La primera consistía en procesar una lista de documentos para encontrar los relevantes. La segunda era una tarea de categorización en la que, por ejemplo, se presentaba un conjunto de 1000 documentos que debían ser agrupados en 10 clases.

- NTCIR (National Institute for Informatics Test Collection for IR): tuvo lugar en 2000, 2002 y 2004 en Japón. Cada año se presentaban 10 sistemas que trabajaban en dos tareas de generación de resúmenes. En el 2000 (NTCIR2, 2001) se utilizó la técnica *extrínseca* de *valoración de la relevancia*, así como evaluación *intrínseca* que para los sistemas extractivos consistía en medir el número de correspondencias entre oraciones seleccionadas por humanos y las extraídas por los sistemas. En Over *et al.* (2005) se hace un resumen del resto de técnicas usadas en otros años en los que el programa tuvo lugar hasta 2007.

- DUC (Document Understanding Congerence): comenzó como un proyecto piloto en el 2000 y la primera evaluación a fondo fue en 2001. La hoja de ruta en sus principio sugería una forma de evaluación *intrínseca*, que más adelante se fuesen introduciendo métodos *extrínsecos*. Se han considerado tanto resúmenes de un solo documento como resúmenes multi-documento, así como genéricos y orientados al usuario. Hoy en día este programa, se conoce como TAC (TAC, 2010).

- TAC (Text Analysis Conference): desde el 2008 DUC se conoce como TAC. Para el 2010 se proponen dos tareas: creación de *resúmenes dirigidos* y *evaluación automática de resúmenes por parejas* (Automatically Evaluating Summaries Of Peers [AESOP]). La creación de resúmenes dirigidos consiste en hacer un resumen de no más de 100 palabras de un conjunto de 10 documentos de un determinado tema y cada tema pertenece a una categoría determinada, cada categoría cubre varios aspectos y dichos aspectos deben encontrarse en el resumen (TAC, 2010). En el caso de AESOP, la tarea consiste en valorar automáticamente resúmenes para una métrica dada. Para obtener más información al respecto consúltese (TAC, 2010). Para la evaluación de los resúmenes dirigidos, el Instituto Nacional de Estándares y Tecnología (National Institute of Standards and Technology [NIST]) valora manualmente el *contenido* siguiendo el método piramidal<sup>1</sup> de la Universidad de Columbia, la legibilidad y la fluidez, y sensibilidad global del resumen.

---

<sup>1</sup> Una pirámide es un modelo que predice la distribución del contenido de la información en los resúmenes tal y como se refleja en los resúmenes escritos por humanos. Para más información consúltese la siguiente referencia: <http://www.l.cs.columbia.edu/~becky/DUC2006/bibliography.html>

---

## 5. Trabajo futuro

Ya desde su trabajo de 1998, [Spark Jones \(1998\)](#) establece un marco para el estudio de los factores que intervienen a la hora de generar un resumen. Insiste en, y revisa, dichos factores en su trabajo de 2007 ([Spark Jones, 2007a](#)). Aún a día de hoy, muchos de los prototipos y sistemas que ven la luz no tienen en cuenta todos estos factores, aunque algunos de ellos si que incorporan algunos en diferentes etapas de la generación de los resúmenes, pero de manera muy limitada. Se consideran tres grandes familias de *factores de contexto*: de *entrada*, de *propósito* y de *salida*:

- Factores de *entrada*:
  - Forma
    - Idioma
    - Registro
    - Medio
    - Estructura
    - Género
    - Extensión
  - Temática
  - Unidades
  - Autor
  - Metadatos
- Factores de *propósito*
  - Uso
  - Audiencia
  - Envoltura
    - Momento
    - Ubicación
    - Formalidad
    - Destinatario
- Factores de *salida*
  - Material
    - Cobertura
    - Condensación
    - Derivación
    - Especialidad
  - Estilo
  - Forma
    - Idioma
    - Registro
    - Medio
    - Estructura
    - Género

Estos factores han de tenerse en cuenta en un futuro cercano tanto en los métodos para resumir automáticamente como en la evaluación de los mismos.

Es posible que los métodos estadísticos estén llegando al punto en el que se requiere demasiado esfuerzo para obtener mejoras nimias en los resultados. Los métodos híbridos que aunan métodos estadísticos y simbólicos dan mejores resultados. Esto parece indicar que se requiere de la incorporación de análisis lingüísticos más profundos. Se debe por tanto avanzar hacia estrategias más ambiciosas que exploten la información semántica y del discurso.

Hasta el momento los métodos abstractivos no han tenido gran éxito debido a sus pobres resultados. Esto no quiere decir que haya que abandonar esta línea de investigación. El avance en técnicas de comprensión del texto seguramente influya significativamente en el progreso de estas técnicas a largo plazo.

De momento, dado el auge de las técnicas extractivas e híbridas, parece razonable hacer especial hincapié en la revisión del resumen generado usando técnicas que tengan en cuenta el contexto para mejorar la coherencia. En este momento estas técnicas son mayoritariamente superficiales, habría que aprovechar más los recursos semánticos que nos ofrecen las bases de conocimiento lingüístico como WordNet o EuroWordNet.

## 6. Conclusiones

Sería interesante hacer una clasificación ordenada de las técnicas de extracción. Pero ante la multitud de criterios a valorar, los diferentes métodos de evaluación y todos los factores que influyen a la hora de generar un resumen automático, esto se hace imposible. Aún así [Mani \(2001\)](#) hace una revisión de las ventajas y debilidades de cada una de las técnicas extractivas.

Como ya se ha mencionado, los métodos extractivos ofrecen resultados aceptables cuando la utilidad del resumen generado es genérica. Son fáciles de implementar y su coste es muy asumible, tanto computacional como económicamente. Las características de ubicación y palabras clave parecen ser en general las más efectivas. Otras características también lo son cuando nos centramos en un género específico, pero con gran variabilidad en diferentes dominios.

Cuando se compara la extracción y la abstracción, más allá de que en los métodos abstractivos se modifique de alguna manera el texto fuente, lo importante es que llevan a cabo una elaboración detallada y organizada de los conceptos del documento fuente. Esto requiere un nivel de entendimiento del texto que de momento no ha dado buenos resultados. Pero esta línea de investigación parece la más prometedora a largo plazo.

En cuanto a la evaluación se puede decir que ha sido la impulsora de los avances en el campo de los resúmenes automáticos. Es más, ha jugado un papel fundamental en el vertiginoso progreso de las tecnologías de análisis del lenguaje. Aunque existen nuevas áreas de la generación de resúmenes automáticos que requieren de nuevos métodos de evaluación como: resúmenes para dispositivos móviles o resúmenes usados como consultas para la recuperación de documentos relevantes.

## 7. Referencias bibliográficas

- Alfonseca, E. & Rodríguez P. (2003). "Generating Extracts with Genetic Algorithms", *Advances In Information Retrieval*, vol. 2633, pp. 511-519.
- Aone, C.; Gorlinsky, J.; Bjornar, L. & Okurowski, M.E. (1999). "A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 72-80, Cambridge MA: MIT Press, 1999.
- Baldwin, B.; Donaway, R.; Hovy, E.; Liddy, E.; Mani, I.; Marcu, D.; McKeown, K.; Mittal, V.; Moens, M.; Radev, D.; Sparck Jones, K.; Sundheim, B.; Teufel, S.; Weischedel, R. & White, M. (2000). "An Evaluation Road Map for Summarization Research" <http://www-nlpir.nist.gov/projects/duc/papers/summarization.roadmap.doc> (Último acceso: 19 de mayo de 2010)
- Bellare, K.; Das Sarma, A.; Das Sarma, A.; Loival, N.; Mehta, V.; Ramakrishnan, G. & Bhattacharya, P. "Generic Text Summarization using WordNet", <http://i.stanford.edu/~anishds/publications/lrec04/lrec04.ps> (Último acceso: 16 de mayo de 2010).
- Brandow, R.; Mitze, K. & Ray, L. (1995). "Automatic Condensation of Electronic Publications by Sentence Selection", *Information Processing & Management*, vol. 31, no. 5, pp 675-685.
- Carbonell, J. & Goldstein, J. (1998). "The Use of MMR and Diversity-based Reranking for Reordering Documents and Producing Summaries", *Proceeding of the 21<sup>st</sup> Annual International ACM-SIGIR conference on Research and Development in Information Retrieval (SIGIR 2001)*, pp. 335-336.
- Edmundson, H.P. (1969). "New Methods in Automatic Extracting", *Journal of the Association for Computing Machinery*, vol. 16, no. 2, pp 264-285.
- Fattah, M.A. & Ren, F. (2008). "Automatic Text Summarization", *Proceedings Of World Academy Of Science, Engineering And Technology*, vol. 27, pp. 192-195.
- Frutelle, R.P. (1999). "Summarization of Diagrams in Documents", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 403-421, Cambridge MA: MIT Press, 1999.
- Hahn, U. & Mani, I. (2000). "The Challenges of Automatic Summarization", *IEEE Computer*, vol. 33, no. 11, pp. 29-36.
- Hovy, E. & Lin, C.Y. (1999). "Automated Text Summarization in SUMMARIST", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 403-421, Cambridge MA: MIT Press, 1999.
- Kupiec, J.; Pedersen, J. & Vhen, F. (1995). "A Trainable Document Summarizer", *Proceedings of the 18<sup>th</sup> ACM-SIGIR Conference*, pp. 68-73.
- Louis, A. & Nenkova, A. (2008). "Automatic Summary Evaluation without Human Models", *Notebook Papers and Results of the Text Analysis Conference (TAC-2008)*.
- Luhn, H.P. (1958). "The Automatic Creation of Literature Abstracts", *IBM Journal of Research*

- Development*, vol. 2, no. 2, pp. 159-165. (Reimpreso en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 15-21, Cambridge MA: MIT Press, 1999)
- Mani, I. & Bloedorn, E. (1999). "Summarizing Similarities and Differences Among Related Documents", *Information Retrieval*, vol. 1, no.1, pp. 1-23.
- Mani, I.; House, D.; Klein, G.; Hirschman, L.; Firmin, T. & Sundheim, B. (1999). "The TIPSTER SUMMAC Text Summarization Evaluation", *Proceedings of the Ninth Conference on European Chapter of the Association For Computational Linguistics*, pp. 77-85.
- Mani, I. & Maybury, M. editors (1999). "Advances in Automatic Text Summarization", Cambridge: Massachusetts: MIT Press.
- Mani, I; Concepcion, K & Van Guilder, L. (2000). "Using Summarization for Automatic Briefing Generation", En *Proceedings of the Workshop on Automatic Summarization*, pp. 98-108. New Brunswick, New Jersey: Association for Computational Linguistics.
- Mani, I. (2001). "Automatic Summarization", Amsterdam: John Benjamins Publishing.
- McCargar, V. (2004). "Statistical Approaches to Automatic Text Summarization", *Bulletin of the American Society for Information Science and Technology*.
- Merlino, A. & Maybury, M. (1999). "An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 391-401, Cambridge MA: MIT Press, 1999.
- Minel, J.L.; Nugier, S. & Piat, G. (1997). "How to Appreciate the Quality of Automatic Text Summarization", *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 25-30. New Brunswick, New Jersey: Association for Computational Linguistics.
- Morris, A.; Kasper, G. & Adams, D. (1992). "The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance", *Information Systems Research*, vol. 3, no. 1, pp. 17-35.
- Myaeng, S.H. & Jang, D.H. (1999). "Development and Evaluation of a Statistically-Based Document Summarization System", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 61-70, Cambridge MA: MIT Press, 1999.
- NTCIR2. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. (2001) <http://research.nii.ac.jp/ntcir/ntcir-ws2/ws-en.html> (Último acceso: 20 de mayo de 2010)
- Over, P.; Dang, H. & Harman D. (2007). "DUC in Context", *Information Processing & Management*, vol. 43, no. 6, pp. 1506.
- Paice, C.D. (1990). "Constructing Literature Abstracts by Computer: Techniques and Projects", *Information Processing & Management*, vol. 26, no. 1, pp. 171-186.
- Pollock, J.J. & Zamora, A. (1975). "Automatic Abstracting Research at Chemical Abstracts Service", *Journal of Chemical Information and Computer Sciences*, vol. 14, no. 4, pp. 226-232.

- Rath, G.J.; Resnick, A. & Savage, T.R. (1961). "The Formation of Abstracts by the Selection of Sentences", *American Documentation*, vol. 2, no. 2, pp. 139-143, (actualmente titulado *Journal of the American Society for Information Science*).
- Salton, G.; Singhal, A.; Mitra M. & Buckley C. (1997). "Automatic Text Structuring and Summarization", *Information Processing & Management*, vol. 33, no. 2, pp. 193-207.
- Silla, C.N.; Pappa, G.L.; Freitas, A.A. & Kaestner, C.A.A. (2004). "Automatic Text Summarization with Genetic Algorithm-Based Attribute Selection", *Lecture Notes in Computer Science*, vol. 3315.
- Simakov, D.; Caspi, Y.; Shechtman, E. & Irani, M. (2008). "Summarizing Visual Data Using Bidirectional Similarity", en *CVPR*, IEEE Computer Society.
- Sparck Jones, K. (1998). "Automatic Summarizing: Factors and Directions", en Mani, I. & Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 1-12, Cambridge MA: MIT Press, 1999.
- Sparck Jones, K. (2007). "Automatic Summarising: The State of the Art", *Information Processing & Management*, vol. 43, no. 6, pp. 1449. (2007a)
- Sparck Jones, K. (2007). "Automatic Summarising: A Review and Discussion of the State of the Art", Technical Report 679, Computer Laboratory, University of Cambridge. (2007b)
- TAC (2010). "TAC 2010 Summarization Track", *Text Analysis Conference*. <http://www.nist.gov/tac/2010/Summarization/index.html>
- Tombros, A. & Sanderson, M. (1998). "Advantages of Query Biased Summaries In Information Retrieval", *Proceedings of the 21<sup>st</sup> International Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 2-10. New York: Association for Computing Machinery.
- Young, S.R. & Hayes, P.J. (1985). "Automatic Classification and Summarisation of Banking Telexes", *Proceedings, Second Conference on Artificial Intelligence Applications*, pp. 402-408. New York, NY: Institute of Electrical and Electronics Engineers, 1985.