

Resúmenes Automáticos

Enfoque extractivo y evaluación

Víctor Márquez Gil

Escuela Politécnica Superior,
Universidad Autónoma de Madrid.

victor.marquez@uam.es

Contenidos

1. Introducción
2. Tipos de resúmenes
3. Enfoque de extracción
4. Evaluación
5. Trabajo futuro
6. Conclusiones
7. Referencias

Introducción

1. Introducción

Objetivo

Definición

Motivación

Campos relacionados

Estrategias de resumen

Arquitectura

2. Tipos de resúmenes

3. Enfoque de extracción

4. Evaluación

5. Trabajo futuro

6. Conclusiones

7. Referencias

Introducción

Objetivo

Dar a conocer el **estado del arte** de:

- los **métodos extractivos**
- la **evaluación**

En el campo de los **resúmenes automáticos**.

Introducción

Definición

Resumir automáticamente es crear un **artefacto software** capaz de:

- tomar una fuente de **información**
- identificar el contenido **relevante**
- presentar dicho contenido al destinatario de manera **condensada**

Introducción

Motivación

- Los **resúmenes** están en todas partes: titulares, trailers, resultados de partidos, en artículos científico-técnicos y libros, etc.
- Vivimos en la **Sociedad de la Información y el Conocimiento**
- **Internet** crece vertiginosamente
- Problema : **sobrecarga**
- Aún así hay que tomar **decisiones**
- **Necesidad**: resúmenes automáticos

Introducción

Campos relacionados

- **Compresión de texto:** condensar el texto para mayor eficiencia de *almacenamiento* y *transmisión*
- **Indexación:** extraer *términos relevantes* para la *recuperación* de información
- **Minería de datos:** detección de información *nueva* o *anómala* para caracterizar *singularidades*

Introducción

Estrategias de resumen

- **Estrategia extractiva:** sólo material *copiado literalmente* del documento fuente
- **Estrategia abstractiva:** *parte* del material presente en el resumen *no se encuentra* en el documento fuente

Introducción

Arquitectura (I)

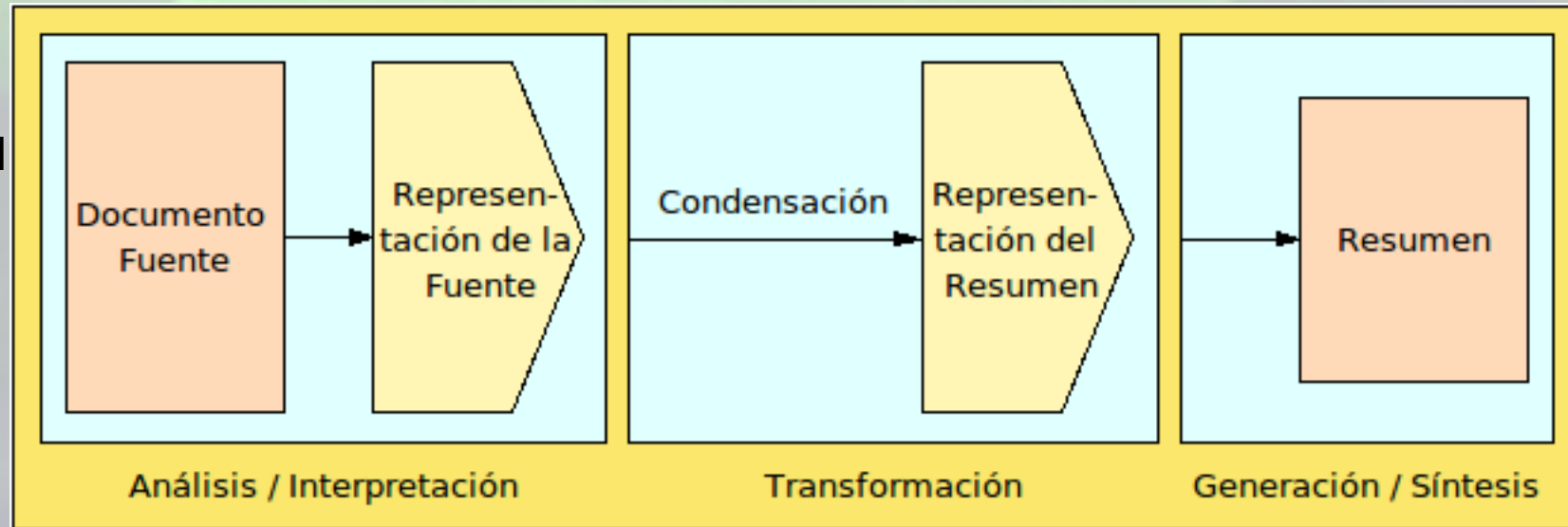
Los sistemas constan de 3 procesos:

- **Análisis:** fuente → representación fuente
- **Transformación:** representación fuente → representación resumen
- **Síntesis:** representación resumen → resumen

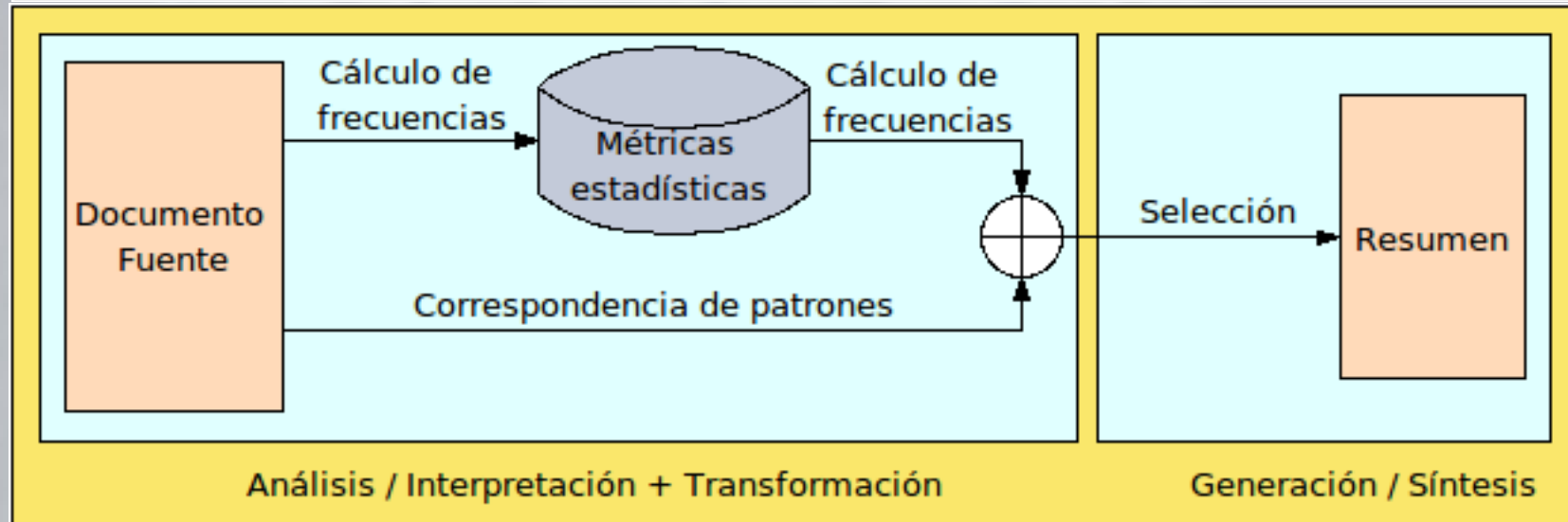
Introducción

Arquitectura (II)

Arquitectura general de los sistemas de resúmenes automáticos



Arquitectura particular de los sistemas de resúmenes automáticos extractivos



Tipos de resúmenes

1. Introducción

2. Tipos de resúmenes

Indicativos/Informativos/Críticos

Genéricos/Orientados al usuario

Un documento/Multi-documento

Multimedia

3. Enfoque de extracción

4. Evaluación

5. Trabajo futuro

6. Conclusiones

7. Referencias

SUMMARIZATION

Tipos de resúmenes

Indicativos/Informativos/Críticos

- **Indicativos:** dan una idea de los *temas relevantes* del texto para que el usuario decida si leer la fuente completa
- **Informativos:** son *sustitutos* de la fuente
- **Críticos:** aportan la *opinión* del escritor del resumen

Tipos de resúmenes

Genéricos/Orientados al usuario

- **Genéricos:** *no están hechos a medida* de ninguna audiencia y el propósito es general
- **Orientados al usuario:** *se adaptan a las necesidades de un usuario*, a través de un modelo de usuario o un simple campo de formulario para realizar una consulta sobre el texto

Tipos de resúmenes

Un documento/Multi-documento

- **Un documento:** el documento fuente a resumir es solamente uno
- **Multi-documento:** se toman varias fuentes de texto y se resumen en una sola

Tipos de resúmenes

Multimedia

- **Multimedia**: se pueden resumir *vídeos*, *imágenes*, grabaciones de *audio* de reuniones o *diagramas* y *combinaciones* de estos elementos como por ejemplo el trabajo de **Merlino & Maybury (1999)** con noticias de informativos de televisión

Enfoque de extracción

1. Introducción

2. Tipos de resúmenes

3. Enfoque de extracción

Orígenes

Métodos estadísticos

Aprendizaje supervisado

Aprendizaje no supervisado

Extracción de hechos

Ventajas e inconvenientes

Revisión

4. Evaluación

5. Trabajo futuro

6. Conclusiones

7. Referencias

Enfoque de extracción

Orígenes

- El primer trabajo es el de [Luhn \(1959\)](#): utiliza **frecuencias de términos** para determinar las oraciones más relevantes del documento fuente.
- Le siguió [Edmundson \(1969\)](#): además añade las características de **expresiones clave**; palabras que aparecen en **títulos y subtítulos**; y la **posición** de la oración en el párrafo.

Enfoque de extracción

Métodos estadísticos

- Los trabajos posteriores seguían el mismo acercamiento que [Luhn \(1959\)](#) y [Edmundson \(1969\)](#) añadiendo nuevas características y aplicándolos a dominios específicos.
- También se han barajado *otras unidades* a extraer en lugar de oraciones como sintagmas, n-gramas u otras ventanas de texto.

Enfoque de extracción

Aprendizaje supervisado

- Se determina la *importancia de las características* mediante un **corpus** de un determinado *género*.
- En el corpus **etiquetados** se empareja un texto fuente con su resumen generado por un humano.
- El sistema puede **aprender** nuevas reglas.
- Ejemplos: [Kupiec et al. \(1995\)](#) y [Mani & Maybury \(1999\)](#)

Enfoque de extracción

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.
- *Características del resumen*, no de la oración: *longitud* del resumen, *orden*, *relación* con perfil de *usuario*, etc.

Enfoque de extracción

Extracción de hechos

- Otros métodos de extracción consisten en rellenar **plantillas** predefinidas con hechos extraídos del documento fuente.
- Solo permiten *un punto de vista*
- Están muy **ceñidos a dominios específicos**
- Ejemplo: **Young & Hayes (1985)** trabaja con telexes bancarios.

Enfoque de extracción

Ventajas e inconvenientes

• Ventajas:

- *Bajo coste* humano, económico y computacional
- Implementación *fácil*
- Consistente y *evita subjetividad*
- *Mejores resultados*

• Inconvenientes:

- Falta de *coherencia*
- *Redundancia*

Enfoque de extracción

Revisión

- La **incoherencia** se da por: *anáforas* no resueltas, *lagunas*, o *entornos estructurados*.
- Se pueden solucionar en algunos casos mediante **revisión** del resumen generado.
- Añadiendo *ventanas de texto* o *eliminando* oraciones con anáforas.
- La **redundancia** se soluciona mediante *MMR*.

Evaluación

1. Introducción
2. Tipos de resúmenes
3. Enfoque de extracción

4. Evaluación

Orígenes

Clasificación

Programas

5. Trabajo futuro
6. Conclusiones
7. Referencias

SUMMARIZATION

Evaluación

Evaluación en la Ciencia

- Parte del **método científico**
- **Evaluar** resultados → construir *argumento a favor* o en *contra* de una *teoría* o *método*
- Prueba para *confirmar* o *refutar* **hipótesis**
- Ayuda a dar lugar a *nuevas hipótesis*
- Proporciona:
 - Estrategia de **investigación**
 - Marco **teórico**

Evaluación

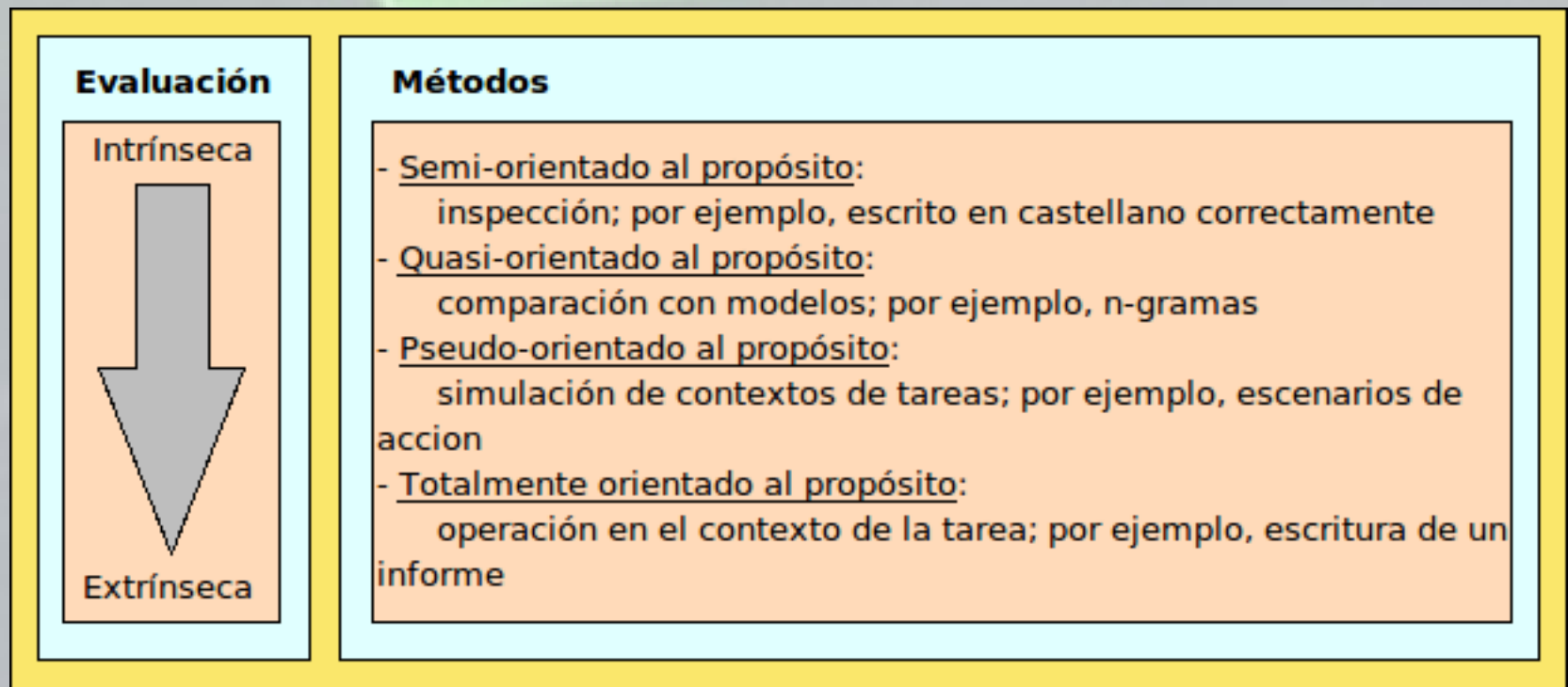
Orígenes

- Métodos **informales** (Pollock & Zamora, 1975)
- Estudios **organizados** (Edmundson, 1969)
- **Comparativas** (Brandow *et al.*, 1995)
- **Programas** de evaluación *SUMMAC*, *DUC*, etc.

Evaluación

Clasificación

- **Intrínseca:** valora la *calidad* del resumen
- **Extrínseca:** centrado en *propósito* de una *tarea*



Gradiente de la evaluación relacionada con el contexto de la tarea

Evaluación

Programas

- **SUMMAC** (1998): primera *evaluación* a gran escala. Valoración *extrínseca*.
- **NTCIR** (2000-2004): *extrínseca* (valoración de la relevancia) e *intrínseca* (contra modelo).
- **DUC** (2000-2007): progresión *intrínseca* → *extrínseca*.
- **TAC** (2008-2010): valoración *manual* del contenido con método *piramidal* (TAC 2010)

Trabajo futuro

1. Introducción
2. Tipos de resúmenes
3. Enfoque de extracción
4. Evaluación

5. Trabajo futuro

6. Conclusiones
7. Referencias

With the **SUMMARIZATION** rapid growth of the World Wide Web and on-line information services, more and more information is available and accessible on-line. This explosion of information has resulted in a well-recognized information

Trabajo futuro (I)

- Tener en cuenta los **factores de contexto**

• 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

Trabajo futuro (II)

- Métodos **híbridos**
- Análisis **lingüísticos** más ligeros *computacionalmente* y más profundos *conceptualmente*
- Tener más en cuenta el **contexto**
- *Aprovechar* más los **recursos** como WordNet y EuroWordNet

Conclusiones

1. Introducción
2. Tipos de resúmenes
3. Enfoque de extracción
4. Evaluación
5. Trabajo futuro

6. Conclusiones

7. Referencias

With the **SUMMARIZATION** rapid growth of the World Wide Web and on-line information services, more and more information is available and accessible on-line. This explosion of information has resulted in a well-recognized information

Conclusiones

- Imposible hacer un ranking de métodos
- Enfoque **extractivo**: buenos *resultados*, *fácil* de implementar, *coste* bajo
- **Características**: *ubicación* y *palabras clave* más efectivas; y otras en dominios específicos
- Tender hacia el **entendimiento del texto**
- **Evaluación**: impulsora de los *avances*
- El campo de los resúmenes automáticos es una **disciplina práctica**, se debe establecer un **marco teórico**

Referencias

1. Introducción
2. Tipos de resúmenes
3. Enfoque de extracción
4. Evaluación
5. Trabajo futuro
6. Conclusiones
- 7. Referencias**

With the **SUMMARIZATION** rapid growth of the World Wide Web and on-line information services, more and more information is available and accessible on-line. This explosion of information has resulted in a well-recognized information

Referencias

- Alfonseca, E. & Rodríguez P. (2003). "Generating Extracts with Genetic Algorithms", Advances In Information Retrieval, vol. 2633, pp. 511-519.
- 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.
- Edmundson, H.P. (1969). "New Methods in Automatic Extracting", Journal of the Association for Computing Machinery, vol. 16, no. 2, pp 264-285.
- Kupiec, J.; Pedersen, J. & Vhen, F. (1995). "A Trainable Document Summarizer", Proceedings of the 18th ACM-SIGIR Conference, pp. 68-73.
- 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. & Maybury, M. editors (1999). "Advances in Automatic Text Summarization", Cambridge: Massachusetts: MIT Press
- 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.
- 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.
- 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.