Creado el 21/01/2014
Última versión: 21/01/2014
Laura Plaza, UNED
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

**Aplicaciones de las TLH/COMPRENSIÓN Y GENERACIÓN DE TEXTO**

# GENERACIÓN DE RESÚMENES (y 2)

## Transcripción del vídeo

Hola, soy Laura Plaza, profesora de la Universidad Nacional de Educación a Distancia y miembro del grupo de investigación de Recuperación de Información y Procesamiento de Lenguaje Natural de esta universidad.

Os doy la bienvenida a la lección de generación automática de resúmenes correspondiente a la unidad de Comprensión y Generación de texto.

¿Qué vamos a ver en esta lección?

En primer lugar, definiremos y motivaremos el problema de la generación automática de resúmenes.

A continuación, repasaremos brevemente las soluciones y técnicas que tradicionalmente se han adoptado para abordar este problema.

Y finalmente, veremos qué criterios podemos utilizar para evaluar la calidad de los resúmenes generados de forma automática.

Para estudiar cómo generar resúmenes automáticamente, es imprescindible saber qué se entiende por resumen: un resumen es un texto producido a partir de uno o varios documentos, que conserva la información importante de los originales, eliminando la información irrelevante, y que presenta una longitud significativamente inferior a la de los documentos originales.

La generación automática de resúmenes surge con la intención de hacer frente a la sobrecarga de información, como una manera de seleccionar automáticamente aquella que realmente interesa al usuario y descartar la que es irrelevante o redundante.

La generación automática de resúmenes es una disciplina que cuenta con más de medio siglo de vida, pero no es hasta finales del siglo XX cuando realmente se populariza, y esto se debe al auge experimentado por Internet y al aumento exponencial de la información disponible, que hace imposible al usuario procesar toda ella en un tiempo razonable.

Un resumen (ya sea generado de forma automática o manual) puede clasificarse de acuerdo a múltiples tipologías. Vamos a ver algunas de ellas:

- Por un lado, podemos distinguir entre resúmenes generados por extracción y resúmenes generados por abstracción. Los primeros se generan a base de extraer y concatenar oraciones completas del texto original, mientras que los segundos implican la realización de un proceso de abstracción, compresión y reescritura.
- Por otro lado, se distingue entre resúmenes indicativos y resúmenes informativos. Los resúmenes indicativos proporcionan al usuario una función de referencia para seleccionar documentos para una lectura más profunda, mientras que los resúmenes informativos cubren toda la información esencial de los textos de entrada, actuando como sustitutos de éstos.
- Dependiendo del número de documentos que intervienen en la generación del resumen cabe hablar de resúmenes mono-documento y multi- documento. Estaremos ante un resumen mono-documento si en su elaboración todo el material procede de una única fuente o documento. Por el contrario, estaremos ante un resumen multi-documento si su contenido procede de distintos documentos que tratan sobre un mismo tema o suceso.
- Finalmente, una clasificación frecuente es aquella que distingue entre resúmenes genéricos y resúmenes adaptados al usuario. Mientras que los resúmenes genéricos recogen los temas principales del documento y van destinados a un amplio y heterogéneo grupo de personas, los resúmenes adaptados al usuario se confeccionan teniendo en cuenta las preferencias o características del usuario a la hora de seleccionar los contenidos del resumen y la forma de presentación.

Conviene realizar más hincapié en la primera taxonomía, la que distingue entre resúmenes por extracción y resúmenes por abstracción, ya que esta distinción es especialmente relevante cuando se quieren generar resúmenes de manera automática.

En la diapositiva, a la izquierda, se muestra un fragmento de una noticia periodística; y a la derecha, su correspondiente resumen generado por extracción. Vemos cómo el resumen se compone de las oraciones primera y última del texto original, extraídas tal cual y sin ninguna modificación, y concatenadas para formar el resumen.

Por lo tanto, un método automático de generación de resúmenes por extracción actuará de acuerdo al siguiente proceso general:

- En primer lugar, se divide el texto en oraciones.
- A continuación, se puntúan individualmente las oraciones de acuerdo con algún criterio o criterios que midan la importancia relativa de las oraciones.
- En tercer lugar, se seleccionan las oraciones mejor puntuadas (en un número que dependerá de la longitud que se desee que tenga el resumen).
- Finalmente, se ordenan las oraciones y se concatenan para componer el resumen final.

En cuanto a las técnicas y soluciones empleadas para generar resúmenes por extracción, cabe agruparlas en tres categorías que veremos en detalle:

- Los enfoques superficiales
- Las técnicas discursivas
- Los enfoques basados en grafos

Los enfoques superficiales puntúan las oraciones utilizando heurísticas sencillas que tratan de estimar la relevancia de las oraciones para la audiencia objetivo del resumen. Entre las heurísticas

más utilizadas están la frecuencia de los términos del documento y la posición de las oraciones en el documento:

- La heurística de la frecuencia de los términos en el documento se basa en el uso de técnicas estadísticas y de recuperación de información para estimar la importancia relativa de las palabras que aparecen en el documento. De esta manera, posteriormente, se seleccionan para el resumen aquellas oraciones que contengan las palabras más importantes.
- La heurística de la localización de las oraciones en el texto se basa en la evidencia empírica de que, en algunos tipos de documentos, la posición de una oración en el texto es indicativa de su relevancia. Así, por ejemplo, las noticias periodísticas se escriben siguiendo una estructura de pirámide invertida, según la cuál los datos de mayor interés se incluyen en primer lugar y, a continuación, se desarrollan aspectos secundarios. Siguiendo esta idea, se asigna mayor puntuación a las oraciones que ocupan las primeras posiciones del documento.

La segunda categoría de técnicas son las técnicas discursivas. Algo más complejas que las anteriores, se basan en un análisis de la estructura del discurso y en la detección de distintos tipos de relaciones entre palabras.

Dentro de estas técnicas, destaca por su difusión el enfoque de las cadenas léxicas. Este método consiste en recorrer el texto buscando enlaces entre oraciones, de tal modo que las oraciones que se encuentren enlazadas conforman una cadena léxica. Cada cadena se identifica con un tema dentro del documento.

Dos oraciones se consideran enlazadas si entre sus palabras se establecen relaciones, pudiendo ser estas relaciones muy diversas (por ejemplo, de sinonimia, homonimia o repetición).

Finalmente, y en lugar de extraer oraciones para el resumen, se extraen cadenas completas. Para ello, se puntúan las cadenas de acuerdo a distintos criterios como su longitud o la frecuencia de sus palabras.

Otra técnica discursiva muy popular es la llamada Teoría de la Estructura Retórica de Mann y Thompson. No se explicará en esta lección, pero se anima al alumno a leer la referencia bibliográfica correspondiente.

Finalmente, las técnicas basadas en grafos representan el texto como una red compleja en la que los nodos representan cada una de las unidades textuales en las que se divide el texto y las aristas representan algún tipo de relación entre estas unidades, generalmente de naturaleza léxica, sintáctica o incluso semántica.

La idea subyacente en este tipo de enfoques es la emergencia en la red de grupos de unidades que guardan estrecha relación entre sí y que determinan la información relevante del documento.

Para ello, se aplica un algoritmo de clustering encaminado a localizar los nodos centrales del grafo en función del número y peso de las aristas que confluyen en los distintos nodos. Finalmente, se seleccionan para el resumen las oraciones más “centrales” del documento.

Pasamos ahora a la generación de resúmenes por abstracción.

En la diapositiva, a la izquierda, se muestra el mismo fragmento de noticia periodística que veíamos anteriormente; y a la derecha, su correspondiente resumen generado por abstracción. Vemos cómo

ahora el resumen se compone de material que no se encuentra explícitamente en el original, sino que ha sido comprimido y reescrito para producir un resumen más compacto y legible.

Indudablemente, los resúmenes generados por abstracción son potencialmente más breves y de mayor calidad gramatical que los generados por extracción. Sin embargo, su generación automática es claramente más compleja y únicamente aplicable a dominios muy concretos y reducidos.

Un método automático de generación de resúmenes por abstracción actuará de acuerdo al siguiente proceso general:

- En primer lugar, se divide el texto en oraciones.
- A continuación, se construye una representación semántica de las oraciones.
- Se realizan operaciones de selección del contenido relevante, agregación y generalización de dicho contenido.
- Finalmente, se traduce la representación resultante a lenguaje natural mediante el uso de técnicas de generación de lenguaje.

En cuanto a las técnicas y soluciones empleadas para generar resúmenes por abstracción, el enfoque más popular es el que utiliza plantillas y técnicas de extracción de información, que veremos en más detalle.

El enfoque de extracción de plantillas se utiliza para generar resúmenes de textos en dominios muy acotados y con una temática muy concreta y conocida de antemano.

Para ello, se utilizan unas plantillas predefinidas, que recogen la información que se considera importante, y se usan técnicas de Extracción de Información para localizar y extraer la información necesaria para rellenar los campos de la plantilla.

Algunos trabajos, además, combinan la extracción de plantillas con el uso de patrones y técnicas de Generación de Lenguaje Natural para redactar el resumen final.

A modo de ejemplo, podemos ver en la parte derecha de la diapositiva las plantillas propuestas por Paice y Jones para extraer el contenido relevante en documentos de investigación sobre cultivos agrícolas. Este tipo de documentos se caracterizan por ser muy estructurados, y en ellos se pueden observar una organización, estilismo y semántica relativamente constante.

También se muestran en la diapositiva los patrones utilizados para generar el resumen final en lenguaje natural.

Ya hemos visto en qué consiste la Generación Automática de Resúmenes y hemos repasado algunas de las técnicas más populares. Veamos ahora cómo podemos evaluar el desempeño de los sistemas automáticos de generación de resúmenes; es decir, la calidad de los resúmenes que producen.

¿Qué criterios se utilizan para medir la calidad de un resumen?

- Por un lado, se ha de evaluar su contenido; es decir, si el resumen preserva el contenido relevante del documento original.
- Por otro lado, se ha de evaluar su forma; es decir, su legibilidad y calidad gramatical.
- Por otro lado, la evaluación puede realizarse de manera manual o automática.
- La evaluación manual es, claramente, la más exacta y fiable, ya que en ella intervienen evaluadores humanos que valoran la calidad de los resúmenes automáticos, tanto desde

el punto de vista de su contenido como de su forma. Sin embargo, este tipo de evaluación es muy costosa, y generalmente es inviable, dada la cantidad de resúmenes que han de evaluarse para obtener una medida significativa del rendimiento de un sistema automático

- Como alternativa, se han propuesto varias métricas y sistemas de evaluación automáticos, el más popular de ellos es ROUGE (Lin, 2004), y que se basan en el cálculo del solapamiento de vocabulario entre los resúmenes automáticos y otros resúmenes "modelo" generados por seres humanos. A mayor solapamiento, mejor calidad se presupone al resumen. Obviamente con estas métricas sólo se puede evaluar, de forma aproximada, el contenido informativo de los resúmenes, pero no su legibilidad.

Para terminar con la lección, te animamos a que realices las actividades y lecturas propuestas.

Así mismo, si deseas profundizar en alguno de los contenidos tratados, puedes consultar las referencias bibliográficas.

Puedes encontrar más información sobre este curso su página web y plantear tus dudas y comentarios a través de correo electrónico o Twitter.

# Referencias

Estas son las fuentes en las que me he basado para confeccionar esta lección:


- Barzilay, R. y M. Elhadad. 1997. Using Lexical Chains for Text Summarization. Proceedings of the Association for Computational Linguistics, Workshop on Intelligent Scalable Text Summarization.
- Baxendale, P.B. 1958. Man-Made Index for Technical Literature: An Experiment. IBM Journal of Research and Development, 2(4):354–361.
- Kupiec, J. et al. 1995. A Trainable Document Summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 68–73.
- Lin, C-Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. Proceedings of the Association for Computational Linguistics, Workshop: Text Summarization Branches Out.
- Mann, W. y S. Thompson. 1988. Rethorical Structure Theory: Towards a Functional Theory of Text Organisation. Text, 8(3):243–281.
- Paice, C. D. y P. A. Jones. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- Laura Plaza y Alberto Díaz. 2011. Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.
- Laura Plaza, et al. 2011. A semantic graph-based approach to biomedical summarisation. Artificial Intelligence in Medicine, 53:1-14.
- Sparck-Jones, K. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, 28(1):11–20.

# Lecturas propuestas

Si quieres ampliar más los contenidos de la lección te recomiendo que te leas estos artículos

- Principios y técnicas de generación de resúmenes
  - Das and Martins, 2007. A survey on automatic text summarization. Disponible en: http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf
  - Spark Jones, 1999. Automatic summarizing: factors and directions. Disponible en: http://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/summbook99.pdf
- Evaluación de resúmenes
  - Mani, 2001. Summarization evaluation: an overview. Disponible en: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf