Creado el 04/09/2013
Última versión: 04/09/2013
Elena Lloret Pastor, Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

moocTLH/PANORAMA ACTUAL

GENERACIÓN DE RESÚMENES

Hola, soy Elena Lloret Pastor, investigadora de la Universidad de Alicante, y os doy la bienvenida a la lección 9 "Generación de resúmenes" de la segunda unidad del curso "Nuevos Retos de las tecnologías del lenguaje humano".

En esta lección os voy a presentar de forma general los aspectos más importantes de esta tarea. Estos incluyen una serie de conceptos básicos, las dos subtareas que conforman la generación de resúmenes, que son la generación propiamente dicha y la evaluación de los mismos. Pasaremos posteriormente a conocer cuáles son las tendencias actuales, y finalmente haré referencia a los materiales adicionales para que podáis consultarlos si tenéis curiosidad y os gustaría profundizar más en los conceptos vistos aquí.

Comencemos pues, con los conceptos básicos. En primer lugar debemos tener claro qué es resumir y en qué consiste la tarea de generación automática de resúmenes. Según el diccionario de la Real Academia de la Lengua Española, resumir es reducir a términos breves y precisos lo esencial de un asunto. Esto implica por un lado, qué debemos saber qué es lo más importante de ese asunto y además, saber expresarlo de manera abreviada. Teniendo estas dos cosas en mente, la generación automática de resúmenes no es más que producir un resumen de manera automática, sin que nosotros nos tengamos que preocupar de qué aspectos son los más importantes y cómo los queremos contar o presentar.

A continuación vamos a ver un ejemplo ilustrativo de lo que sería un resumen de una noticia. Como vemos, tenemos un titular de una noticia y su texto correspondiente. En la parte derecha vemos un posible resumen generado a partir de esta noticia.

En este otro ejemplo, vemos cómo a partir de un conjunto de noticias que tratan todas ellas del mismo tema, también podemos generar un resumen, identificando la información más importante de cada noticia, asegurándonos de que no se repita la misma información en el resumen y uniéndola para formar el resumen final.

Una vez ya que sabemos qué es la tarea de generación de resúmenes, nos podemos preguntar: ¿Para qué nos puede servir?. Actualmente, la gran cantidad de información disponible, sobre todo en Internet, hace que sea prácticamente imposible e inviable poder leerla toda y procesarla de acuerdo a nuestros intereses. Las tecnologías del Lenguaje

Humano, y en particular la generación de resúmenes nos van a proporcionar herramientas y recursos para que toda esa tarea de procesar la información la podamos realizar de manera más sencilla. En concreto, si somos capaces de disponer de una herramienta que genere resúmenes automáticamente, esto nos permitirá: ahorrar tiempo de lectura, saber si una información es de interés o no, conocer los aspectos más relevantes de un asunto y estar actualizados sobre un tema. Y todo esto sin necesidad de que nosotros procesemos nada, el ordenador se encargará de todo.

Ahora que ya hemos visto los conceptos básicos, es interesante que sepamos que aunque la tarea se conoce como "Generación automática de resúmenes", en realidad dentro de ésta, podemos encontrar dos subtareas diferenciadas, pero a la vez relacionadas entre sí: la tarea de generación y la tarea de evaluación. Esto quiere decir que nos podemos centrar en investigar y proponer métodos y técnicas para generar el resumen en sí, o bien podemos abordar la tarea de evaluación de un resumen una vez que ya está generado. Ambas son tareas complicadas, debido al alto contenido de subjetividad que tiene un resumen.

Seguidamente, vamos a profundizar un poquito más en cada una de estas subtareas. Centrándonos en la generación de un resumen, lo primero que debemos tener en cuenta es que generalmente el proceso se compone de tres etapas. Una primera etapa de identificación del tópico, donde nuestro objetivo será saber de qué trata el documento. Una segunda fase de interpretación, en la que tenemos que entender ese documento, y finalmente la fase propia de generación del resumen, donde produciremos el resumen final, y si queremos utilizar palabras distintas de las de los documentos de entrada, tendremos que utilizar también, técnicas de generación del lenguaje natural.

Otro aspecto importante, que ya adelantamos anteriormente, es que no encontramos un solo tipo de resúmenes, sino muchos de ellos. Dependiendo de lo que queramos conseguir con el resumen, a quién vaya dirigido, cómo lo vamos a presentar, en qué idioma, etc. veremos que existe una amplia variedad de tipos de resúmenes, de los que mostramos aquí los más importantes.

Por último, es importante conocer qué técnicas y enfoques podemos utilizar para poder generar un resumen de forma automática. Dentro del curso, la lección 3 de la unidad 10 se centra en esto y en ella podréis conocer más detalles.

Una vez generado el resumen de forma automática, una parte muy importante, como ya hemos dicho anteriormente, es evaluarlo, es decir, determinar cómo de bueno o de malo es. Para ello, existen dos tipos básicos de evaluación. Podemos realizar una evaluación intrínseca, y evaluar el resumen en sí, en base a la información que contiene o a ciertas características, o bien podemos llevar a cabo una evaluación extrínseca, evaluando el resumen pero aplicándolo a un tarea concreta, como por ejemplo, la recuperación de información, búsqueda de respuestas, etc. De esta manera, comprobaremos si el resumen es útil para mejorar el rendimiento de la tarea en cuestión.

Al igual que en la parte de generación, en la lección 3 de la unidad 10 encontraréis más información sobre distintos criterios para evaluar un resumen, así como también las herramientas más comunes que se suelen utilizar.

Una vez que ya hemos situado la generación automática de resúmenes en contexto y hemos dado unas pequeñas pinceladas respecto a la generación y a la evaluación, en esta última parte de la lección vamos a obtener una visión actualizada de lo que se está cociendo actualmente. En primer lugar, cabe destacar que en los últimos años han aparecido nuevos escenarios donde la generación de resúmenes ha cobrado una gran relevancia. Tales escenarios hacen posible que podamos generar de forma automática resúmenes o sinopsis de libros, de patentes, podemos también generar resúmenes que nos sirvan para apoyar una imagen con texto, e incluso, aprovechar todos los nuevos géneros textuales que se han creado con la Web 2.0 para producir resúmenes de blogs, de reseñas o reviews, así como de tweets.

Estos nuevos escenarios, lo que nos han permitido también es que se originen nuevos tipos de resúmenes.

Si partimos de los tipos presentados anteriormente, en los últimos años, las tendencias de investigación se han orientado principalmente a generar resúmenes a partir de textos de la Web (hipertexto), en diferentes idiomas (multilingües y cross-lingües), personalizados según las necesidades y preferencias de los usuarios, resúmenes de opinión, para que sirvan de ayuda a los usuarios en el proceso de toma de decisiones, o también de actualización, para mantener al usuario informado acerca de los hechos más recientes y relevantes de un tema, similar a tener la "última hora" de una noticia, por ejemplo.

Pero además, han aparecido otros nuevos tipos, como son los resúmenes multigénero, los micro-resúmenes, o los resúmenes que sirven para generar artículos para la Wikipedia de forma automática.

Un resumen multigénero es aquel que ha sido generado a partir de información de varios documentos que tratan sobre lo mismo, pero con la particularidad de que estos documentos pertenecen a géneros textuales distintos, donde cada uno tiene unas características y no tienen por qué seguir un formato estándar ni homogéneo (por ejemplo, tweets, comentarios de un blog, y reseñas). No debemos confundirlos con los resúmenes multidocumento, que se generan a partir de varios documentos del mismo género, como pueden ser las noticias periodísticas. Como vemos en el siguiente ejemplo, imaginemos que queremos generar un resumen sobre el "iphone". Para ello, en la Web encontramos mucha información, procedente de blogs, de páginas de opiniones, de páginas que comparan productos, de Twitter, etc. El sistema de generación de resúmenes deberá seleccionar de cada sitio la información común entre todos los documentos, determinar la más relevante, y deberá finalmente generalizar y combinar dicha información para obtener algo de estilo a esto.

Otro tipo de resúmenes novedosos, son los micro-resúmenes, que consisten en producir un resumen muy corto, del estilo de un titular de prensa, pero que normalmente los trabajos recientes adaptan estos titulares para que quepan en un tweet (140 caracteres), puesto que

actualmente, Twitter se ha convertido en una de las redes sociales más usadas y de mayor relevancia a nivel mundial.

Finalmente, también se ha aplicado la generación de resúmenes para crear artículos de la Wikipedia. Esto consistiría en coger información de varias fuentes, generar el resumen más o menos largo con la información más importante o los hechos y eventos que se deseen destacar y presentarlo en forma de un artículo de la Wikipedia. Trabajos realizados en relación a esto han demostrado que esta tarea es factible, añadiendo incluso entradas a la Wikipedia generadas de forma automática, sin que sean rechazadas por parte de los revisores.

Con esto, hemos finalizado esta lección. Si quieres ampliar tus conocimientos sobre este tema tan interesante y con tanta aplicabilidad en la actualidad, puedes consultar los materiales complementarios y la bibliografía recomendada. Además, te invitamos a ver la lección 3 de la unidad 10 y a realizar las actividades propuestas para que afiances los conceptos aprendidos.

También puedes encontrar más información sobre este curso a través del correo electrónico en la dirección moocTLH@gmail.com, Twitter (@moocTLH) o en Google Plus. Finalmente, te agradecemos tu interés por la lección y la atención prestada.