Índice.

Hemos divido el trabajo en estas secciones de las que hablaremos a continuación

Introducción

Podemos encontrar las primeras investigaciones sobre sistemas de resumenes de texto a finales de los años 50 de mano de Luhn en su paper “La creación automática de resúmenes de literatura”.

El resumen de textos está incluido en lo que llamamos las “Tecnologías del lenguaje humano” “HLT” en inglés. Este conjunto de tecnologías tiene como objetivo permitir la comunicación de las personas con las máquinas utilizando habilidades de comunicación natural.

Con el rápido crecimiento de internet, este tipo de tecnologías se han vuelto cada vez más importantes por la inmensa cantidad de información en diferentes formatos (texto, vídeo, imágenes) que podemos encontrar. Para cada uno de estos tipos tenemos que trabajar de forma diferente para extraer su información, lo que complica aún más si cabe la tarea.

**Evaluación de resumen de texto**

Dos métodos

* Intrínseco: evalúa el resumen (por ejemplo, el contenido)
  + Evaluar la calidad: respecto a diferentes criterios como la gramaticalidad.
  + La información.
  + Fidelidad de la fuente: si el resumen tiene los mismos conceptos relevantes que el documento fuente. Problema -> como dar con los conceptos claves.

El enfoque intrínseco más común es evaluar la información del resumen comparándolo con uno hecho por un humano.

* Extrínseco: probar la eficacia de un resumen.
  + The Shannon Game: cuantificar contenido adivinando fichas para intentar recrear el documento fuente.
  + The Question Game: comprensión del lector del resumen y la capacidad de este de transmitir el contenido.
  + The Clasification Game: determinar categorías tanto del resumen como del original y medir la correspondencia entre ellos.
  + Keyword Association: hay una lista de palabras claves y se verifica si los resúmenes poseen estas palabras.
  + Reading comprehension: después de leer el resumen, responder una prueba de opción múltiple.

**Evaluación de la información**

Métricas de recuperación de información (vistas en clase)

* Precisión
* Recall
* F-media

Al usar estas métricas se pueden juzgar mal dos extractos buenos para un mismo documento.

Se proponen entonces:

* Utilidad: varios jueces clasifican cada oración dándole un valor de 1 a 10 para que sea útil para el resumen. Cuanto más alto en la clasificación mejor para el resumen.
* Factoid score(unidad atómica de información que representan el significado de una oración): usa varios resúmenes como patrón y mide la superposición entre ellos.
* Pyramid: similar pero los resúmenes hechos por humanos. La información SCU (Summary Content Units) tiene un peso donde se determina la importancia.

Factoid score y Pyramid requieren un esfuerzo humano muy importante.

Herramientas para evaluar los resúmenes de forma automática que pueden ser automáticos o semi-automáticos.

* ROGUE(Recall-Oriented Understudy for Gisting Evaluation)
* QARLA: proporciona QUEEN, KING y JACK
* Elementos Básicos (BE): unidades muy pequeñas de contenido de una oración intentando mejorar los inconvenientes de comparar palabras de n-gramas fijos. Tripletes de palabras que consisten en una cabeza y un modificador o argumento.
* Gramáticas de texto (text grammars): forma de describir una estructura de texto válida de manera formal por medio de las relaciones entre oraciones (microestructura) y la estructura del texto entero(macroestructura). Tiene tres medidas de puntuación: la información desinformación y t-gramaticalidad. Aplicando este método a resúmenes automáticos y humanos, los humanos obtienen mejores puntuaciones.
* AutoSummENG: tiene un alto paralelismo con los juicios humanos. Se diferencia de los demás métodos por el tipo de información estadística extraído, representación extraída y el método para calcular la similitud; se construyen primeros gráficos de caracteres de n-gramas y luego se compara sus representaciones para establecer un grado de similitud.
* DEPEVAL: basada en la dependencia. Parecido a BE.
* GEMS:
* HowNet eval.: método de evaluación para otros idiomas, como el chino.

**Evaluación de la calidad**

Los métodos automáticos solo evalúan la calidad según su contenido, pero no por su coherencia o su redundancia.

Para tener en cuenta otros aspectos, se han diseñado algunos protocolos como:

* FAN Protocol: evaluar la calidad de un resumen independientemente del texto fuente y la información que contiene. 4 criterios
* MLUCE Protocol: permitir que los usuarios potenciales evaluen los resúmenes, dependiendo de para que querian el resumen. Tb manual.

**Dificultades o limitaciones**

Es complicado decidir si el resumen es bueno o no debido a la subjetividad que estos presentan. Es difícil decir que dos resúmenes no son buenos para un mismo documento fuente.