Resumen automático de textos basado en un modelo unificado del resumen extractivo y abstractivo.

Juan José López Condori [1]

UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

12 de diciembre de 2018

Sumario

Introducción

- El desarrollo de las Tecnologías de la Información y, especialmente, Internet ha desestabilizado por completo lo que conocíamos como producción de documentos y, por ende, todo el proceso documental.
- Una de las cuestiones que tiene planteadas la Minería de Texto, es el resumen automático de documentos.
- El resumen de textos es la tarea de condensar automáticamente un trozo de texto a una versión más corta manteniendo los puntos importantes.

Problema

- El objetivo final del resumen es salvar un tiempo de lectura necesario para localizar la información requerida en un determinado momento.El problema es que la elaboración de resúmenes consume abundantes recursos humanos.
- La mayoría de estudios se enfocan solo en uno de los tipos de resúmenes.

Objetivos

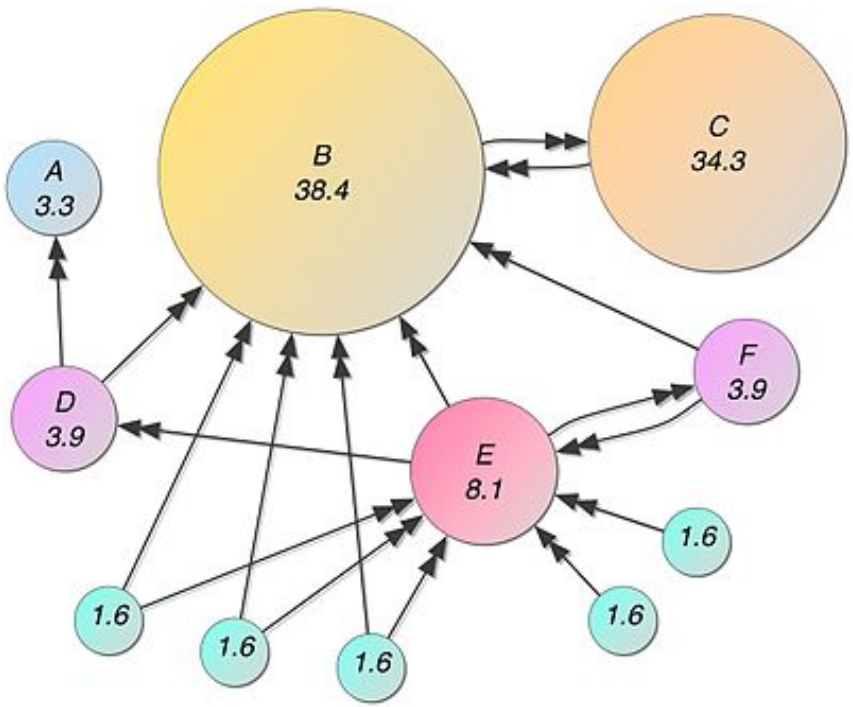
- Aplicar un modelo para unificar el resumen extractivo y abstractivo y así aprovechar sus características propias de cada uno, logrando un resumen más informativo y legible.
- Comparar las distintas técnicas del resumen extractivo y abstractivo respectivamente.
- Evaluar y comparar el método que aprovechará las carácterísticas del resumen extractivo y abstractivo.

Resumen Extractivo

- La sumarización extractiva genera un resumen seleccionando los segmentos más representativos (usualmente sentencias) de los textos fuente, sin hacer ningún cambio en los segmentos.
- En la selección de las oraciones más relevantes, los métodos extractivos utilizan un mecanismo de los rangos para obtener las sentencias con las mejores puntuaciones.

TextRank

- El TextRank, es una adaptación del PageRank a un problema muy distinto al anterior: la extracción de palabras clave y resúmenes de textos en el ámbito del lenguaje natural.



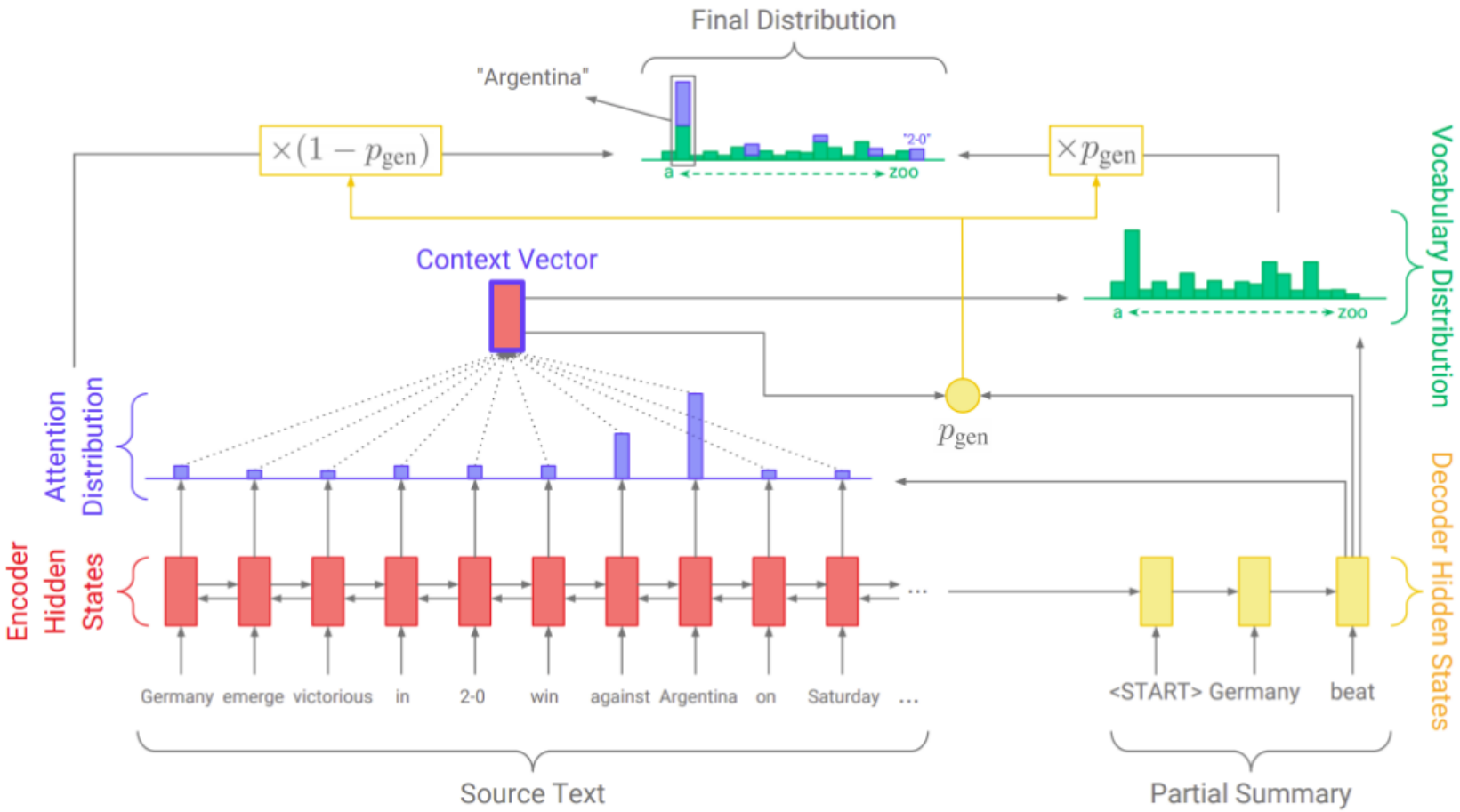
TextRank

$$TR(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{v_k \in Out(V_j)} w_{jk}} TR(V_j)$$

Resumen Abstractivo

- A diferencia del enfoque extractivo, el resumen abstractivo no sólo selecciona las sentencias de los textos fuente, analiza los documentos y automáticamente genera nuevas sentencias.
- Este enfoque intenta producir nuevos textos a partir de los fragmentos originales identificables como importantes

Pointer-Generater



Propuesta

Aplicar un método que combina explícitamente la atencion al nivel de oración β_n y a nivel de palabra α_m^t . La atención de la palabra actualizado α'^t_m es:

$$\alpha'^t_m = \frac{\alpha^t_m \times \beta_{n(m)}}{\sum_m \alpha^t_m \times \beta_{n(m)}}$$

.

Experimentos y Resultados

Para los experimentos se utilizaron los siguientes criterios:

- Evaluamos nuestros modelos en el conjunto de datos de CNN / Daily Mail que contiene noticias en los sitios web de CNN y Daily Mail.
- Cada artículo en este conjunto de datos está emparejado con un resumen de oraciones múltiples escrito por el hombre.
- Exactamente: $287,113$ pares de entrenamiento, $13,368$ pares de validación y 11,490 pares de prueba.

Experimentos y Resultados

Metodo	ROUGE-1	ROUGE-2	ROUGE-L
Lead ([23])	45.9	18.0	42.1
Mead ([25])	44.5	20.0	41.0
TextRank([19])	47.0	19.5	44.07

Cuadro 5.1: Evaluación del extractor con otros modelos mediante las puntuaciones ROUGE

Experimentos y Resultados

Metodo	ROUGE-1	ROUGE-2	ROUGE-L
HierAttn ([27])	32.75	12.21	29.01
DeepRL ([29])	39.87	15.82	36.90
GAN ([12])	39.92	17.65	36.71
Pointer-generator	39.53	17.28	36.38

Cuadro 5.2: Evaluación del abstractor con otros modelos mediante las puntuaciones ROUGE

Experimentos y Resultados

Metodo	ROUGE-1	ROUGE-2	ROUGE-L
Modelo Unificado	39.53	17.28	36.38

Cuadro 5.4: Puntuaciones ROUGE del modelo unificado

Conclusiones

- En la presente tesis utilizó un modelo que combina atenciones en dos niveles, a nivel de oración referente al resumen extractivo y a nivel de palabra referente al resumen abstractivo, aprovechando cada una de sus características.
- Se logra un modelo competitivo con los métodos de la literatura y, en varios casos, una calificación ROUGE mejor que otros estudios, siendo el resumen más informativo y legible sobre el conjunto de datos CNN / Daily Mail en una evaluación humana sólida.

Referencias

- *Ramesh Nallapati, Bowen Zhou, and Mingbo Ma, Neural architectures for extractive document summarization,2016a*
- *Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li,Incorporating copying mechanism in sequence-to-sequence learning, 2016*
- *Abigail See, Peter J Liu, and Christopher D Manning. ,Get to the point: Summarization with pointergenerator networks*
- *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin, Attention is all you need. In Advances in Neural Information Processing Systems, 2017*

Resumen automático de textos basado en un modelo unificado del resumen extractivo y abstractivo.

Juan José López Condori [1]

UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

12 de diciembre de 2018