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

Juan José López Condori

Orientador:

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Diciembre de 2018

Índice

Índice de cuadros

1. Motivación y Contexto

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.

El problema ahora no es sólo la cantidad de documentos, sino otra serie de factores añadidos como la volatilidad, la falta de tipología, las autorías de nuevo tipo (colaborativas, y otras), nuevas estructuras documentales, y un largo etcétera.

Una de las cuestiones que tiene planteadas la Minería de Texto, es el resumen automático de documentos, un tema de suma importancia que recibe continuamente contribuciones científicas.

El resumen de textos es la tarea de condensar automáticamente un trozo de texto a una versión más corta manteniendo los puntos importantes. La capacidad de condensar información de texto puede ayudar a muchas aplicaciones tales como la creación de compendios de noticias, presentando resultados de búsqueda, y generación de informes.

Hay principalmente dos tipos de enfoques: extractivo y abstractivo. Los enfoques extractivos reúnen resúmenes directamente de el texto de origen típicamente selecciona una oración completa a la vez. En contraste, los enfoques abstractivos pueden generar nuevas palabras y frases que no son copiados del texto de origen.

La mayoría de estudios [17], [25], [21], [3], [14], [9] y [15] se enfocan solo en uno de ellos, es por esto que la presente tesis describe un método unificado del resumen extractivo y abstractivo para la generación automática de resúmenes de un texto determinado.

2. Definición del Problema

El propósito de los resúmenes es facilitar y acelerar la identificación de los temas interesantes de entre una gran cantidad de documentos. El objetivo final 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 aplicación de ordenadores a esta tarea viene siendo estudiada desde hace varias décadas; ya en 1958 se empiezan a proponer sistemas de resumen automático [11] como solución al creciente número de textos técnicos publicados.

Pero es en la actualidad, sobre todo con el crecimiento espectacular de Internet, cuando se ha hecho más perentoria la necesidad de disponer de esta tecnología, lo que ha potenciado su desarrollo.

Una solución a este problema se plantea en la presente tesis que es el unificar dos enfoques de resumen: el extractivo y abstractivo para la generación de un resumen más informativo y legible.

3. Justificación

Esta tesis pretente explotar tanto las características del resumen extractivo como del resumen abstractivo, adaptando un modelo que combina atenciones en dos niveles: oración (extractivo) y palabra (abstractivo) para los datos CNN / Daily Mail especialmente en textos de noticias, además de aplicarlo para contribuir en la creación de compendios de noticias, presentando resultados de búsqueda, y generación de informes.

4. Objetivos

4.1. Objetivo General

Aplicar un modelo para unificar el resumen extractivo y abstractivo y así aprovechar características propias de cada uno, logrando un resumen más informativo y legible.

4.2. Objetivos específicos

- Comparar las distintas técnicas del resumen extractivo y abstractivo respectivamente.
- Evaluar y comparar el método que aprovechará las carácterísticas del resumen extractivo y abstractivo.

5. Trabajos Relacionados

El resumen automático de textos es la tarea que consiste en la producción automática de una versión más corta (conocida como sumario o resumen) de uno o más textos-fuente [12]. El resumen debe contener la información más relevante de los textos fuente. Muchos enfoques se han desarrollado para determinar de manera automática la información debe incluirse en el resumen. [13] y [18], los resúmenes pueden clasificarse de varias formas. En cuanto a la información que contienen, los resúmenes pueden clasificarse en tres tipos: indicativos, informativos y críticos / evaluadores. Los resúmenes indicativos contienen sólo los temas esenciales de los textos fuente, no necesariamente conteniendo detalles de resultados, argumentos y conclusiones. Por ejemplo, los índices son sumarios indicativos. Los resumenes informativos, que son los más tradicionales, se consideran sustitutos de los textos, debiendo contener toda la información principal. Los abstractos de artículos son ejemplos de este tipo de resumen. Los sumarios críticos, además de resumir el contenido de los los textos fuente, añaden crítica en función del contenido. Las reseñas de libros son ejemplos de los restos críticos. La sumarización automática también puede clasificarse

como monodocumento o multidocumento. En la primera clase, el resumen se genera automáticamente a partir de un único texto-fuente. En la segunda categoría, se produce un resumen a partir de un conjunto de textos sobre un mismo tema. En cuanto a la forma de composición de los resumenes, se tiene el resumen extractivo (cuando se seleccionan segmentos textuales enteros) y el abstractivo (cuando se realizan operaciones de reescritura). A continuación, estos enfoques se explican, así como se mencionan brevemente algunos trabajos desarrollados para estos enfoques.

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. La idea de este enfoque es extraer las oraciones que contengan mucha información y novedad. 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, es decir, las más importantes.

Se usan redes neuronales para mapear oraciones en vectores en [17] , [25] y seleccionar oraciones basadas en esos vectores. [6] , [21] y [23], además de aplicar redes neuronales recurrentes para leer el artículo. y obtener las representaciones de las oraciones y el artículo para seleccionar oraciones.

Se puede utilizar información lateral (es decir, leyendas de imágenes y títulos) para ayudar al clasificador de oraciones a elegir oraciones.[8].

Un estudio reciente [16] combina redes neuronales recurrentes con redes convolucionales gráficas para computar la importancia de cada oración. Mientras que algunos métodos de resumen extractivo obtienen altas puntuaciones de ROUGE, todos ellos sufren de baja legibilidad.

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. Con esta característica, este enfoque puede solucionar el problema de falta de cohesión de los enfoques extractivos. A pesar de no ser una cuestión nueva, hay relativamente pocos trabajos sobre la sumarización abstracta.

En [3] utiliza un codificador basado en la atención para leer el texto de entrada y generar el resumen. Basados en ellos, [15] emplea un codificador automático variacional y [22] utiliza un modelo de secuencia a secuencia más potente. Además, [22] crea un nuevo conjunto de datos de resumen de nivel de artículo llamado *CNN / Daily Mail* adaptando el conjunto de datos de preguntas y respuestas de DeepMind [9].

Se cambia el método de entrenamiento tradicional para optimizar directamente las métricas de evaluación (por ejemplo, BLEU y ROUGE) en [14] .[7], [1] y [24] combinan redes de punteros [19] en sus modelos para tratar con palabras fuera de vocabulario (OOV). Para disminuir frases repetidas en el resumen generado. [20] y [2] restringen sus modelos de prestar atención a la misma palabra.

La investigación [4] aplica el modelo de secuencia a secuencia convolucional y diseña varias tareas nuevas para el resumen. [10] logra una alta puntuación de legibilidad en la evaluación humana utilizando redes adversas generativas.

6. Propuesta

En esta presente tesis se propone un modelo unificado que combina el potencial del resumen extractivo y abstractivo. Por un lado, un simple modelo extractivo puede obtener atención a nivel de la oración con puntuaciones ROUGE altas pero menos legibles basado en el algoritmo de *TextRank*. Por otro lado, un modelo abstractivo más complicado puede obtener atención dinámica a nivel de palabra para generar un párrafo más legible. En el modelo unificado se basa en [5], la atención a nivel de oración (resumen extractivo) se usa para modular la atención a nivel de palabra (resumen abstractivo) tal que es menos probable que se generen palabras en oraciones menos concurridas y asi generar un resumen más informativo y legible.

7. Cronograma de Actividades

Ejemplo: *En el Cuadro 1 se puede observar un ejemplo de un Cronograma de Actividades.*

Cuadro 1: Posible Cronograma de Actividades para el desarrollo de la Tesis.

Actividad	Inicio Aprox.	Fin Aprox.
Elaboración del Plan de Tesis	*02-setiembre-18*	*09-setiembre-18*
Presentación y Refinación del Plan de tesis	*09-setiembre-18*	*16-setiembre-18*
Redacción de la Parte Teórica de la Tesis	*16-octubre-18*	*17-octubre-18*
Implementación parcial de las técnicas a ser usadas	*17-octubre-18*	*20-noviembre-18*
Implementación parcial de las técnicas a ser usadas	*20-octubre-18*	*20-octubre-18*
Realización de las pruebas y escritura de resultados	*20-noviembre-18*	*20-noviembre-18*
Implementación total de las técnicas a ser usadas	*20-noviembre-18*	*20-nov-18*
Presentación del Borrador de Tesis	*20-nov-18*	*04-dic-18*
Correción de observaciones al Borrador de Tesis	*04-dic-18*	*19-dic-18*
Presentación de la Versión Final de la Tesis	*19-dic-13*	*01-ene-19*

8. Indice Tentivo de la Tesis

La tesis a ser desarrollada como producto de este Plan de Tesis tendra tentantivamente el siguiente Indice de Contenidos:

1. *Resumen.*
2. *Abstract.*
3. *Introducción.*
    a) *Motivación y Contexto.*
    b) *Definición del Problema.*
    c) *Objetivos.*
        1) *Objetivo Principal.*
        2) *Objetivos Específicos.*
4. *Conceptos sobre Bases de Datos.*
5. *Trabajos Relacionados.*
6. *Propuesta.*
7. *Experimentos y Resultados.*
8. *Conclusiones.*
9. *Recomendaciones.*
10. *Trabajos Futuros.*

Referencias

[1] Abigail See, P. J. L., and Manning., C. D. Get to the point: Summarization with pointergenerator networks. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), 1073–1083.

[2] Abigail See, P. J. L., and Manning., C. D. Get to the point: Summarization with pointergenerator networks. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), 1073–1083.

[3] Alexander M Rush, S. C., and Weston, J. A neural attention model for abstractive sentence summarization. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), 379–389.

[4] Angela Fan, D. G., and Auli, M. Controllable abstractive summarization.

[5] Ashish Vaswani, Noam Shazeer, N. P. J. U. L. J. A. N. G. u. K., and Polosukhin., I. Attention is all you need. in advances in neural information processing systems. 6000–6010.

[6] Cheng, J., and Lapata, M. Neural summarization by extracting sentences and words. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), 484–494.

[7] Jiatao Gu, Zhengdong Lu, H. L., and Li, V. O. Incorporating copying mechanism in sequence-to-sequence learning. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), 1631–1640.

[8] JShashi Narayan, Nikos Papasarantopoulos, M. L., and Cohen, S. B. Neural extractive summarization with side information.

[9] Karl Moritz Hermann, Tomas Kocisky, E. G. L. E. W. K. M. S., and Blunsom, P. Teaching machines to read and comprehend. in advances in neural information processing systems. 1693–1701.

[10] Linqing Liu, Yao Lu, M. Y. Q. Q. J. Z., and Li, H. Generative adversarial network for abstractive text summarization. *n Proceddings of the 2018 Association for the Advancement of Artificial Intelligence* (2017).

[11] Luhn, H. The automatic creation of literature abstracts. *IBM Journal of Research and Development* (1958), 159–165.

[12] Mani, I. Automatic summarization . natural language processing.

[13] Mani, I., M. M. Introduction. in : Advances in automatic text summarization.

[14] Marc'Aurelio Ranzato, Sumit Chopra, M. A., and Zaremba, W. Sequence level training with recurrent neural networks. 1693–1701.

[15] Miao, Y., and Blunsom, P. Language as a latent variable: Discrete generative models for sentence compression. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), 319–328.

[16] Michihiro Yasunaga, Rui Zhang, K. M. A. P. K. S., and Radev, D. Graph-based neural multi-document summarization. *In Proceedings of the 21st Conference on Computational Natural Language Learning* (2017), 452–462.

[17] Mikael Kageback, Olof Mogren, N. T., and Dubhashi, D. Extractive summarization using continuous vector space models. *In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)* (2014), 31–39.

[18] Nenkova, A.; McKeown, K. Automatic summarization. foundations and trends in information retrieval.

[19] Oriol Vinyals, M. F., and Jaitly, N. Pointer networks. *In Advances in Neural Information Processing Systems* (2015), 2692–2700.

[20] Qian Chen, Xiaodan Zhu, Z. L. S. W., and Jiang, H. Distraction-based neural networks for modeling documents. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016).

[21] Ramesh Nallapati, B. Z., and Ma, M. Neural architectures for extractive document summarization.

[22] Ramesh Nallapati, Bowen Zhou, C. G. B. X. Abstractive text summarization using sequence-to-sequence rnns and beyond. 280–290.

[23] Ramesh Nallapati, F. Z., and Zhou, B. A recurrent neural network based sequence model for extractive summarization of documents. *Proceedings of the 31st AAAI conference* (2017).

[24] Romain Paulus, C. X., and Socher, R. A deep reinforced model for abstractive summarization. *In Proceedings of the 2018 International Conference on Learning Representations* (2018).

[25] Yin, W., and Pei, Y. Optimizing sentence modeling and selection for document summarization. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence* (2015), 1383–1389.