Document Summarization using a Scoring-Based Representation

Augusto Villa Monte*, Laura Lanzarini†
Instituto de Investigación en Informática LIDI
Facultad de Informática, Universidad Nacional de La Plata
La Plata, Buenos Aires, Argentina
{avillamonte, laural}@lidi.info.unlp.edu.ar

Luis Rojas Flores‡
Facultad de Ingeniería
Universidad Nacional de la Patagonia San Juan Bosco
Tierra del Fuego, Argentina
luisf.09@gmail.com

José A. Olivas Varela§
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ciudad Real, España
JoseAngel.Olivas@uclm.es

***Abstract*—Currently, data repositories contain a plethora of information in different formats, most of which consists of text. This situation has raised interest in the study of techniques to automate the identification o f t he m ost r elevant s entences o f a document with the goal of generating a text summary.**

This article presents a technique for extracting the most representative sentences in a document, employing a user-defined criteria. The criteria is learned by the system using an optimization technique and a training document where the user has ranked the sentences according to their relevance.

The proposed method has been applied to a five-chapter thesis with good results. At the end of this paper we provide some conclusions as well as ideas for future work.

***Keywords*—Text Summarization, Extractive Summaries, Sentence Scoring Methods, Particle Swarm Optimization.**

I. Introducción

El avance de la tecnología ha favorecido la generación de grandes volúmenes de datos provenientes del registro automático de numerosos procesos. El objetivo con el cual fueron generados estos repositorios ha ido cambiando a lo largo de los años. Lo que inicialmente fue un simple registro de la actividad realizada se convirtió en una gran cantidad de información capaz de detallar cada decisión tomada. Fue este concepto lo que llevó a buscar la manera de identificar patrones capaces de explicar el comportamiento registrado dando origen así al proceso de extracción de conocimiento que, a través de las distintas técnicas de Minería de Datos, ha tratado de dar respuesta a este problema.

Hoy en día, si bien se utiliza información en diferentes formatos, en la mayoría de los sistemas y más aún en Internet, se almacena información textual. El uso de herramientas automáticas para su procesamiento es esencial ya que sin ellas no sería posible explotar toda la información disponible, organizarla y recuperarla para tomar decisiones.

La Minería de Texto es un área de la Minería de Datos que trabaja con datos no estructurados y cuyas técnicas son capaces de procesar lenguaje natural. Dentro de este campo, la generación de resúmenes de documentos de texto es una tema de suma importancia que recibe continuamente contribuciones científicas. Tanto las técnicas de Aprendizaje Automático como las de Recuperación de Información junto con la Minería de Texto, campos estrechamente relacionados, se han adaptado con éxito para contribuir al resumen automático.

Resumir texto es un proceso a través del cual se construye una versión más corta de uno o varios documentos de texto. Esta es una acción que el ser humano sabe resolver muy bien por tener, entre otras cosas, capacidad de abstracción e interpretación, poder expresar lo mismo con otras palabras y además de manejar la ambigüedad.

La construcción del resumen de un texto en forma automática busca reducir su tamaño preservando la información. Esto permite obtener el contenido principal de un documento en menor tiempo del que llevaría hacerlo a partir del texto completo de forma manual. Hoy con la tecnología que rodea la Minería de Texto se dispone de una mayor capacidad de cómputo que la que se podría conseguir poniendo a trabajar muchos cerebros humanos en simultáneo. Sin embargo, no todo está resuelto y por ello resulta fundamental el estudio, investigación y desarrollo de este tema.

El objetivo de este artículo es aprender el criterio utilizado por una persona al resumir un texto para luego poderlo aplicar sobre otros documentos y darle como respuesta un resumen similar al que hubiera realizado en forma manual pero en un tiempo de respuesta mucho menor. Para ello se requiere contar con un breve documento resumido por la persona en cuestión, tal como lo hubiera realizado utilizando un resaltador para

*Becario de Doctorado UNLP Tipo B
*Ayudante Diplomado Dedicación Semi-Exclusiva
†Profesor Titular Dedicación Exclusiva
‡Ayudante Alumno Dedicación Simple
§Profesor Titular de Universidad
978-1-5090-1633-4/16/$31.00

marcar todas las partes que considera importantes.

Por otro lado, utilizando un conjunto de métricas se construye una representación vectorial para cada una de las sentencias que forman el conjunto de documentos a resumir. Luego, mediante una técnica de optimización se identifican las características principales de las sentencias que mejor aproximan al resumen que hubiese realizado la persona así como la manera en que se relacionan para formar el criterio buscado. Finalmente, aplicando este resultado al resto de los documentos se logra obtener el resumen completo con el mismo criterio que el utilizado para el documento modelo. Más adelante se desarrollará en qué consiste este "criterio" del que se habla.

En la sección II se mencionan brevemente varios trabajos relacionados. En la sección III se desarrolla la metodología propuesta y se describe la representación utilizada para los documentos. En la sección IV se detallan los resultados obtenidos y por último en la V se presentan las conclusiones junto con líneas de trabajo futuro.

II. Trabajos relacionados

La obtención automática de resúmenes de documentos de texto es un tema de investigación que, a pesar de haberse comenzado a desarrollar hace muchos años, aún continúa vigente. En la literatura existen numerosos trabajos relacionados, que si bien presentan diferentes vistas del mismo problema, todos tienen como objetivo principal conseguir un conjunto de frases o "ideas" que conserven la información contenida en el documento pero que sean de menor longitud que el texto original.

La generación de resúmenes tiene distintos enfoques. El primero de ellos tiene que ver con el conjunto de datos y la manera en que se los procesa. Establece si se trabaja con un único documento o múltiples documentos interpretados como uno solo o con varios documentos tratados individualmente por separado (pero con cierta relación entre ellos) [1] [2]. El segundo enfoque determina el modo de obtención del resumen, el cual puede conseguirse reescribiendo el texto por completo o extrayendo partes del mismo tal cual aparecen en él (párrafos u oraciones, hasta secuencias de palabras separadas por algún signo de puntuación que no necesariamente tiene que ser el punto). Este último enfoque se denomina resumen extractivo y es uno de los más utilizados en la literatura ya que simplifica en cierta manera el problema dejando de lado aspectos semánticos que requieren el uso de estructuras adicionales como diccionarios, tesauros y ontologías [3]. Por otro lado, hay otros enfoques no tan mencionados en la literatura que se refieren a si el resumen se consigue de manera supervizada, si se utiliza un umbral (por cantidad o porcentaje) para fijar el tamaño del resumen a conseguir o si el humano interviene en el proceso automático de generación de resúmenes [4].

En lo que refiere al resumen extractivo o basado en extracción de sentencias, algunos trabajos encaran este problema interpretando el resumen como el resultado de una operación de búsqueda sobre una base de datos [5]. En estos casos el resumen se construye determinando lo que es relevante para la consulta realizada. Muchos otros trabajos se centran en la obtención de un resumen genérico que para construirlo se seleccionan sentencias completas del documento con algún criterio. Para ello, se les asigna una especie de puntaje a través de distintos métodos de "sentence scoring". Este puntaje se calcula a través de métricas que analizan determinadas características presentes en el texto. Este enfoque fue utilizado por primera vez en 1968 [6]. Los métodos van desde identificar determinadas expresiones en el texto (como "lo más importante", "en resumen" o "el artículo describe" entre otras) hasta calcular la centralidad de la sentencia determinando la superposición entre una sentencia y las otras sentencias en el documento (vocabulario en común).

En [7] con el objetivo de asistir a los estudiantes con dificultades de lectura, además de predecir las sentencias importantes de un texto utilizando algunas de las métricas de la literatura basadas en características del texto, se proponen características que dependen del estudiante de manera que el resumen sea fácil de entender por éste.

En [8] se realizó un estudio interesante de la relación entre el resumen humano y las partes del texto que una persona seleccionaría para formar parte del resumen construido. En este artículo los autores mencionan que los algoritmos para elaborar resúmenes automáticos utilizan medidas simples sobre el texto las cuales permiten imitar la selección humana sin problemas. Sin embargo, para aumentar la precisión del algoritmo, ellos sugieren completar estas técnicas con estrategias similares a las utilizadas por los seres humanos. Ellos aseguran que los seres humanos explotan la estructura retórica del texto al identificar las frases más destacadas que usarán luego en la elaboración del resumen.

En [9] se utilizan varias métricas de la literatura y se proponen otras para determinar si al resumir texto en dos idiomas, inglés y hebreo, las diez sentencias más importantes que componen el resumen son las mismas. En [10] analizan si la calidad del resumen obtenido con combinaciones de métodos de scoring depende de la clase de documento (noticias, blogs y artículos). Los resultados que obtuvieron les permitió afirmar que determinadas técnicas son más eficaces en algunos de los contextos estudiados. Hace unos años atrás en [11] se realizó un estudio similar a estos dos últimos pero combinando ambas ideas: distintos corpus y lenguajes.

El objetivo del presente artículo se relaciona en algo con estos últimos trabajos aunque el enfoque es otro. Se trata de desarrollar un sistema que asista al usuario en la tarea de resumir, indicándole las sentencias más representativas según el criterio aprendido previamente. Dicho criterio fue aprendido a partir de un breve resumen realizado por el usuario y consiste de la selección de las métricas principales junto con su grado de participación. Ambos aspectos resultan fundamentales para determinar la importancia de cada sentencia en el resumen final.

III. Propuesta metodológica

Es importante recordar que el enfoque extractivo elegido para generar el resumen no garantiza la coherencia narrativa de las sentencias seleccionadas. Sin embargo, en cualquiera de

los casos, siempre se reduce el tamaño del documento dejando únicamente el contenido relevante. Esto tiene tres ventajas:

- se puede controlar el tamaño del resumen
- el contenido del resumen se obtiene de forma precisa
- la relación entre el resumen y el texto original es inmediata

En términos generales el proceso de resumir extrayendo partes de un texto consiste en tres grandes etapas: (1) crear una representación intermedia para el texto original, (2) valorar cada una de las sentencias a través de un puntaje y (3) seleccionar el conjunto de sentencias que formarán parte del resumen.

En la figura 1 se muestra un diagrama que resume la metodología propuesta en este artículo, la cual se desarrollará a continuación. En la subsección III-A se describirá el pre-procesamiento realizado a los documentos. En la subsección III-B se detallará la representación utilizada para las sentencias del texto. Por último en III-C se hará mención al método empleado para aprender el criterio con el que se resume.

III-A. Pre-procesamiento del texto

Toda tarea de Minería de Textos comienza con la etapa de preprocesamiento. Dentro de esta etapa la tarea más importante es la segmentación del documento y su posterior representación. Esta tarea consiste en dividir el texto en porciones más pequeñas tal cual aparecen en el mismo y con algún criterio a partir de uno o varios delimitadores. Generalmente a estas porciones del texto se las llama "sentencias" en forma genérica y se las obtiene reconociendo signos ortográficos como el punto y sus derivados (punto y seguido, punto y a parte, punto final). En el caso del punto las sentencias representan oraciones.

Una oración es una unidad del lenguaje que tiene sentido por si misma y que está formada por un verbo. También pueden utilizarse otros signos como la coma, el punto y coma, los dos puntos y los paréntesis. El uso de cada uno de ellos tendrá consideraciones especiales tal como también los tiene el punto en relación a sus otros usos: en abreviaturas, siglas y números.

Una vez que se dispone de las sentencias se procede a dividirlas en términos y almacenarlos adecuadamente de forma tal que pueda reconstruirse en cualquier momento el documento original. Esto último es fundamental para que se pueda calcular correctamente para cada sentencia cada una de las métricas descritas en la sección III-B. En principio, todas las palabras que aparecen en el texto se consideran posibles términos. Las palabras son cualquier cadena de caracteres que inicia la sentencia, que le sigue a un espacio en blanco o que termina en un espacio en blanco, entre otros posibles delimitadores.

En esta etapa, con el fin de reducir el vocabulario con el que se trabaja es frecuente eliminar con algún criterio ciertos términos. Por un lado, se suele convertir todo el texto a mayúsculas o minúsculas ya que en general resulta intrascendente el formato de la letra. Sin embargo, eventualmente surge la necesidad de considerar excepciones como los nombres propios por ejemplo. Por otro lado, se procede a eliminar palabras vacías de contenido semántico como artículos, preposiciones, conjunciones, adverbios y adjetivos. También es habitual descartar algunos verbos como "ser" y "estar" además de las palabras de menor longitud (1 a 3 caracteres). Por último, se extraen las raíces de todas las palabras (stemming). Esta es la reducción de términos más significativa ya que el número de palabras derivadas de la misma raíz es muy elevado. Sin embargo, si bien la raíz de un término aporta mayor contenido semántico, utilizarla aumenta el ruido y genera una vez más la necesidad de considerar excepciones. No sólo por haber términos que poseen la misma raíz con diferente significado sino por existir situaciones en las que, por ejemplo, los sufijos de género son significativos (como lo es por ejemplo al analizar la posición de los medios periodísticos respecto a la violencia de género).

III-B. Representación de los documentos

Existen muchas maneras de caracterizar las sentencias de un documento. En este artículo se han seleccionado de la literatura 17 métricas para calificar cada sentencia y representar con ellas a cada documento como un conjunto de vectores numéricos. Estos valores están basados en cantidades calculadas a partir de frecuencias, posiciones, longitudes y palabras en común que tienen las sentencias en el texto. En la tabla I se define cada una de ellas en base a la sentencia S_i de un documento D. De esta manera cada documento estará representado por una matriz de tantas filas como sentencias tenga dicho documento y tantas columnas como métricas se utilicen (en este caso 17).

III-C. Aprendizaje del criterio del usuario

La preferencia del usuario por una sentencia viene dada por la puntuación que le asigne. Se trata de un valor entero positivo proporcional al grado de importancia estimado. Aquellas sentencias que reciban 0 como puntaje serán interpretadas como irrelevantes mientras que las que reciban los valores más altos serán las más significativas.

Por otro lado, dado que se cuenta con varias métricas calculadas para cada sentencia del documento se espera que una combinación lineal de ellas represente el criterio del usuario. Por lo tanto, el problema a resolver consiste en hallar los coeficientes $c_1, c_2, \ldots, c_k$ tales que aplicados a las métricas de la sentencia S_i, $m_{i1}, m_{i2}, \ldots, m_{ik}$, permitan aproximar el puntaje indicado por el usuario tal como se indica en la ecuación 1 donde $puntaje_i$ es el puntaje indicado por el usuario para la sentencia i.

$$puntaje_i = \sum_{t=1}^{k} (c_t * m_{it}) \tag{1}$$

Para resolver esto se utilizó la técnica de optimización PSO (Particle Swarm Optimization), una metaheurística poblacional propuesta por Kennedy y Eberhart [20] donde cada individuo de la población, denominado partícula, representa una posible solución del problema y se adapta siguiendo tres factores: su conocimiento sobre el entorno (su valor de aptitud), su conocimiento histórico o experiencias anteriores (su memoria)

Tabla I
DETALLE DE LAS MÉTRICAS CONSIDERADOS EN ESTE TRABAJO

Tipo	Métrica	Descripción	Fuente
Posición	POS_L $(S_i) = i$ POS_F $(S_i) = \frac{1}{i}$ POS_B $(S_i) = \text{máx}\left(\frac{1}{i}, \frac{1}{n-i+1}\right)$	Estas métricas miden la cercanía de la sentencia S_i al comienzo del documento, al final y a los bordes respectivamente siendo n el número total de sentencias de D e i un número entre 1 y n asignado a cada sentencia secuencialmente según su aparición dentro del documento de principio a fin.	[12]
Longitud	LEN_W $(S_i) = \lvert words(S_i)\rvert$ LEN_CH $(S_i) = \lvert characters(S_i)\rvert$	Estas métricas miden la cantidad de palabras y caracteres que posee la sentencia S_i respectivamente. Es decir que $\lvert.\rvert$ indica la cardinalidad del conjunto.	[13]
Frecuencia	LUHN $(S_i) = \frac{\lvert keywords(c_i)\rvert^2}{\lvert c_i\rvert}$	Donde c_i es la secuencia más grande de palabras consecutivas que comience y termine con palabras consideradas clave con algún criterio para el documento.	[14]
	KEY $(S_i) = \sum_{k \in keywords(S_i)} tf_k$	La métrica suma las frecuencias de todas las palabras clave que la sentencia S_i contenga, siendo tf_k la frecuencia de la palabra clave k en el documento.	[6]
	COV $(S_i) = \frac{\lvert keywords(S_i)\rvert}{\lvert keywords(D)\rvert}$	Mide la proporción de palabras clave del documento D contenidas en la sentencia S_i.	[15]
	TF $(S_i) = \frac{\sum_{w \in words(S_i)} tf_w}{\lvert words(S_i)\rvert}$	Calcula la frecuencia promedio de las palabras de la sentencia S_i.	[16]
	TFISF $(S_i) = \sum_{w \in words(S_i)} tf_w \times isf_w$	$isf_w = 1 - \frac{log(n_i)}{log(n)}$ donde n_i es el número de sentencias que contienen la palabra i.	[17]
	SVD $(S_i) = \sqrt{\sum_{j=1}^{k} diag(S,j)^2 \times V(j,i)^2}$	Donde S y V son las matrices obtenidas de descomponer en valores singulares la matriz de frecuencia binaria de n filas (sentencias) por tantas columnas como términos tenga el documento. k es el número de conceptos o tópicos del documento representados por los k valores singulares en orden descendiente.	[18]
Título	TITLE_O $(S_i) = \frac{\lvert words(S_i) \cap words(T_i)\rvert}{\text{mín}\left(\lvert S_i\rvert, \lvert T_i\rvert\right)}$ TITLE_J $(S_i) = \frac{\lvert words(S_i) \cap words(T_i)\rvert}{\lvert words(S_i) \cup words(T_i)\rvert}$ TITLE_C $(S_i) = \frac{\vec{S_i} \times \vec{T_i}}{\lvert\vec{S_i}\rvert \times \lvert\vec{T_i}\rvert}$	Estas métricas miden la similitud de la sentencia S_i con el título T_i asociado a ella utilizando tres medidas: superposición, Jaccard y coseno respectivamente. En todos los casos la similitud se define a través de las palabras en común que posean la sentencia y el título. Sin embargo para calcular coseno se necesita una matriz de frecuencia de dos filas, una para el título T_i y otra para la sentencia S_i, por tantas columnas como palabras distintas tengan entre las dos.	[6]
Cobertura	D_COV_O $(S_i) = \frac{\lvert words(S_i) \cap words(T_i)\rvert}{\text{mín}\left(\lvert S_i\rvert, \lvert D - S_i\rvert\right)}$ D_COV_J $(S_i) = \frac{\lvert words(S_i) \cap words(T_i)\rvert}{\lvert words(S_i) \cup words(D - S_i)\rvert}$ D_COV_C $(S_i) = \frac{\vec{S_i} \times \vec{D - S_i}}{\lvert\vec{S_i}\rvert \times \lvert\vec{D - S_i}\rvert}$	Estas métricas también utilizan las tres medidas de similitud anteriormente mencionadas pero estableciendo la similitud de la sentencia S_i con las restantes sentencias del documento $(D - S_i)$.	[19]

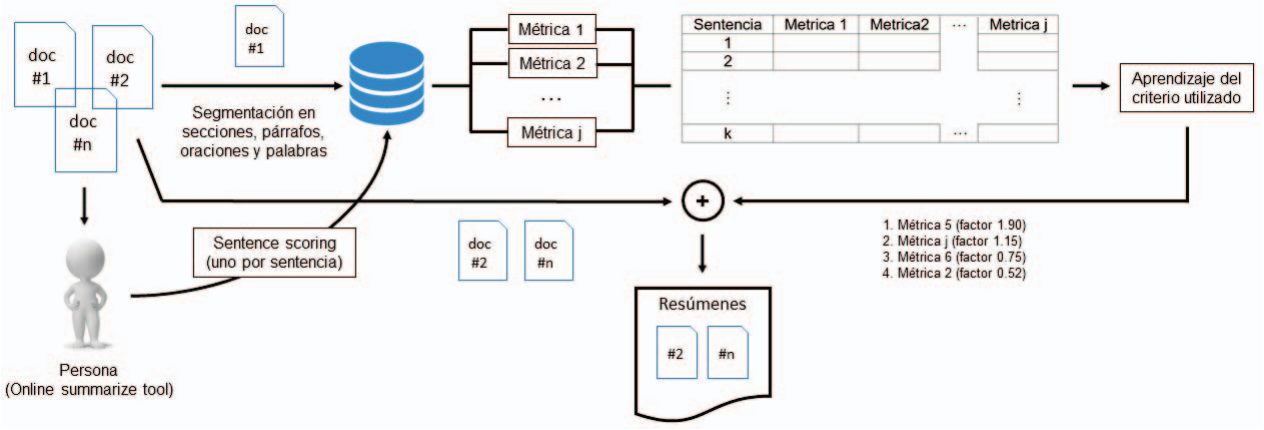

Figura 1. Esquema de la metodología propuesta en este artículo para la generación de resúmenes de documentos

Tabla II
TASA DE ACIERTO PROMEDIO OBTENIDA POR EL MÉTODO PROPUESTO DURANTE LA VALIDACION CRUZADA

Corte	Métricas utilizadas	
	Seleccionadas	Todas (las 17)
5 %	91 %±0.00	92 %±0.01
10 %	83 %±0.01	85 %±0.01
15 %	77 %±0.02	80 %±0.02
20 %	72 %±0.03	75 %±0.02
25 %	68 %±0.03	72 %±0.02
30 %	64 %±0.03	68 %±0.02
35 %	63 %±0.04	67 %±0.02
40 %	60 %±0.04	66 %±0.01

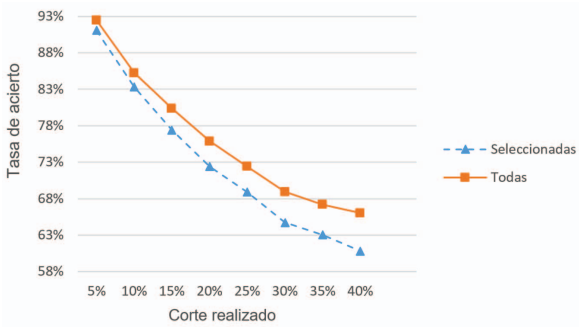

Figura 2. Tasa de acierto promedio obtenida por el método propuesto durante la validación cruzada

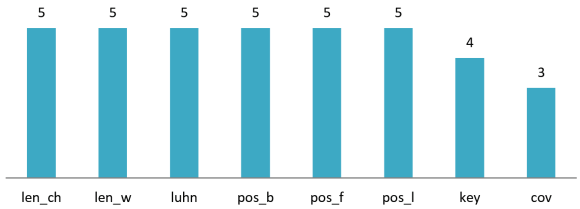

Figura 3. Cantidad de veces que cada métrica fue seleccionada durante la validación cruzada. Las métricas que no aparecen no fueron seleccionadas

y el conocimiento histórico o experiencias anteriores de los individuos situados en su vecindario (su conocimiento social). Existen distintas versiones de esta técnica según si se trabaja con un tamaño de población fija o variable [21] o si el espacio de entrada es continuo o binario [22]. Si bien en esta etapa se utilizó PSO, pudo haberse aplicado cualquier otra técnica de optimización.

Para identificar las características representativas del criterio del usuario se utilizó la versión continua de población fija con un tamaño de población de 30 individuos donde cada uno está formado por tantos coeficientes como métricas se han utilizado para puntuar.

El fitness del individuo es el error cuadrático medio entre el puntaje calculado por el individuo y el estimado por el usuario para cada sentencia tal como se indica en la ecuación (2)

$$fitness_j = (\sum_{i=1}^{T}(puntaje_i - \sum_{t=1}^{k}(c_{jt} * m_{it}))^2/n \quad (2)$$

siendo n la cantidad de sentencias del documento, $puntaje_i$ el puntaje asignado por el usuario a la i-ésima sentencia escalado linealmente entre 0 y 1, c_{jt} el coeficiente que utilizará el individuo j para ponderar a la métrica t de la sentencia i indicado como m_{it}.

IV. RESULTADOS EXPERIMENTALES

Para comprobar el funcionamiento de la metodología propuesta en este artículo se utilizaron cinco capítulos de una Tesina de Licenciatura de la Universidad Nacional de La Plata [23] donde cada capítulo está formado por aproximadamente 80 párrafos, cada párrafo contiene entre 2 y 3 oraciones y cada oración tiene en promedio unas 24 palabras.

Tal como se explicó en la sección III, para aplicar la técnica de optimización es preciso contar con un documento resumido por el usuario. Esto fue resuelto en forma automática utilizando una aplicación web. Luego de analizar varias aplicaciones que realizan resúmenes online a partir de un texto se decidió utilizar [24] ya que de las disponibles en Internet fue la única que cumplió todos los requisitos: (1) cada sentencia corresponde a una oración del texto del documento, (2) permite rankear la totalidad de las sentencias del documento, (3) establece el ranking de sentencias asignándoles un puntaje a cada una y (4) dispone de una interfaz Web que pudo integrarse al desarrollo realizado.

Dado que el objetivo de este trabajo es utilizar un documento resumido de una colección a partir del cuál se aprende el criterio utilizado para seleccionar las sentencias y que a través de una herramienta web se cuenta con algunos documentos resumidos automáticamente, se decidió realizar una validación cruzada con los cinco documentos de la colección utilizando en cada caso uno de ellos para aprender el criterio y los cuatro restantes para testear si dicho criterio permite obtener el resumen esperado. Dado que el resultado de la técnica de optimización depende de la inicialización de la población se realizaron 30 ejecuciones independientes de cada caso y se promediaron los resultados obtenidos.

Antes de comenzar cada uno de los entrenamientos se escalaron linealmente los datos entre 0 y 1. La población inicial fue inicializada en forma aleatoria con distribución uniforme. Los límites de velocidad fueron fijados entre $-0{,}5$ y $0{,}5$ y los límites de variables entre 0 y 3. Esto último permite obtener individuos con coeficientes positivos evitando así que el fitness se ajuste restando los valores de las métricas.

La tabla II resume los resultados obtenidos luego de efectuar una validación cruzada de cinco documentos. Los valores mostrados corresponden al promedio de 30 corridas independientes para cada uno de los umbrales de corte.

Como puede verse tanto en la tabla II como en la figura 2, a medida que se incrementa el umbral de corte se observa que es necesario contar con más métricas para predecir el criterio del usuario con una mayor precisión.

Con respecto a la selección de las métricas, si bien se observa que depende del documento que se tome para realizar el entrenamiento, se identifica un núcleo común formado por las métricas: len_ch, len_w, luhn, pos_b, pos_f, pos_l, key y cov. Estas métricas estuvieron presentes en todos los conjuntos resultantes. La figura 3 ilustra la cantidad de veces que cada métrica fue seleccionada en cada parte de la validación cruzada. La cantidad promedio de métricas seleccionada fue 8, es decir, menos del 50 % de la cantidad total de métricas.

V. Conclusiones y trabajos futuros

Se ha presentado una técnica capaz de aprender el criterio utilizado por un usuario para seleccionar las sentencias más representativas de un documento. Es decir que, a través de un umbral es posible realizar un resumen extractivo con las partes más relevantes según un criterio dado.

Los resultados experimentales muestran que el método propuesto es efectivo. La combinación de métodos de puntuación con PSO permite identificar las métricas más utilizadas por la persona al momento de resumir.

Es importante destacar que la calidad del resumen obtenido depende de la combinación de los métodos de puntuación de sentencias que se utilicen.

La comparación de performance entre las métricas seleccionadas y el conjunto de métricas completo arroja resultados casi iguales en lo que se refiere a identificar hasta el 20 % de las sentencias más significativas según el criterio del usuario. Esto permite afirmar que la selección realizada ha sido correcta.

A futuro se propone analizar si existe relación entre las métricas seleccionadas y el tipo de documento o la temática. También interesa analizar la importancia del lenguaje en el que está escrito el documento así como del estilo utilizado por el autor. Finalmente se repetirán estas mismas pruebas utilizando otras marcaciones para los documentos, incluyendo no sólo las nuevas que se puedan generar en forma automática sino también otras construidas manualmente.

Referencias

[1] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 362–370. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620754.1620807

[2] E. Lloret and M. Palomar, "Text summarisation in progress: a literature review," *Artificial Intelligence Review*, vol. 37, no. 1, pp. 1–41, 2012.

[3] K. S. Jones, "Automatic summarising: Factors and directions," in *Advances in Automatic Text Summarization*. MIT Press, 1998, pp. 1–12.

[4] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011.

[5] J. Xu, S. Zhou, H. Chen, and P. Li, "A sample partition method for learning to rank based on query-level vector extraction," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[6] H. P. Edmundson, "New methods in automatic extracting," *J. ACM*, vol. 16, no. 2, pp. 264–285, Apr. 1969. [Online]. Available: http://doi.acm.org/10.1145/321510.321519

[7] K. Nandhini and S. Balasundaram, "Improving readability through extractive summarization for learners with reading difficulties," *Egyptian Informatics Journal*, vol. 14, no. 3, pp. 195 – 204, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1110866513000376

[8] J. Z. Self, R. Zeitz, C. North, and A. L. Breitler, "Auto-highlighter: Identifying salient sentences in text," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, June 2013, pp. 260–262.

[9] M. Litvak, H. Lipman, A. Ben Gur, M. Last, S. Kisilevich, and D. Keim, "Towards multi-lingual summarization : a comparative analysis of sentence extraction methods on english and hebrew corpora," in *Proceedings of the 4th International Workshop on Cross Lingual Information Access*. Beijing: COLING, 2010, pp. 61–69, paper presented at: 4th International Workshop On Cross Lingual Information Access, Beijing, China, 2010, http://www.aclweb.org/anthology/W10-4010.

[10] R. Ferreira, F. Freitas, L. d. S. Cabral, R. D. Lins, R. Lima, G. França, S. J. Simske, and L. Favaro, "A context based text summarization system," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, April 2014, pp. 66–70.

[11] C. Nobata, S. Sekine, and H. Isahara, "Evaluation of features for sentence extraction on different types of corpora," in *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, ser. MultiSumQA '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 29–36. [Online]. Available: http://dx.doi.org/10.3115/1119312.1119316

[12] P. B. Baxendale, "Machine-made index for technical literature—an experiment," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354–361, Oct 1958.

[13] C. Nobata, S. Sekine, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence extraction system assembling multiple evidence," in *In Proceedings of the Second NTCIR Workshop Meeting*, 2001, pp. 5–213.

[14] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, Apr. 1958. [Online]. Available: http://dx.doi.org/10.1147/rd.22.0159

[15] F. J. Kallel, M. Jaoua, L. B. Hadrich, and A. B. Hamadou, "Summarization at laris laboratory," in *In Proceedings of the Document Understanding Conference*, 2004.

[16] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1606–1618, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2007.01.023

[17] J. Larocca Neto, A. D. Santos, C. A. Kaestner, and A. A. Freitas, *Advances in Artificial Intelligence: International Joint Conference 7th Ibero-American Conference on AI 15th Brazilian Symposium on AI IBERAMIA-SBIA 2000 Atibaia, SP, Brazil, November 19–22, 2000 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, ch. Generating Text Summaries through the Relative Importance of Topics, pp. 300–309.

[18] J. Steinberger and K. Ježek, *Advances in Information Systems: Third International Conference, ADVIS 2004, Izmir, Turkey, October 20-22, 2004. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ch. Text Summarization and Singular Value Decomposition, pp. 245–254.

[19] M. Litvak, M. Last, and M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 927–936. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858776

[20] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, 1995, pp. 1942–1948.

[21] L. Lanzarini, V. Leza, and A. De Giusti, "Particle swarm optimization with variable population size," in *Artificial Intelligence and Soft Computing - ICAISC 2008*, ser. Lecture Notes in Computer Science, L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds. Springer Berlin Heidelberg, 2008, vol. 5097, pp. 438–449.

[22] L. Lanzarini, J. López, J. A. Maulini, and A. De Giusti, "A new binary pso with velocity control," in *Advances in Swarm Intelligence*, ser. Lecture Notes in Computer Science, Y. Tan, Y. Shi, Y. Chai, and G. Wang, Eds. Springer Berlin Heidelberg, 2011, vol. 6728, pp. 111–119.

[23] A. Villa Monte. (2013, Mar.) Obtención de reglas de clasificación utilizando estrategias adaptativas. http://sedici.unlp.edu.ar/handle/10915/47056.

[24] Online summarize tool. https://www.tools4noobs.com/summarize/.