Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 103-110 recibido 12-04-2019 revisado 15-05-2019 aceptado 17-05-2019

Generación de resúmenes extractivos de múltiples documentos usando grafos semánticos

Multi-document extractive summarization using semantic graph

Oleyda del Camino Valle[1], Alfredo Simón-Cuevas[2], Eduardo Valladares-Valdés[2], José A. Olivas[3], Francisco P. Romero[3]

[1] Empresa Nacional de Software (DESOFT), La Habana, Cuba
[2] Universidad Tecnológica de La Habana "José Antonio Echeverría",
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
[3] Universidad de Castilla La Mancha, Paseo de la Universidad, 4, Ciudad Real, España
oleyda.camino@cfr.desoft.cu, {asimon, evalladares}@ceis.cujae.edu.cu
{JoseAngel.Olivas, FranciscoP.Romero}@uclm.es

Resumen: La generación automática de resúmenes consiste en sintetizar en un texto corto la información más relevante contenida en documentos, y permite reducir los problemas generados por la sobrecarga de información. En este trabajo se presenta un método no supervisado de generación de resúmenes extractivos a partir de múltiples documentos. En esta propuesta, la conceptualización y estructura semántica subyacente del contenido textual se representa en un grafo semántico usando WordNet y se aplica un algoritmo de agrupamiento de conceptos para identificar los tópicos tratados en los documentos, con los cuales se evalúa la relevancia de las oraciones para construir el resumen. El método fue evaluado con corpus de textos de MultiLing 2015, y se usaron métricas de ROUGE para medir la calidad de los resúmenes generados. Los resultados obtenidos se compararon con los de otros sistemas participantes en MultiLing 2015, evidenciándose mejoras en la mayoría de los casos.
Palabras clave: Generación de resúmenes multi-documentos, grafos semánticos, desambiguación, agrupamiento de conceptos

Abstract: The automatic texts summarization consists in synthesizing in a short text the most relevant information contained in text documents, and allows to reduce the generated problems by the information overload. In this paper, an unsupervised method for extractive multi-document summarization is presented. In this proposal, the conceptualization and underlying semantics structure of the textual content is represented in a semantic graph using WordNet, and a concept clustering algorithm is applied to identifying the topics of the documents set, with which the relevance of the sentences is evaluated to build the summary. The method was evaluated with texts corpus from MultiLing 2015, and ROUGE metrics were used to measure the quality of the generated summaries. The obtained results were compared with those other participant systems in MultiLing 2015, evidencing improves in most of the cases.
Keywords: multi-document summarization, semantic graph, word sense disambiguation, concept clustering

1 Introducción

En la actualidad, el crecimiento exponencial de la información textual disponible se presenta como una amenaza para su uso efectivo en la toma de decisiones, ya que, si bien su acceso parece estar garantizado, no ocurre lo mismo con el tiempo necesario para su análisis y procesamiento. Por tanto, la obtención de información relevante dentro de ese gran volumen de textos constituye un importante y desafiante objetivo alcanzar en ese escenario. La generación automática de resúmenes de textos es una de las líneas que actualmente se trabaja para enfrentar este desafío y reducir el impacto de la sobrecarga de información.

La generación automática de resúmenes tiene el propósito de condensar en un texto

ISSN 1135-5948. DOI 10.26342/2019-63-11

corto la información más relevante y esencial contenida en uno o varios documentos de textos (Das y Martins, 2007), considerando los criterios de concisión por parte del usuario (Padmapriya y Rajasekaran, 2012; Kumar y Salim, 2012), y preservando el contenido principal de la información, así como su significado general (Das y Martins, 2007; Padmapriya y Rajasekaran, 2012). Los resúmenes se pueden obtener mediante métodos extractivos (seleccionan las oraciones más importantes de los textos) o abstractivos (usan palabras o frases diferentes a las incluidas en los documentos originales) (Bhatia y Jaiswal, 2015), y a partir de enfoques supervisados, no supervisados o semi-supervisados (Gambhir y Gupta, 2017). Esta problemática ha sido ampliamente abordada, sin embargo, sigue siendo una tarea desafiante, sobre todo cuando se quiere obtener el resumen de múltiples documentos (Gambhir y Gupta, 2017).

En el trabajo se presenta un método no supervisado que combina enfoques basados en grafos y conceptos (Moratanch y Chitrakala, 2017) para generar resúmenes extractivos de múltiples documentos (o multi-documentos). En esta nueva propuesta, la conceptualización y estructura semántica subyacente del contenido de los documentos se representa mediante grafos semántico generados automáticamente, a partir de la identificación de conceptos y relaciones semánticas entre ellos desde WordNet (Miller y Fellbaum, 1998). Estos grafos se fusionan para obtener una única representación integrada del contenido de los documentos, soportado en la aplicación de un algoritmo de desambiguación para resolver las ambigüedades presentes. A partir del grafo que representa el conjunto de documentos, se aplica un algoritmo de agrupamiento de conceptos para identificar los tópicos más relevantes tratados en ese contenido, con los cuales se mide la relevancia de las oraciones a través de evaluar su similitud con los clústeres que representan los tópicos con un enfoque semántico. El uso de este tipo de grafos, inspirado en la propuesta de generación de resúmenes mono-documentos de Plaza y Díaz (2011), permite lograr mayor granularidad en el procesamiento semántico del contenido de los documentos, respecto a otras soluciones reportadas, así como identificar los temas principales en el contenido y generar resúmenes que se correspondan más a esos temas. También la inclusión de un proceso de desambiguación, como soporte a la integración del contenido de los documentos, constituye otra contribución a este ámbito, dado que ha sido poco tratado en otras propuestas reportadas.

El método propuesto fue evaluado con corpus de textos en español e inglés ofrecidos en MultiLing 2015, y la calidad de los resúmenes obtenidos fue medida usando métricas de ROUGE-N (Lin, 2004). Los resultados del método fueron comparados con los obtenidos por otros sistemas evaluados con esos corpus en el contexto de MultiLing2015, resultando mejores en la mayoría de los casos.

El trabajo se ha organizado de la siguiente forma: en la Sección 2 se sintetiza el análisis de los trabajos relacionados; en la Sección 3 se describe el método propuesto; en la Sección 4 se muestran y analizan los resultados experimentales obtenidos; y en la Sección 5 se exponen las conclusiones arribadas y líneas de trabajo futuro.

2 Trabajos relacionados

La generación de resúmenes multi-documentos consiste en crear de forma automática un texto corto con la información más relevante y esencial de un conjunto de documentos, en principio, relacionados a un tema específico. En estos procesos surgen problemáticas adicionales que repercuten en mayores complejidades y desafíos, respecto a la generación de resúmenes mono-documento, por ejemplo: mezcla de información inconexa y no relacionada (los documentos pueden no compartir los mismos temas), redundancia (los documentos pueden contener información en común), falta de coherencia (las unidades de información provienen de diferentes fuentes), entre otras (Gambhir y Gupta, 2017).

Los métodos extractivos de generación de resúmenes se han diseñado usando técnicas supervisadas y no supervisadas, y estas últimas han sido aplicadas con diferentes enfoques (Moratanch y Chitrakala, 2017). En la generación de resúmenes multi-documentos específicamente se han aplicado: técnicas basadas en grafos (los documentos se representan en modelos de grafos) (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013; Ferreira et al. 2014; Zore y Deshpande, 2014; Mirchev y Last, 2014; Yan y Wan, 2014; Al-Saleh y Menai, 2018); basadas en conceptos (se identifican conceptos de fuentes externas con los que se realiza el

análisis de relevancia) (Baralis et al., 2013; Sankarasubramaniam, Ramanathan, y Ghosh, 2014); lógica difusa (algunas características del contenido textual se procesan usando técnicas difusa, tales como: sistemas basados en conjuntos difusos y reglas de inferencia) (Bhoir y Gulati, 2015); LSA (*Latent Semantic Analysis*) (para procesar la estructura semántica subyacente de oraciones y palabras) (Steinberger, 2013); evaluación de características del contenido textual (Bhoir y Gulati, 2015; Naserasa, Khosravi, y Sadegh, 2018); métodos de agrupamiento (Zhong et al., 2017; Puspaningrum et al. 2018); entre otras. Según esta caracterización, el uso de grafos constituye uno de los enfoques más aplicados, siendo común en estas soluciones la representación de las oraciones como vértices en los grafos (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013; Ferreira et al. 2014; Zore y Deshpande, 2014; Al-Saleh y Menai, 2018), y el establecimiento de relaciones entre ellas a partir de medidas de similitud, en su mayoría basadas en aspectos sintácticos del contenido (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013), y no considerando aspectos semánticos (Ferreira et al. 2014). Sin embargo, este tipo de esquema de representación limita el poder realizar un procesamiento computacional del contenido con un mayor nivel de granularidad, por ejemplo, a nivel de conceptos. Esto resulta ser poco conveniente en soluciones que incluyen el modelado de tópicos como parte de la evaluación de la relevancia de las oraciones, donde los resúmenes se construyen con las oraciones que mayor representen los tópicos o temas principales tratados en los documentos, así como para darle solución los problemas de ambigüedad presentes en el contenido textual.

La ambigüedad de las palabras es otro gran problema que aún no se ha resuelto en el área de la construcción automática de resúmenes (Soriyan y Omodunbi, 2014), aunque este problema puede afectar en menor medida en la construcción de resúmenes mono-documentos. En el contexto de resúmenes multi-documentos, tal situación puede ocurrir con mayor probabilidad y, por tanto, tener mayor impacto en la calidad de los resultados (Ferreira et al. 2013). Inicialmente se puede identificar que una oración en un documento A puede tener un alto grado de similitud a otra del documento B (sin tener en cuenta las posibles ambigüedades en el contenido), y esto incluso puede sugerir que esa oración sea una buena candidata para el resumen, pero solo una de ellas debe ser incluida. Sin embargo, el tratamiento de la ambigüedad, puede resultar que esa semejanza aparente no sea tal, ya que se pueden estar usando términos con diferentes significados, logrando con ello una mayor precisión en el procesamiento y reduciendo la ocurrencia de redundancias. A pesar de lo argumentando, en la mayoría de las propuestas estudiadas no se incluye la desambiguación como parte del proceso de generación de resúmenes, esto solo se reporta en (Baralis et al., 2013). En (Baralis et al., 2013) los conceptos presentes en el texto son identificados y desambiguados usando la ontología Yago. El proceso de desambiguación se realiza analizando el grado de pertinencia de cada entidad de la ontología al contexto en el que se encuentran los conceptos ambiguos en el texto. El impacto de ambigüedad en la calidad de los resúmenes generados también se puede reducir usando medidas de similitud semántica, por ejemplo, entre oraciones en las soluciones basadas en grafos, pero esto tampoco ha sido común en este tipo de propuestas.

3 Método propuesto

El método fue concebido teniendo en cuenta los procesos que suelen considerarse en el diseño de este tipo de soluciones (Allahyari et al., 2017): (1) representación de los documentos en grafos semánticos; (2) fusión de los grafos semánticos; (3) agrupamiento de conceptos; (4) evaluación de relevancia de las oraciones; y (5) construcción del resumen.

3.1 Representación de los documentos en grafos semánticos

En esta fase se construye automáticamente un grafo semántico de cada uno de los documentos, en los cuales se representan conceptos presentes en el texto y las relaciones semánticas entre ellos, usando WordNet (Miller y Fellbaum, 1998) como referencia para la captura de esa información. Previo a la generación de los grafos, se aplican un conjunto de tareas básicas de procesamiento de lenguaje natural sobre los documentos, tales como: extracción del texto plano, segmentación en oraciones, análisis morfo-sintáctico y la eliminación de palabras desprovistas de significado (o *stop words*). En este pre-procesamiento se utiliza el analizador sintáctico

Freeling, dado que brinda soporte para procesar textos en español e inglés.

La construcción de los grafos parte de extraer de WordNet los sentidos (*synset*) de los conceptos incluidos en cada una de las oraciones, así como las relaciones semánticas de hiperonimia, hiponimia, meronimia y holonimia existentes entre ellos. A partir de cada uno de los *synset* identificados también se extraen aquellos *synset* incluidos en una vecindad de radio 2 con los cuales están relacionados (según los tipos de relación mencionados). Estos elementos capturados son usados para obtener el grafo que representa a cada oración, donde los *synset* constituyen los nodos (los conceptos ambiguos estarían representados en más de un nodo por sus *synset*) y las relaciones se establecen según los vínculos identificados en WordNet. Luego, los grafos generados son integrados para obtener un único grafo del documento, cuyo proceso se realiza unificando los *synset* comunes en los diferentes grafos de las oraciones.

3.2 Fusión de los grafos semánticos

El propósito de este proceso es obtener una representación del contenido del conjunto de documentos en un único grafo, a partir del cual se identifican los tópicos más relevantes con el algoritmo de agrupamiento que se aplica en la siguiente fase. Este proceso se lleva a cabo mediante la integración de los grafos semánticos que representan a cada uno de los documentos. La fusión de los grafos está basada en la identificación e integración de los conceptos representados que tienen el mismo significado. Por lo tanto, como paso previo es necesario resolver las ambigüedades presentes en los conceptos representados en esos grafos. En este sentido, se aplica el algoritmo de desambiguación reportado en (Hojas et al. 2018) sobre cada uno de los grafos de los documentos. Este algoritmo combina heurísticas de dominio, contexto y glosa para determinar el sentido más apropiado de un concepto representado en un grafo usando WordNet. A partir del análisis de cada heurística se le otorga un valor de peso a cada sentido posible, cuyos valores son agregados en una suma ponderada para obtener un valor de peso global, siendo el de mayor peso el seleccionado como sentido del concepto. Esta combinación de heurísticas reduce las limitaciones derivadas de aplicar una sola de ellas en este proceso (Navigli, 2009), como ocurre en el caso del algoritmo basado en glosa usado en (Plaza y Díaz, 2011).

El proceso de desambiguación incluido en esta propuesta constituye un factor clave para mejorar los resultados en la obtención de los resúmenes multi-documentos, aun cuando la mayoría de las soluciones similares estudiadas no lo consideran. Esto no solo permite evitar posibles redundancias (Ferreira et al. 2013), sino también contribuye a realizar un uso más eficaz de Wordnet, y con ello reducir los efectos de otros problemas que pueden padecer este tipo de soluciones como: mezcla de información inconexa o no relacionada, y la falta de coherencia en el resumen resultante.

Al concluir la desambiguación, son refinados los grafos que representan los documentos, eliminándose aquellos *synset* de los términos ambiguos que recibieron una menor votación como resultado del algoritmo aplicado. Finalmente, todos los grafos de los documentos son fusionados mediante la integración de los vértices que representen los mismos *synset*, obteniéndose un único grafo semántico que representa el contenido del conjunto documentos.

3.3 Agrupamiento de conceptos

En esta fase, los conceptos representados en el grafo resultante del proceso anterior son agrupados en clústeres, los cuales representan los tópicos principales abordados en el conjunto de documentos. Los conceptos centroides en estos clústeres son los que aportan la información más relevante sobre cada tópico. Este proceso se lleva a cabo usando un algoritmo de agrupamiento basado en la conectividad, similar al aplicado en (Plaza y Díaz, 2011). Los conceptos son agrupados de acuerdo al grado de conectividad entre ellos, considerando que, en el tipo de grafo que se genera en esta propuesta, unos pocos vértices estarán altamente conectados, mientras que el grado de conectividad del resto de los vértices será relativamente más bajo.

El algoritmo de agrupamiento inicia identificando el grado de conectividad de cada vértice del grafo, otorgando un valor de relevancia para cada uno mediante la suma de sus vértices adyacentes; con lo que se genera un ranking de vértices. A partir de este ranking son seleccionados lo *n* vértices de mayor relevancia (mayor grado de conectividad), siendo el valor

de *n* determinado por un usuario. Los vértices de mayor relevancia son denominados vértices *HUB* y son agrupados en conjuntos denominados *HVS* (*HUB Vertex Sets*), considerando que *HVS* es un conjunto de vértices fuertemente conectados entre sí y constituyen los centroides de los clústeres. El algoritmo de agrupamiento construye los *HVS* buscando, iterativamente y para cada vértice *HUB*, aquel vértice *HUB* más conectado a este, y los agrupa en un único *HVS*. Posteriormente, se comprueba para cada par de *HVS* si sus conectividades internas son menores que la conectividad entre ellos, agrupándose ambos *HVS* si esta condición se cumple. Finalmente, en un proceso iterativo se agrupan los vértices no clasificados como *HUB* al *HVS* correspondiente (con el que existe una mayor conectividad) para obtener los clústeres finales.

3.4 Evaluación de relevancia de las oraciones

En esta fase es evaluada la relevancia de cada oración presente en los documentos, obteniéndose un ranking a partir del cual se seleccionan las oraciones que conformarán el resumen. Este proceso se lleva a cabo mediante una evaluación de la similitud existente entre el contenido de cada una de las oraciones y los clústeres generados en la fase anterior, con un enfoque basado en el análisis semántico.

La similitud entre la oración y el clúster es evaluada combinando el uso de la medida de similitud semántica de Lin (1998) (incluida en el paquete de *WordNet::Similarity* (Pedersen, Patwardhan y Michelizzi, 2004)) para evaluar la similitud entre los conceptos de la oración y los incluidos en el clúster, con un mecanismo de ponderación similar al usado en (Plaza y Díaz, 2011), y calculada según la ecuación (1). En este caso, para cada uno de los conceptos (c_i) de la oración, se calcula su similitud semántica con cada uno de los conceptos (v_j) del clúster (C_j) ($sim\_semántica(c_i,v_j)$), ponderado con un peso ($w_{i,j}$) diferente, según el tipo de vértice del clúster con el que se realiza el análisis. Considerando que los conceptos pertenecientes al *HVS* tienen mayor importancia que el resto, se considera el peso 1.0 ($w_{i,j} = 1$) si v_j pertenece al *HVS* del clúster y la mitad del peso ($w_{i,j} = 0.5$) en caso contrario. Los valores de similitud obtenidos son promediados para obtener el valor de relevancia global de las oraciones.

$$similitud(O_i, C_j) = \sum_{c_i / c_i \in O_i} \sum_{v_j / v_j \in C_j} w_{i,j} * sim\_semántica(c_i, v_j) \quad (1)$$

3.5 Construcción del resumen

Luego de haber calculado la relevancia de las oraciones, en esta fase se procede a la construcción del resumen seleccionando las N oraciones con mayor relevancia, siendo N un parámetro que representa la tasa de compresión deseada. En este proceso de selección son considerados otros dos aspectos: el orden de los documentos en el que aparecen las oraciones más relevantes, y luego el orden en que aparecen dentro del documento. Estos criterios contribuirían a evitar inconsistencias en el resumen resultante.

4 Resultados experimentales

El método propuesto fue evaluado con los corpus de textos en español e inglés ofrecidos en MultiLing 2015 para evaluar soluciones en la tarea MMS (*Multilingual Multi-document Summarization*), siendo la última edición en la que ha sido incluida la generación de resúmenes multi-documentos como tarea de evaluación en este escenario. Los corpus seleccionados están formados por artículos de noticias provenientes de *WikiNews* asociados a 15 tópicos, y los mismos se han agrupado en dos colecciones: ENCol y SPCol, respectivamente. Ambas colecciones se caracterizan en la Tabla 1.

	Colecciones	
Características	**ENCol**	**SPCol**
Idioma	Inglés	Español
Cantidad de Corpus	15	15
Documentos en cada Corpus	10	10
Cantidad de Oraciones Totales	4469	5000
Prom. de Oraciones en Corpus	297,9	333,3
Prom. de Oraciones en Doc.	28,92	33,34

Tabla 1: Caracterización de las colecciones

Los resúmenes obtenidos son evaluados usando las métricas de precisión (P), *recall* (R) y medida-F (F), en el contexto de ROUGE. Específicamente, ROUGE-1 y ROUGE-2, teniendo en cuenta que ambas medidas funcionan relativamente bien en este tipo de tarea, según se reporta en (Lin, 2004). La Tabla 2 muestra los resultados obtenidos y su comparación con las soluciones participantes en la tarea MMS de MultiLing 2015 y evaluados

con los corpus seleccionados. También se incluye los resultados del sistema *Lead*, el cual fue propuesto como *baseline* para esta tarea.

Según se aprecia en la Tabla 2, los resultados obtenidos por el método propuesto superan los valores del *baseline* en cada una de las métricas y colecciones de prueba. Los valores obtenidos para ROUGE-1 también mejoran los del resto de los sistemas en ambas colecciones, lográndose un incremento superior al 20 %, con respecto a los valores promedio obtenidos por esos sistemas en cada una de las métricas. Constituye un resultado muy positivo también los valores superiores al 50 % obtenidos de *recall* y medida-F en las pruebas con SPCol. En el caso de las mediciones con ROUGE-2, los valores de *recall* y medida-F obtenidos en ambas colecciones fueron superiores a los reportados por el resto de los sistemas. En este caso, la precisión obtenida por el método no fue la mayor, pero el valor resultante es muy cercano al obtenido por los sistemas que mejores resultados reportaron, específicamente, BGU-MUSE (Litvak et al., 2016) y UWB (Steinberger, 2013). BGU-MUSE (Litvak et al., 2016) es un método supervisado, basado en la aplicación de algoritmos genéticos y entrenado con colecciones ofrecidas en MultiLing 2015. Esta cualidad constituye una desventaja frente al acercamiento no supervisado propuesto, si se desea generalizar su aplicación en otros dominios donde no se disponga de colecciones de entrenamiento. Por otra parte, UWB (Steinberger, 2013) se basa en la técnica *LSA* para modelar los tópicos y el tamaño del vector de las oraciones es usado como medida para evaluar su relevancia dentro de los tópicos. Aunque los resultados que obtiene son buenos, esta estrategia puede requerir que más de una oración exprese toda la información asociada a los tópicos, lo que puede constituir una limitación en algunas colecciones (Nenkova y McKeown, 2012). El método propuesto no se ve afectado por este problema dado que los temas son modelados de una forma más granular, representando e integrando estructuras semánticas de conceptos y aplicando luego un algoritmo de agrupamiento de conceptos.

Sistemas	**ENCol**						**SPCol**					
	ROUGE-1			**ROUGE-2**			**ROUGE-1**			**ROUGE-2**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SCE-Poly	.207	.196	.200	.127	.117	.121						
BUPT-CIST	.125	.110	.116	.019	.015	.017	.130	.122	.124	.028	.027	.028
BGU-MUSE	.294	.277	.284	**.188**	.173	.180						
NCSR/SCIFY	.154	.135	.142	.053	.043	.046	.226	.202	.211	.068	.059	.062
UJF-Grenoble	.142	.121	.129	.043	.034	.038						
UWB	**.344**	**.327**	**.333**	.186	**.178**	**.181**	**.412**	**.390**	**.399**	**.261**	**.247**	**.253**
ExB	.229	.227	.227	.084	.086	.085	.253	.242	.245	.098	.097	.097
ESIAllSummr	.159	.150	.153	.039	.034	.036	.208	.202	.204	.054	.055	.054
IDAOCCAMS	.230	.230	.229	.066	.068	.067	.252	.244	.246	.085	.081	.082
GiauUngVan	.147	.123	.134	.037	.027	.031						
Lead (baseline)	*.391*	*.409*	*.401*	*.111*	*.126*	*.116*	*.451*	*.458*	*.453*	*.158*	*.170*	*.165*
Solución propuesta	**.427**	**.460**	**.442**	.186	**.185**	**.185**	**.485**	**.558**	**.518**	.259	**.285**	**.271**

Tabla 2: Resultados experimentales

5 Conclusiones y trabajo futuro

En el trabajo se ha presentado un nuevo método no supervisado de generación de resúmenes extractivos multi-documentos. En este método se usan grafos semánticos capturados de WordNet para representar el contenido del conjunto de documentos, incluyendo la aplicación de un algoritmo de desambiguación que combina varias heurísticas. La aplicación de este algoritmo, no solo posibilita reducir las ambigüedades presentes en el contenido, sino también lograr mayor precisión en la integración del contenido de los documentos representados en los grafos y reducir la ocurrencia de redundancias en el resumen resultante. El uso de los grafos semánticos

basado en conceptos y relaciones entre ellos permitió lograr mayor granularidad en el procesamiento del contenido textual, así como a reducir la mezcla de información inconexa o no relacionada, y la falta de coherencia en los resúmenes generados. Por otra parte, también facilitó la identificación de los temas principales tratados en los documentos, a través del algoritmo de agrupamiento de conceptos aplicado, y con ello poder generar resúmenes sobre la base de la selección de oraciones relevantes para esos temas. Los experimentos realizados sobre colecciones de textos en español e inglés, arrojaron como resultado mejoras en la calidad de los resúmenes generados por esta propuesta, con respecto a otros sistemas. En todas las pruebas se superaron los valores de ROUGE del *baseline*, y en el caso de ROUGE-1 se superaron los resultados de todos los sistemas (en ambos idiomas). En el caso de ROUGE-2, los resultados obtenidos en ambas colecciones fueron, en general, superiores al resto de los sistemas, destacándose mejoras en la cobertura y en la medida-F, y alcanzándose una precisión muy cercana a la obtenida por los sistemas que mejores resultados reportaron.

En trabajos futuros se pretende evaluar el uso de otros recursos externos, como BabelNet, para generar los grafos, con el objetivo de lograr una mayor cobertura en la representación del contenido de los documentos. Además, se evaluarán otras alternativas para identificar los temas principales del contenido a partir de los grafos generados, por ejemplo, aplicando otros algoritmos dirigidos a determinar nodos relevantes en grafos. En este sentido, se considerarían también variantes de ponderación de las relaciones representadas en el grafo en función de su significado semántico, de tal forma que se logre un mayor aprovechamiento de esa información en este proceso.

Agradecimientos

Este trabajo ha sido parcialmente soportado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio Español de Economía y Competitividad, bajo la subvención del proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET) Ref. TIN2016-76843-C4-2-R (AEI/FEDER, UE)

Bibliografía

Allahyari, M., S., Pouriyeh, S., Safaei, E. D., Trippe, J. B., Gutierrez, y K., Kochut. 2017. Text Summarization Techniques: A Brief Survey, *Int. J. of Advanced Computer Science and Applications*, 8(10): 397-405.

Al-Saleh, A. y M. El B. Menai. 2018. Solving Multi-Document Summarization as an Orienteering Problem. *Algorithms*, 11(7):1-27.

Baralis, E., L. Cagliero, S. Jabeen, A. Fiori, y S. Shah. 2013. Multi-document summarization based on the Yago ontology, *Expert Systems with Applications*, 40:6976–6984.

Bhatia, N., y A., Jaiswal. 2015. Trends in Extractive and Abstractive Techniques in Text Summarization, *International Journal of Computer Applications*, 117(6):21-24.

Bhoir, A. S., y A. Gulati. 2015. A Multi-document Hindi Text Summarization Technique using Fuzzy Logic. *Int. J. of Advance Research in Science and Engineering*, 4(1):468-476.

Das, D. y A. F. Martins. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4: 192-195.

Erkan, G., y D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research*, 22:457-479.

Ferreira, R., L. S. Cabral, R. D. Lins, G. Pereira e Silva, F. Freitas, G. D.C. Cavalcanti, R. Lima, S. J. Simske, y L. Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, 40:5755-5764.

Ferreira, R., L. S. Cabral, F. Freitas, R. D. Lins, G. F. Silva, S. J. Simske, y L. Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41:5780–5787.

Gambhir, M., y V. Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1-66.

Hariharan, S., T. Ramkumar, y R. Srinivasan. 2013. Enhanced graph based approach for

multi document summarization. *Int. Arab J. Inf. Technol.*, 10:334-341.

Hojas, W., A. Simón, M. de la Iglesia, F. P. Romero, J. A. Olivas. 2018. A Concept-Based Text Analysis Approach Using Knowledge Graph. *Communications in Computer and Information Science*, 854:696-708.

Kumar, Y. J., y N. Salim. 2012. Automatic Multi Document Summarization Approaches. *Journal of Computer Science*, 8(1):133-140.

Lin, C.-Y. 2004. ROUGE: a package for automatic evaluation of summaries. En *Proceedings of the ACL-04 workshop*, páginas 74-81.

Lin, D. 1998. An information-theoretic definition of similarity. En *Proceedings of the International Conference on Machine Learning*.

Litvak, M., N. Vanetik, M. Last, y E. Churkin. 2016. MUSEEC: A Multi-lingual Text Summarization Tool. En *Proceedings of the 54th Annual Meeting of the ACL - System Demonstrations*, páginas 73 - 78.

Miller, G. y C. Fellbaum (Eds.). 1998. WordNet: An Electronic Lexical Database, The MIT Press: Cambridge, MA. 1998.

Mirchev, U., y M. Last. 2014. Multi-document summarization by extended graph text representation and importance refinement. En A. Fiori (Ed): Summarization Techniques: Revolutionizing Knowledge Understanding, IGI Global, páginas 28-53.

Moratanch, N. y S. Chitrakala. 2017. A Survey on Extractive Text Summarization. En *Proceedings of the IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)*.

Navigli, R. 2009. Word sense disambiguation: A survey, *ACM Computing Surveys*, 41(2):1-69.

Naserasa, A., H. Khosravi, y F. Sadegh. 2018. Extractive multi-document summarization based on textual entailment and sentence compression via knapsack problem. *Natural Language Engineering*, 25 (1):121–146.

Nenkova, A. y K. McKeown. 2012. A Survey of Text Summarization Techniques. En C.C. Aggarwal and C.X. Zhai (eds.), Mining Text Data, Springer, páginas 44-76.

Padmapriya, K. D. G., y V. G. Rajasekaran. 2012. A View On Natural Language Processing and Text Summarization. *Int. Journal of Communications and Engineering*.

Pedersen, T., S. Patwardhan y J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. En *Proceedings of the AAAI-04,* páginas 1024-1025.

Plaza, L., y A. Diaz. 2011. Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. *Procesamiento del Lenguaje Natural*, 47:97-105.

Puspaningrum, A., A. Nurilham, E. F. Bisono, K. Umam, y A. Z. Arifin. 2018. Inter and Intra Cluster on Self-Adaptive Differential Evolution for Multi-Document Summarization. *Journal of a Science and Information*, 11(2):86-94.

Soriyan, A., y T. Omodunbi. 2014. Trends in Multi-document Summarization System Methods. *International Journal of Computer Applications*, 97(16):46-52.

Sankarasubramaniam, Y., K. Ramanathan, y S. Ghosh. 2014. Text summarization using Wikipedia. *Information Processing & Management*, 50(3):443-461.

Steinberger, J. 2013. The UWB Summariser at Multiling 2013. En *Proc. of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, páginas 50-54.

Yan, S., y X. Wan. 2014. SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization. *IEEE/ACM Transactions on Audio, Speech, and Learning Processing,* 22(12):2048-2058.

Zhong, Y., Z. Tang, X. Ding, L. Zhu, Y. Le, K. Li, y K. Li. 2017. An Improved LDA Multi-Document Summarization Model Based on TensorFlow. En *Proceedings of the 2017 International Conference on Tools with Artificial Intelligence*, páginas 255-259.

Zore, A. S, y A. Deshpande. 2014. Extractive Multi Document Summarizer Algorithm. *IJCSIT*, 5(4):5245-5248.