Procesamiento del Lenguaje Natural
ISSN: 1135-5948
secretaria.sepln@ujaen.es
Sociedad Española para el Procesamiento del Lenguaje Natural
España

Pérez-Guadarramas, Yamel; Rodríguez-Blanco, Aramis; Simón-Cuevas, Alfredo; Hojas-Mazo, Wenny; Olivas, José Ángel

Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos

Procesamiento del Lenguaje Natural, núm. 59, 2017, pp. 39-46
Sociedad Española para el Procesamiento del Lenguaje Natural
Jaén, España

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 39-46 recibido 31-03-2017 revisado 11-05-2017 aceptado 25-05-2017

Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos

Combining lexical-syntactic patterns and topic analysis for automatic keyphrase extraction from texts

Yamel Pérez-Guadarramas[1], Aramis Rodríguez-Blanco[1], Alfredo Simón-Cuevas[1], Wenny Hojas-Mazo[1], José Ángel Olivas[2]
[1]Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
{yperezg, aridriguezb, asimon, whojas}@ceis.cujae.edu.cu
[2]Universidad de Castilla La Mancha
Paseo de la Universidad, 4, Ciudad Real, España
JoseAngel.Olivas@uclm.es

Resumen: La extracción automática de frases relevantes constituye una tarea de gran importancia para muchas soluciones computacionales en el área del procesamiento de lenguaje natural y la minería de texto. En este trabajo se propone un nuevo método no supervisado para la extracción de frases relevantes en textos, en el cual se combina el uso de patrones léxico-sintácticos con una estrategia de análisis de tópicos basada en grafo. El método fue evaluado con los corpus SemEval-2010 e INSPEC y comparado con otras propuestas del estado del arte, obteniéndose resultados muy prometedores.
Palabras claves: Extracción automática de frases relevantes, minería de texto, procesamiento de lenguaje natural

Abstract: The automatic keyphrases extraction is a useful task for many computational solutions in the natural language processing and text mining areas. In this paper, a new unsupervised method for keyphrase extraction from texts is proposed, in which the use of lexical-syntactic patterns is combined with a graph-based topic analysis strategy. The method was evaluated with the SemEval-2010 and INSPEC corpus, and compared with other state-of-the-art proposals, obtaining promising results.
Keywords: Automatic keyphrase extraction, text mining, natural language processing

1 Introducción

Actualmente, es notable la cantidad de información textual disponible en formato digital, sobre todo en el ámbito de Internet, ya sea en forma de noticias, opiniones, artículos, u otros. En este escenario, surge la minería de texto (MT) como el proceso de descubrimiento de conocimientos en colecciones de textos a partir de la identificación y exploración de patrones interesantes, con el objetivo incrementar el aprovechamiento de esa información. Este proceso puede estar orientado a diferentes tipos de soluciones: construcción de resúmenes, clasificación y agrupamiento de documentos, recuperación de información, minería de opinión, entre otras.

A través de las palabras o frases relevantes se puede alcanzar un alto nivel de descripción de un documento, por a su relación con el o los temas principales que se abordan en el mismo, por lo que su extracción de forma automática constituye una tarea de gran utilidad para la MT (Hasan y Ng, 2014; Merrouni, Frikh, y Ouhbi, 2016). Por otra parte, también la extracción de frases relevantes facilita la construcción de modelos de representación de textos, por ejemplo, en forma de grafo, lo cual constituye otro aspecto relevante para la MT (Chang y Kim, 2014).

Se han reportado varias soluciones a la extracción automática de frases relevantes, con enfoques supervisados (Hulth, 2003; Grineva, Grinev y Lizorkin, 2009; López y Romary, 2010) y no supervisados (Mihalcea y Tarau, 2004; Liu et al., 2009; Bougouin, Boudin y Daille, 2013; Thi, Nguyen y Shimazu, 2016; Martínez, Araujo y Fernández, 2016). Sin

ISSN 1135-5948

embargo, estas soluciones aún muestran bajas tasas de precisión y bajo rendimiento (Hasan y Ng, 2014; Merrouni, Frikh, y Ouhbi, 2016), por lo que la solución a esta problemática aún constituye un espacio propicio para la innovación.

En este trabajo se propone un nuevo método no supervisado para extraer frases relevantes en textos, cuya principal contribución está en combinar el uso de patrones léxico-sintácticos para extraer las frases candidatas con una estrategia mejorada (respecto a soluciones similares del estado del arte) de análisis de tópicos para determinar las frases relevantes. El método se diseñó en cuatro fases y ofrece la capacidad de procesar textos en español e inglés. Se evaluaron dos variantes del método con los corpus SemEval-2010 e INSPEC y se compararon los resultados con los obtenidos por otras propuestas del estado del arte. Los resultados obtenidos superan los reportados por la mayoría de las propuestas incluidas en la comparación, y son muy similares a los de la mejor propuesta en cada corpus.

Este artículo está organizado de la siguiente forma: en la Sección 2 se resume el análisis de los trabajos del estado del arte relacionados con la propuesta, en la Sección 3 se describe el método propuesto, en la Sección 4 se analizan los resultados de los experimentos realizados, y en la Sección 5 se exponen las conclusiones.

2 Trabajos relacionados

Las soluciones que automatizan la extracción de frases relevantes en textos suelen diseñarse en 4 fases: pre-procesamiento, identificación y selección de frases candidatas, determinación de frases relevantes y evaluación (Merrouni, Frikh y Ouhbi, 2016). Estas soluciones se clasifican según el enfoque que implementan para determinar las frases relevantes a partir de las frases candidatas identificadas, siendo estos: supervisado y no supervisado (Hasan y Ng, 2014; Merrouni, Frikh y Ouhbi, 2016).

Los métodos supervisados se caracterizan por aplicar algún tipo de algoritmo de aprendizaje automático (Hulth, 2003), y algunos también utilizan recursos de conocimiento externo, como Wikipedia (Grineva, Grinev y Lizorkin, 2009; López y Romary, 2010). Este enfoque responde a un modelo de predicción de frases relevantes en nuevos documentos, a partir de otros documentos, en los cuales han sido identificadas manualmente las frases relevantes. *HUMB* (López y Romary, 2010) es uno de los métodos supervisados más conocidos por los buenos resultados que ha obtenido con diferentes *dataset*, aunque está orientado a extraer frases relevantes en artículos científicos. En este método se identifican y procesan solo las principales secciones de los artículos para identificar los términos candidatos, siendo estos los términos que poseen hasta cinco palabras y no empiezan ni terminan con palabras vacías. Se utilizan las bases de conocimiento *GRISP* y *Wikipedia* para extraer características léxico/semánticas de los términos y los árboles de decisión para evaluar los términos y seleccionar las frases relevantes. Aunque este tipo de métodos, generalmente, logran mejores tasas de precisión, tienen entre sus limitaciones: ser dependientes de un dominio, y requerir corpus de entrenamiento para los algoritmos de aprendizaje, lo que implica que si se cambia el dominio de aplicación se necesita invertir tiempo en el reentrenamiento de esos algoritmos (Merrouni, Frikh y Ouhbi, 2016).

Los métodos no supervisados tienen la ventaja de utilizar solo la información contenida en los documentos de entrada para determinar las frases relevantes. Ejemplos de estos métodos se reportan en (Thi, Nguyen y Shimazu, 2016; Martínez, Araujo y Fernández, 2016; Bougouin, Boudin y Daille, 2013; Liu et al., 2009; Mihalcea y Tarau, 2004).

En *TextRank* (Mihalcea y Tarau, 2004) los términos candidatos y sus relaciones se representan en un grafo, cuyos vértices representan los términos y los arcos representan relaciones de co-ocurrencia entre ellos. Luego se construye un grafo no ponderado y no dirigido, sobre el cual se aplica un algoritmo similar a *PageRank* (Brin y Page, 1998) para determinar la relevancia de cada vértice. Posteriormente, se seleccionan los N mejores vértices, siendo N la tercera parte de los vértices del grafo. Finalmente, los términos relevantes son marcados en el texto y las secuencias de palabras adyacentes son seleccionadas como frases relevantes.

En *TopicRank* (Bougouin, Boudin y Daille, 2013) se propone una estrategia basada en la identificación y análisis de tópicos para extraer las frases relevantes, con muy buenos resultados. En este método se extraen las secuencias más largas de sustantivos y

adjetivos del texto como frases candidatas. Las frases sustantivas similares se agrupan en una sola entidad, tratada como un tema o tópico, usando un algoritmo de Agrupamiento Aglomerativo Jerárquico (HAC, por sus siglas en inglés) (Müllner, 2011). Luego, se construye un grafo donde cada vértice representa un tema y los arcos (etiquetados con un peso) representan sus relaciones. El peso del arco representa la fuerza de la relación semántica existente entre un par de temas, entendiéndose aquí relación semántica como la cercanía existente en el texto entre las frases candidatas que agrupa un tema con respecto a las que se agrupan en otro tema. Luego, se selecciona una frase relevante por cada tema, según uno de los siguientes criterios: la frase candidata con mayor frecuencia, la que primero aparece en el texto o la que tiene el rol de centroide. La selección de una frase relevante por cada tema constituye una limitación en esta propuesta, ya que alrededor de un tema se pueden agrupar más de una frase relevante en un mismo texto. Liu et al. (2009) también consideran el agrupamiento de frases candidatas como parte de la extracción de frases relevantes, pero este se realiza a partir del análisis de una medida de distancia semántica.

Martínez, Araujo y Fernández (2016) proponen un método para extraer frases relevantes a partir de artículos científicos, considerando solo las secciones de: título, resumen, introducción, trabajos relacionados y conclusiones. En este método se identifican las frases sustantivas como frases candidatas, y se representan como vértices en un grafo, donde los arcos representan el nivel de relación semántica existente entre cada par de frases. Para extraer las frases relevantes se seleccionan las posibles mayores secuencias de palabras sin solapamiento, descartando las que tienen 3 o 4 términos, sin contar las palabras vacías: ‘de’, ‘por’ y ‘a’, y ponderando el peso de las frases extraídas del título, el resumen y la introducción. Las frases relevantes se determinan usando el algoritmo *PageRank* (Brin y Page, 1998), y analizando la frecuencia de aparición en el texto de las frases candidatas.

En (Thi, Nguyen y Shimazu, 2016) se propone el uso de patrones sintácticos para identificar frases candidatas, los cuales tienen como base las frases sustantivas, pero también incorporan verbos y participios. Utilizan el método TF-IDF como parte de la evaluación de la relevancia de esas frases, seleccionándose al final las quince mejores evaluadas como frases relevantes. En esta propuesta se aprecian los beneficios del uso patrones sintácticos para incrementar de la capacidad de extracción de frases candidatas, no obstante, aunque obtiene buenos resultados, estos no logran ser mejores que los obtenidos por otras soluciones.

Según Merrouni, Frikh y Ouhbi (2016), los métodos no supervisados ofrecen mayores fortalezas que los supervisados, pero tienen como debilidad que los basados en grafos no garantizan que todos los temas principales del documento sean representados por las frases relevantes extraídas y no logran alcanzar una buena cobertura del documento. Precisamente, en el nuevo método que se propone, se combinan un conjunto de elementos dirigidos a reducir estas debilidades y mejorar los resultados en la extracción de frases relevantes.

3 Método propuesto

El método fue concebido sobre la base de combinar el uso de patrones léxico-sintácticos con una estrategia basada en el análisis de tópicos. El mismo se diseñó en cuatro fases: pre-procesamiento, identificación de temas, evaluación de temas y selección de frases relevantes. En el método se incluyó el uso de patrones léxico-sintácticos para extraer frases candidatas en la fase de pre-procesamiento, se propone un nuevo criterio de agrupamiento y dos condiciones de parada para éste en la fase de identificación de temas, tomando como referencia el método *TopicRank*. Además, se incorpora un mecanismo mejorado de selección de frases relevantes que permite extraer más de una frase relevante por cada tema, para resolver la limitación identificada en *TopicRank*.

3.1 Pre-procesamiento

En esta fase se ejecutan diferentes tareas de PLN con el objetivo de extraer la información sintáctica del texto de entrada requerida en el proceso de extracción de frases candidatas. La fase se inicia con la extracción del texto plano del fichero de entrada, el cual puede estar en diferentes formatos. El texto extraído es segmentado en párrafos y oraciones, y cada oración es fragmentada en el conjunto de *tokens* que la componen (ej. palabras, números, signos de puntuación, etc.). Posteriormente, se

realiza el análisis sintáctico superficial del texto usando el analizador sintáctico *Freeling*. El uso de *Freeling* ofrece al método propuesto la ventaja de procesar textos en inglés y español. El proceso concluye con la obtención del árbol sintáctico del texto, a partir del cual se obtienen las frases candidatas.

La extracción de frases candidatas se basa en la identificación de aquellas frases que puedan constituir conceptos, para lo cual fueron definidos un conjunto de patrones léxico-sintácticos, los que se muestran en la Tabla 1.

Categorías	Patrones
"sn" (sintagma nominal)	[D | P] + [<s-adj>] + NC
	[D | P] + NC + [<s-adj>]
	[D] + NP
	NC
"s-adj" (sintagma adjetivo)	([R] + [A | VPN])
	<s-adj> + (C |Fc) + <s-adj>
"sadv" (sintagma adverbial)	R
Leyenda: NC: sustantivo común; NP: sustantivo propio; D: determinante; VPN: verbo participio; P: pronombre; C: conjunción; R: adverbio; A: adjetivo; Fc: coma; +: concatenación; |: disyunción; <>: categoría sintáctica; []: opcional; (): agrupación	

Tabla 1: Patrones léxico-sintácticos

Estos patrones han sido formalizados a partir del etiquetado gramatical que realiza *Freeling* y en ellos se combinan un conjunto de categorías gramaticales relevantes en la composición de frases conceptuales. Tienen sus orígenes en el trabajo de Rodríguez y Simón (2013), así como en los patrones más frecuentes a partir de los cuales están formados los conceptos incluidos en la ontología del proyecto *DBpedia* (Lehmann et al., 2012). Las frases candidatas son extraídas a partir de la identificación de estos patrones en el árbol sintáctico del texto. Los autores consideran que con los patrones definidos se incrementan las capacidades para la extracción de frases candidatas, con respecto a otras propuestas que solo tienen en cuenta frases sustantivas, y se contribuye a lograr una mayor cobertura del documento. Según lo reportado en (Thi, Nguyen, y Shimazu, 2016), el uso de otros tipos de palabras, además de las que incorporan las frases sustantivas, puede mejorar la evaluación de los métodos de extracción de frases relevantes.

3.2 Identificación de temas

La identificación de los temas iniciales está basada en un mecanismo de agrupamiento de frases candidatas, en el cual se tiene en cuenta el peso de las relaciones entre cada par de frases candidatas. Este proceso se lleva a cabo de manera similar a *TopicRank*, proponiéndose otro criterio para realizar el agrupamiento. Específicamente, además del cálculo del peso de las relaciones mediante la similitud sintáctica usada en *TopicRank*, se incorpora el cálculo de la distancia entre palabras en el texto. Por lo general, las palabras que se encuentran cerca en el texto, o en un mismo contexto, suelen estar relacionadas a un mismo tema, por tanto, el uso de la distancia en palabras entre frases como criterio de agrupamiento propicia la formación de grupos de frases que tengan un fuerte vínculo contextual y con ello se logra una mejor representación e identificación de temas. En el cálculo de la similitud sintáctica entre dos frases se plantea que: dos frases son similares sintácticamente si tienen al menos un 25% de palabras traslapadas (Bougouin, Boudin y Daille, 2013). Por otra parte, la distancia promedio en palabras que existe entre cada par de frases se calcula según la fórmula (1).

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} D(a,b) \qquad (1)$$

Donde *a* es cada palabra dentro de la frase candidata FC_i (con una cantidad *A* de palabras) y *b* es cada palabra de la frase FC_j (con una cantidad de *B* palabras), por lo que las frases que estén cerca en el texto entre si tendrían una menor distancia.

El agrupamiento de frases candidatas en temas se ejecuta usando el algoritmo HAC, en correspondencia con lo usado en *TopicRank*, y teniendo en cuenta los criterios definidos para el tratamiento de las relaciones entre cada par de frases candidatas. Este proceso se lleva a cabo mediante la creación de una matriz cuadrada simétrica de tamaño *n* (número de frases candidatas), como se muestra en la Figura 1, donde *A*, *B*, *C*, …, *G* constituyen los temas. Inicialmente, se considera a cada frase como un tema y como se aprecia en la Figura 1, cada tema identifica una fila y una columna. La intersección entre cada par de temas contiene el peso de la relación del par de frases que representen los temas correspondientemente.

	A	B	C	D	E	F	G
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

Figura 1: Matriz cuadrada simétrica de temas

En cada iteración se agrupan el par de temas cuyas relaciones sean las que tengan mayor valor de peso, si se aplica similitud sintáctica como criterio de agrupamiento, y las que tengan menor valor de peso, cuando se aplica el criterio de distancia en palabras. Entre las estrategias de vinculación más usadas, en *TopicRank* (Bougouin, Boudin y Daille, 2013) se propone usar la vinculación promedio porque representa una compensación entre la vinculación completa y la simple. Mediante el uso de la vinculación promedio, el peso de la relación entre el nuevo tema T_x y el tema T_k, denotado por $R(T_x, T_k)$, se calcula según la fórmula (2).

$$R(T_x, T_k) = \frac{R(T_i, T_k) + R(T_j, T_k)}{2} \quad (2)$$

Siendo T_i y T_j los temas que se unifican para conformar T_x. En cada iteración, al agrupar dos temas en un nuevo tema se recalculan las distancias asociadas a sus relaciones con el resto de los temas. En la Figura 2 se muestra, como ejemplo, el resultado de la iteración 1, a partir de la matriz de la Figura 1 (iteración 0).

	A	B	(C,E)	D	F	G
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

Figura 2: Matriz con el agrupamiento de dos temas y el recalculo de relaciones

El agrupamiento de frases candidatas se detiene cuando se cumpla alguna de las condiciones de parada definidas en el método propuesto. Este método incorpora dos condiciones de parada a considerar en el caso que se utilice como criterio de agrupamiento el cálculo de la distancia en palabras entre frases, y una condición de parada cuando el criterio de agrupamiento esté guiado por la similitud semántica; aspecto no especificado claramente en (Bougouin, Boudin y Daille, 2013).

Con relación al primer caso: la primera condición consiste en agrupar mientras la menor distancia sea mayor o igual a la distancia promedio entre cada par de frases candidatas del texto, y la segunda condición consiste en agrupar mientras que el nuevo tema a formar tenga una cantidad de frases menor o igual que la cantidad promedio de frases por párrafo en el texto. Con relación al segundo caso: la condición de parada definida consiste en agrupar mientras la mayor similitud sea mayor o igual que un 25 %, teniendo en cuenta el criterio propuesto por Bougouin, Boudin y Daille (2013) para determinar si dos frases son similares sintácticamente.

La fase concluye con una representación del texto, mediante un grafo completo, en el cual los temas se representan como vértices y estos se conectan mediante arcos etiquetados con el peso de la relación entre los temas. Cada peso representa la fuerza de la relación semántica existente entre el par de temas. El tema *A* y el tema *B* tiene una fuerte relación semántica si las frases candidatas que agrupan cada uno aparecen cerca en el texto con frecuencia. Considerando esto, el peso W_{ij} de un arco se calcula según las fórmulas (3) y (4). La fórmula (4) hace referencia a la distancia recíproca entre las posiciones de las frases candidatas c_i y c_j en el texto, donde $pos(c_i)$ representa todas las posiciones (p_i) de la frase candidata c_i.

$$W_{i,j} = \sum_{c_i \in T_i} \sum_{c_j \in T_j} D(c_i, c_j) \quad (3)$$

$$D(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|} \quad (4)$$

3.3 Evaluación de temas

A partir del grafo construido en la fase anterior se procede a evaluar cada tema, teniendo en cuenta como modelo de evaluación lo propuesto en *TextRank* (Mihalcea y Tarau, 2004), y usando la fórmula (5).

$$S(T_i) = (1 - \lambda) + \lambda * \sum_{T_j \in V_i} \frac{W_{i,j} * S(T_i)}{\sum_{T_k \in V_j} W_{j,k}} \quad (5)$$

donde V_i constituye el conjunto de temas adyacentes a T_i en el grafo, que son los temas

que aportan a su evaluación, y λ es un factor de amortiguado que generalmente es 0,85 (Brin y Page, 1998). En este modelo se asigna una puntuación de significación para cada tema, basado en el concepto de "votación": los temas con mayor puntuación contribuyen más a la evaluación del tema T_i conectado.

3.4 Selección de las frases relevantes

En el método se propone realizar la selección de frases relevantes asociadas a cada tema según el uso (independiente o combinado) de los siguientes criterios:

- frase candidata que primero aparece en el texto.
- frase candidata más frecuentemente usada.
- frase candidata que más relación tiene con las demás de cada tema (rol de centroide).

Estos criterios también se usan en *TopicRank*, pero de forma poco flexible porque solo se puede tener en cuenta uno de ellos para seleccionar una frase relevante por cada tema. Aunque esto evita la ocurrencia de redundancias (Bougouin, Boudin y Daille, 2013), también puede afectar la cobertura en el proceso de extracción de frases relevantes. En este nuevo método se implementa un mecanismo que posibilita combinar los tres criterios mencionados, según los intereses del usuario, dando la posibilidad de extraer más de una frase relevante por cada tema. En el caso específico de existir más de una frase con la mayor frecuencia en un tema, se toman todas, si la frecuencia no es 1, porque en ese caso solo se tomaría la primera frase que aparece en el texto y se descarta el criterio de la frecuencia. En este sentido, el método propuesto fue implementado de tal forma que la selección de los criterios de agrupamiento y las condiciones de parada definidas puedan ser configurables por un usuario, para ofrecer una mayor flexibilidad en su ejecución.

4 Resultados experimentales

El método propuesto fue evaluado utilizando los corpus de prueba SemEval-2010 (Kim et al., 2010) e INSPEC (Hulth, 2003) y los resultados fueron medidos usando las métricas de Precisión (P), Cobertura (C), y la medida-F (F). Los textos contenidos en estos corpus están escritos en inglés, y en la Tabla 2 se resume una caracterización de cada uno.

Corpus	Textos	Tipos	Frases Relevantes
SemEval-2010	100	Artículos científicos	1482 (aprox. 15 por texto)
INSPEC	500	Resúmenes de artículos científicos	4913 (aprox. 10 por texto)

Tabla 2: Caracterización de los corpus

Los experimentos se realizaron con dos variantes implementadas del método, donde en cada una de ellas se combinan el uso de los tres criterios de selección de frases relevantes, pero se utiliza un criterio de agrupamiento diferente. Esto permite tener una percepción más clara de los aportes de cada criterio de agrupamiento por separado. Variantes evaluadas:

- Propuesta (V1): variante que usa el criterio de similitud sintáctica entre frases;
- Propuesta (V2): variante que usa el criterio de distancia en palabras entre frases.

Ambas variantes se evaluaron con cada uno de los corpus y los resultados se compararon con los obtenidos por otros métodos del estado del arte, que reportaban los mejores resultados con esos corpus; según la bibliografía consultada e independientemente del enfoque. La mayoría de las propuestas incluidas en estas comparaciones son no supervisadas, aunque también se incluye el método *HUMB*. Las Tablas 3 y 4 muestran los resultados obtenidos con SemEval-2010 e INSPEC, respectivamente. Los métodos evaluados fueron ordenados según los valores de la medida-F, en correspondencia con la estrategia de *ranking* usada en SemEval-2010.

Según se aprecia en las Tablas 3 y 4, el método propuesto obtiene muy buenos resultados de forma general, ya que en cada uno de los corpus una de las dos variantes evaluadas ha quedado en segunda posición, mejorando los resultados de la mayoría de los métodos incluidos en la comparación. En ambos corpus se logran mejorar los resultados del método *TopicRank*, siendo significativa esta mejora en el caso de la evaluación con INSPEC.

Es de destacar el estrecho margen que se aprecia entre los resultados de los métodos que mejores resultados reportan con cada corpus, con respecto a los obtenidos por alguna de las variantes evaluadas. En las pruebas realizadas con SemEval-2010, el método de Martínez,

Araujo y Fernández (2016) solo supera en 1,3% el valor de medida-F obtenido por V2, siendo muy similares también los valores obtenidos de precisión y cobertura. Este resultado tiene mayor relevancia considerando que esa propuesta está diseñada específicamente para extraer frases relevantes en artículos científicos, lo que no ocurre con el método propuesto. En las pruebas realizadas con INSPEC, el método de Liu et al. (2009) supera en apenas 0,4% el valor de la medida-F obtenido por V1, aunque la precisión alcanzada por V1 es ligeramente superior. No obstante, la propuesta de Liu et al. (2009) tiene la ventaja de utilizar Wikipedia, mientras que el método propuesto no requiere el uso de algún recurso de conocimiento externo. En la evaluación realizada con este corpus, se destaca el valor de precisión alcanzado por V2, superior a todas las propuestas incluidas en la comparación, pero su cobertura resultó ser baja.

Métodos	P (%)	C (%)	F (%)
(Martínez, Araujo y Fernández, 2016)	32,2	**33,2**	**32,8**
Propuesta (V2)	30,8	32,3	31,5
(Bougouin, Boudin y Daille, 2013)	**37,6**	25,8	30,3
Propuesta (V1)	36,4	23,2	28,3
(López y Romary, 2010)	27,2	27,8	27,5
(Samhaa y Rafea, 2010)	24,9	25,5	25,2

Tabla 3: Resultados con SemEval-2010

Métodos	P (%)	C (%)	F (%)
(Liu et al. 2009)	35,0	**66,0**	**45,7**
Propuesta (V1)	35,2	63,8	45,3
(Thi, Nguyen, y Shimazu, 2016)	38,1	46,1	41,7
Propuesta (V2)	**55,8**	30,1	39,1
(Mihalcea y Tarau, 2004)	31,2	43,1	36,2
(Bougouin, Boudin y Daille, 2013)	36,4	39,0	35,6

Tabla 4: Resultados con INSPEC

Otras conclusiones a mencionar, son: (1) con el uso de la distancia en palabras entre frases como criterio de agrupamiento (V2) se obtienen mejores resultados sobre textos extensos; y (2) con el uso de la similitud sintáctica entre frases (V1) se obtienen mejores resultados sobre textos cortos. Mediante la aplicación de V2 sobre textos cortos se extrae menor cantidad de frases relevantes, con respecto a V1, ya que en esta última variante el índice de agrupamiento de frases candidatas es menor que en V2 y por tanto se crean una mayor cantidad de temas. Esto propicia que, en textos cortos, con V2 se obtenga mayor precisión y menor cobertura, sucediendo lo contrario con V1. Por otro lado, con textos extensos esto no ocurre de la misma manera ya que, en este escenario, con V2 se extrae una mayor cantidad de frases relevantes que con V1 y por tanto su precisión resulta ser más fácilmente afectada, aunque su cobertura resulta ser potenciada. Son muy positivos los valores de cobertura obtenidos por las variantes del método mejor evaluadas con cada corpus, siendo muy similares a los mejores resultados reportados, lo cual se debe, en gran medida, a los patrones léxico-sintácticos definidos.

En general, los resultados expuestos demuestran la utilidad de combinar el uso de los patrones léxico-sintácticos definidos, con la estrategia de análisis de tópicos sustentada en *TopicRank*. A través de esos patrones, se incrementan las capacidades para extraer las frases candidatas en los textos, incorporando otros tipos de palabras, como adverbios y participios, en la identificación de esas frases; elementos no incluidos en otras propuestas. También resultó ser ventajosa la flexibilización de la utilización de los criterios definidos para seleccionar las frases relevantes de cada tema.

5 Conclusiones

En este trabajo se presentó un nuevo método no supervisado para la extracción de frases relevantes en textos en español e inglés, en el cual se combinó el uso de patrones léxico-sintácticos para extraer las frases candidatas con una estrategia mejorada de análisis de tópicos para determinar las frases relevantes. El uso de esos patrones posibilitó incrementar las capacidades de extracción de frases candidatas en los textos, e incrementar la cobertura del documento. Las mejoras incorporadas a la estrategia de análisis de tópicos, respecto a su extensión y flexibilización, también propiciaron que se alcanzaran mejores resultados que propuestas similares. Los resultados obtenidos demuestran la validez de la propuesta realizada, colocándola entre los métodos de

mejores resultados con los corpus de Semeval-2010 e INSPEC, respecto a los incluidos en la comparación realizada.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET), financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Economía y Competitividad (MINECO), Ref. TIN2016-76843-C4-2-R.

Bibliografía

Brin, S., y L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30(1-7):107–117.

Bougouin, A., F. Boudin, y B. Daille. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. En *Proceedings of 6th Int. Joint Conf. on NLP*, páginas 543–551.

Chang, J. Y., y I. M. Kim. 2014. Research Trends on Graph-Based Text Mining. *Int. Journal of Software Engineering and Its Applications*, 8(4):37-50.

Grineva, M., Grinev, y D., Lizorkin. 2009. Extracting Key Terms From Noisy and Multi-theme Documents. En *Proceedings of the 18th Int. Conf. on WWW*. páginas 661-670.

Hasan, K. S. y V. Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. En *Proceedings of the 52nd Annual Meeting of the ACL*. páginas 1262–1273.

Hulth, A. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. En *Proceedings of the 2003 Conf. on Empirical Methods in NLP*, páginas 216–223.

Kim, S. N., O. Medelyan, M. Y. Kan, y T. Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval'10)*, páginas 21-26.

Liu, Z., P. Li, Y. Zheng, y M., Sun. 2009. Clustering to Find Exemplar Terms for Keyphrase Extraction. En *Proceedings of the 2009 Conf. on Empirical Methods in NLP*, páginas 257-266.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 1:1-27.

López, P., y L. Romary. 2010. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval'10)*, páginas 248–251.

Martínez, J., L. Araujo, y A. D. Fernández. 2016. SemGraph: Extracting Keyphrases Following a Novel Semantic Graph-Based Approach. *Journal of the Assoc. for Info. Science and Technology*, 67(1): 71–82.

Merrouni, Z. A., B. Frikh, y B. Ouhbi. 2016. Automatic Keyphrase Extraction: An Overview Of The State Of The Art. En *Proceedings of the 4th IEEE Int. Colloquium on Information Science and Technology (CiSt)*, páginas 306-313.

Mihalcea, R., y P. Tarau. 2004. TextRank: Bringing Order into Texts. En *Proceedings of the 2004 Conf. on Empirical Methods in NLP*. páginas 404-411.

Müllner, D. 2011. Modern hierarchical, agglomerative clustering algorithms. *CoRR*, abs/1109.2378.

Rodríguez, A., y A. Simón. 2013. Método para la extracción de información estructurada desde textos. *RCCI*, 7(1): 55-67.

Samhaa, R. El-B. y A. Rafea. 2010. KP-Miner: Participation in SemEval-2. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval '10)*. páginas 190–193.

Thi, T., M. L. Nguyen, y A. Shimazu. 2016. Unsupervised Keyphrase Extraction: Introducing New Kinds of Words to Keyphrases. *AI'16, LNCS 9992*. páginas 665–671.