ISSN 1870-4069

Extracción automática de frases clave para idioma inglés utilizando patrones léxicos

Yanet Hernández-Casimiro[1,2], Yulia Ledeneva[1],
René-Arnulfo García-Hernández[1], José-Luis Tapia-Fabela[1]

[1] Universidad Autónoma del Estado de México,
Instituto Literario,
México

[2] Cátedras COMECyT,
México

{yhernandezcs,joseluis.fabela}@gmail.com
yledeneva@yahoo.com, renearnulfo@hotmail.com

Resumen. En este artículo se presenta un método para la tarea de extracción automática de frases clave utilizando secuencias frecuentes maximales y patrones léxicos para el idioma inglés. El conjunto de datos utilizado para la evaluación del método es Inspect. Primeramente, se realiza el proceso de preparación de los textos para poder utilizarlos en el método de generación de patrones léxicos, ya que estos son obtenidos de los datos de entrenamiento y posteriormente convertidos en patrones de búsqueda, mediante el uso de expresiones regulares, permitiendo considerar secuencias de caracteres (signos de puntuación y *stopwords*) que determinen la forma en la que aparece una frase clave, sin depender de sus características sintácticas o semánticas y obtener un listado de posibles frases candidatas que puedan representar el contenido del documento. Además, para seleccionar las mejores, se consideraron únicamente aquellas que obtuvieron una ponderación alta. Se puede observar que los resultados por el método propuestos fueron satisfactorios, comparados con los métodos del estado del arte, obteniendo un rendimiento de 91.40% en precisión y 29.47% en recuerdo para el top 5, mientras que para el top 10 un 32.60% en precisión y 21.02% en recuerdo.

Palabras clave: Frases clave, patrones léxicos, secuencias frecuentes maximales.

Automatic Keyphrases Extraction in English language Using Lexical Patterns

Abstract. In this paper we present a method for the automatic extraction of key phrases using maximum frequent sequences and lexical patterns for the English language. The data set used for the evaluation of the method is Inspect. Firstly, the process of preparing the texts is carried out in order to be able to use them in the method of generating lexical patterns, since these are obtained from the

training data and later converted into search patterns, through the use of regular expressions, allowing to consider sequences of characters (punctuation marks and stopwords) that determine the form in which a key phrase appears, without depending on its syntactic or semantic characteristics, and obtain a list of possible candidate phrases that can represent the content of the document. In addition, to select the best, only those that will obtain a high weight will be considered. The results for the proposed method were satisfactory, compared with the methods of the state of the art, obtaining a performance of 91.40% in precision and 29.47% in recall for the top 5, while for the top 10 a 32.60% in precision and 21.02% in recall.

Keywords: Key phrases, lexical patterns, maximal frequent sequences.

1. Introducción

En la actualidad, la generación de documentos digitales disponibles en la web ha crecido de forma exponencial debido al uso de la tecnología y digitalización, lo que ha provocado una necesidad en la búsqueda y gestión de información relevante para proporcionar un resumen del contenido principal de un documento.

Una manera de poder resolver este problema es realizar la asignación de frases, palabras o términos clave a una colección de documentos o un solo documento, que proporcionen una forma de identificar y caracterizar su contenido de manera concisa, para ayudar al lector a decidir si la información que encuentra es la adecuada para él [1, 2, 3, 4].

Una frase clave es un conjunto representativo de palabras, frases o términos que pueden resumir el contenido de un documento e identificar sus tópicos principales. Además, de acuerdo con la posición y tipo de texto en el que se encuentren, cumplen con un determinado objetivo; cuando son colocadas en la primera página de un artículo, estas apoyan en la generación de un breve resumen al lector y son asignadas por su autor, que van desde cinco hasta quince frases; si estas son encontradas dentro del índice de una revista, se puede considerar como un tema de indexación, para encontrar de forma rápida un artículo relevante.

Sin embargo, cuando se encuentran dentro de un motor de búsqueda, el objetivo es realizar "búsquedas" de documentos de forma precisas que coincidan con la frase.

Estos tres objetivos mencionados, cumplen con una característica en común, considerar una lista de frases que capturen los tópicos principales en los documentos [5, 6, 7, 8, 9, 10].

El proceso de asignación de frases clave es una tarea considerada como complicada, ya que es un proceso manual, el cual implica demasiados esfuerzos en tiempo, costo y recursos humanos; siendo esta última, llevada a cabo por un experto de determinado domino, que implementa tesauros de acuerdo con la temática del documento, para poder normalizar las palabras que se asignan a cada texto. Por lo que algunas de las palabras asignadas al documento no se pueden encontrar en él y esto provoca confusiones o malas interpretaciones por parte del lector, de ahí la importancia de contar con una forma resumida que represente el contenido principal de un documento y un proceso automatizado para generarlas [11, 12].

Esta actividad es llamada Extracción Automática de Frases Clave o en inglés *Automatic Keyphrase Extraction* (AKE) , en la cual se generan frases a partir del texto de un documento fuente, que facilita la búsqueda e indexación de un gran volumen de colecciones de textos en formato digital cuyas mejoras en esta tarea podrían conducir a aumentar el rendimiento en algunas otras de la rama del Procesamiento de Lenguaje Natural (PLN) como: Extracción de información [13], clasificación [14], agrupación [15], traducción automática [16] y algunas otras tareas de procesamiento de información [2, 10, 17, 18, 19, 20].

Debido a su importancia y apoyo en tareas de PLN, se han propuesto diferentes enfoques para la AKE que van desde métodos supervisado y no supervisado; donde los primeros consideran la tarea como un problema de clasificación, mientras que los segundos usan otro tipo de técnicas como el agrupamiento [21, 22, 23].

En este artículo, se propone un sistema no supervisado para la extracción automática de frases clave, que comprende dos actividades principales; la primera es el proceso de descubrimiento de patrones léxicos, mediante Secuencias Frecuentes Maximales (SFM) y la segunda se encarga de la selección de frases clave, a partir de la implementación de una serie de pesados, derivados de la evaluación de cada patrón léxico. Los resultados de este artículo fueron comparados con los obtenidos por los métodos del estado del arte.

Las contribuciones de este artículo son la: propuesta de pesados para la selección de las mejores frases clave que representen el contenido principal de un documento. Así como la implementación del método en un conjunto de datos en idioma inglés.

El resto del artículo está organizado de la siguiente manera: sección 2, se realiza un análisis de los trabajos relacionados; sección 3, se da una descripción del método propuesto; sección 4, se presentan los resultados experimentales y, por último, en la sección 5, mostramos las conclusiones y trabajos futuros.

2. Trabajos relacionados

La tarea de AKE se clasifica en dos enfoques principales: supervisado y no supervisado. En los cuales se implementan diferentes subprocesos y modelos.

Un proceso general para poder llevar a cabo esta tarea está constituido por 4 pasos principales: (1) pre-procesamiento, (2) selección de candidatas, (3) ponderación y (4) construcción de un listado de frases clave final [24, 25, 26]. A continuación, se abordan trabajos desde estos dos enfoques.

Dentro de los enfoques no supervisado se puede considerar el uso de la frecuencia del término [27, 28]; frases que coincidan con etiquetado POS [1] implementando medidas de centralidad.

También existen otros métodos como los basados en patrones léxicos [25, 29] en el cual se realiza una etapa de pre-procesamiento de los textos, para previamente obtener un conjunto de patrones léxicos que pueden identificar las frases clave de un documento, mediante la asignación de pesados que determinen la importancia de una palabra, generando un listado de palabras asignadas a un documento.

donde se demuestra que el uso de patrones léxicos no dependen de la característica de frecuencia para determinar que una frase es importante a diferencia de algunos otros

Tabla 1. Patrones léxicos de acuerdo con su longitud.

Longitud l	Frase clave	Patrón léxico
l^1	OF *RELATIONSHIP*.	OF<KP>.
l^2	FOR *WEBSIT DEVELOP*,	FOR<KP><KP>,
l^3	THE *AVAIL CHEMIC DIRECTORI*,	THE<KP><KP><KP>,
l^4	OF *MICROWEL POLYMERAS CHAIN REACTION* (	OF<KP><KP><KP><KP> (

trabajos, donde ésta es una de las características básicas para considerar si una palabra es clave [10, 15, 31, 30].

También existen trabajos basados en grafos [18, 21, 13, 32, 33] donde su principal característica es la construcción de grafos donde cada uno de los nodos (vértices) es considerado como una palabra o frase candidatas, mientras que los enlaces o aristas determina la relación que existen entre un nodo y otro.

Estas relaciones se basan en la ocurrencia o relación semántica. Además de los métodos anteriores, existen aquellos que pueden combinar técnicas como la implementación de patrones léxico-sintácticos con un modelo basado en grafos, unida a la perspectiva de lógica difusa [34], donde se requiere de medidas semánticas y sintácticas, además del operador OWA (*Ordered Weighted Averaging*).

Para trabajos con enfoques supervisados, la selección de frases clave se puede considerar como un problema de clasificación. Turney [5] implementa un sistema de aprendizaje automático llamado Genex que consta de dos secciones; un algoritmo genético llamado Genitor y el Extractor, constituido por doce parámetros que son ajustados por el algoritmo genético, para luego generar una lista de frases clave.

Otro de los trabajos reconocidos es KEA (*Keyphrase Extraction Algorithm*) [35], un algoritmo simple y eficaz que se basa en un aprendizaje automático implementando Naïve Bayes, para el entrenamiento y la extracción de frases clave que consta de dos fases; (1) Entrenamiento: se crea un modelo para identificar frases clave, utilizando documentos de entrenamiento donde se conocen las frases clave del autor, (2) Extracción: elije frases clave de un documento de prueba utilizando el modelo anterior.

3. Método propuesto

El método descrito en este artículo está constituido por cuatro fases importantes para la extracción automática de frases clave, donde se utilizan las SFM para el proceso de descubrimiento de patrones léxicos, que ayudarán en la generación de una lista de frases clave candidatas.

3.1. Pre-procesamiento

Esta fase es importante para la preparación de los datos, ya que el método propuesto se basa en el uso del algoritmo de SFM [36] que solo acepta, letras y el símbolo "@", para ello es necesaria la eliminación de los caracteres especiales, puesto que no proporcionan información relevante en el texto, por ejemplo: √ *(raíz cuadrada), β (beta), Θ (theta), Σ (sigma), etc.* ; así como la limpieza, reestructuración y codificación de signos de puntuación, *stop words* y números, donde éstos últimos son transformados en una etiqueta, por ejemplo: "(" → @PaA (paréntesis que abre); ":" → @DP (dos puntos); "?" → @PreC (signo de interrogación que cierra); etc., que ayudan en el proceso de la obtención de patrones léxicos. También se aplicó un proceso de lematización utilizando el algoritmo de Porter [37].

3.2. Descubrimiento de patrones léxicos

En esta fase, se crean tres conjuntos, partiendo del entrenamiento (*train*) $\boldsymbol{D_{train}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3} \dots \boldsymbol{d_j}\}$ y prueba (*test*) $\boldsymbol{D_{test}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3} \dots \boldsymbol{d_j}\}$: El conjunto ${\boldsymbol{D_{formato}}}^{test|train}$ se obtiene de haber aplicado la fase de pre-procesamiento en $\boldsymbol{D_{test|train}}$. Y los conjuntos ${\boldsymbol{D_{búsqueda}}}^{test}$ $\boldsymbol{y}$ ${\boldsymbol{D_{contexto}}}^{\boldsymbol{train}}$, parten de haber identificado un conjunto de frases clave generadas por un experto de acuerdo con cada documento d_j y longitud l^n de la frase, ${\boldsymbol{KP_{experto}}}_{\boldsymbol{d_j}}^{\boldsymbol{l}\ test|train}$ en cada uno de los conjuntos ${\boldsymbol{D_{formato}}}^{test|train}$. Donde ${\boldsymbol{KP_{experto}}}_{\boldsymbol{d_j}}^{\boldsymbol{l}\ test|train}$ es transformado en frases clave de búsqueda ${\boldsymbol{KP_{búsqueda}}}_{\boldsymbol{d_j}'}^{\boldsymbol{l}\ test|train}$ mediante expresiones regulares.

Una vez creados estos conjuntos, se aplica el algoritmo de SFM en ${D_{contexto}}^{train}$. Para la creación de ${D_{contexto}}^{train}$ se identificaron las ${\boldsymbol{KP_{búsqueda}}}_{\boldsymbol{d_j}}^{\boldsymbol{l}\ train}$ consideradas como semillas ${\boldsymbol{KP_{semilla}}}_{\boldsymbol{k\ d_j}}^{\boldsymbol{l}\ train}$ y representadas con la etiqueta "<KP>" en el texto de cada documento de ${\boldsymbol{D_{formato}}}^{train}$, de tal manera que ${D_{contexto}}^{train}$ está constituido por un listado de las ${\boldsymbol{KP_{semilla}}}_{\boldsymbol{k\ d_j}}^{\boldsymbol{l}\ train}$ encontradas.

Para implementar el algoritmo de SFM, es necesario considerar un umbral β de aparición, en el que se toma un porcentaje aleatoria del 0 al 1 para asignar los valores de β, de acuerdo con la cantidad de ${\boldsymbol{KP_{semilla}}}_{\boldsymbol{k\ d_j}}^{\boldsymbol{l}\ train}$ identificadas en ${D_{contexto}}^{train}$ y longitud de la frase.

Si se considera que la cantidad de semillas encontradas por l^n de frase en el conjunto es de $\boldsymbol{l^{1^{train}}}$ = 28,350; $\boldsymbol{l^{2^{train}}}$ = 12,320 y $\boldsymbol{l^{3^{train}}}$ = 456; tomando un umbral β al 0.250%, los valores serían: $\boldsymbol{l^{1^{train}}}$ = 70.89≈71; $\boldsymbol{l^{2^{train}}}$ = 30.8≈31 y $\boldsymbol{l^{3^{train}}}$ = 1.14≈3, este β representa el número de apariciones que debe presentar una ${\boldsymbol{kp}_{semilla}}_{\boldsymbol{k\ d_j}}^{\boldsymbol{l}\ train}$ para poder considerarla como un patrón léxico de acuerdo a la cantidad de semillas por l^n en el texto, es por ello que en $\boldsymbol{l^{3^{train}}}$ no se puede considerar

Algoritmo 1. EVALUACIÓN DE PATRONES LÉXICOS

1: **procedimiento** EVALUACIÓN DE PATRONES LÉXICOS

2: **para cada** $l^n \in \boldsymbol{D_{train}}$ **hacer**

3: $P_{búsqueda} \leftarrow$ ER (Expresiones Regulares) **aplicar** $P_{léxicos}$

4: $KP_{ident} \leftarrow P_{búsqueda}$ **aplicar** $\mathrm{D}_{\mathrm{format}}$

5: **evaluar** resultado ← *gold estándar* **comparar** KP_{ident}

Algoritmo 2 SELECCIÓN Y EVALUACIÓN DE FRASES CLAVE

1: **procedimiento** SELECCIÓN Y EVALUACIÓN DE FRASES CLAVE

2: **para cada** d ∈ D_{test} **hacer**

3: **para cada** $\mathrm{p}_{\mathrm{búsqueda}} \in P_{búsqueda}$ **hacer**

4: **para cada** $\mathrm{kp}_{\mathrm{ident}} \in KP_{ident}$ **hacer**

5: Precisión (P) $kp_{ident} \leftarrow P_{búsqueda}$ Precisión

6: Recuerdo (R) $kp_{ident} \leftarrow P_{búsqueda}$ Recuerdo

7: F-Measure (F-M) $kp_{ident} \leftarrow P_{búsqueda}$ F-Measure

8: Booleano (B) $kp_{ident} \leftarrow$ Booleano **donde** 1 = aparece o 0 = no aparece

9: **para cada** $\mathrm{kp}_{\mathrm{ident}} \in KP_{ident}$ **hacer**

10: **sumar** peso ← kp_{ident}

el valor de uno, porque encontraría todas las semillas del conjunto, por lo que en este trabajo el β mínimo de frecuencia es considerado como tres.

Para obtener $\boldsymbol{P_{léxicos}}^{l^{train}}$, únicamente son elegidas las SFM que estén constituidas por al menos una $\boldsymbol{kp}_{semilla}{}^{\boldsymbol{l}}_{\boldsymbol{k\,d_j}}{}^{train}$ y que cumplan con la siguiente estructura: $contexto_{izq}/ < KP > /contexto_{der}$.

El contexto es considerado como la secuencia de por lo menos 1 a 20 palabras a la derecha o izquierda de una $\boldsymbol{kp}_{semilla}{}^{\boldsymbol{l}}_{\boldsymbol{k\,d_j}}{}^{train}$ identificada, que posteriormente es considerada como un patrón léxico. Estas palabras pueden estar constituidas por signos de puntuación, números, *stop words* o palabras del texto. Las semillas serán colocadas en un listado de acuerdo con la longitud de la $\boldsymbol{kp}_{semilla}{}^{\boldsymbol{l}}_{\boldsymbol{k\,d_j}}{}^{train}$, cabe señalar que la cantidad de etiquetas "<KP>" determina la longitud de la frase. En la tabla 1 se muestran algunos ejemplos de los patrones léxicos encontrados de acuerdo con su longitud.

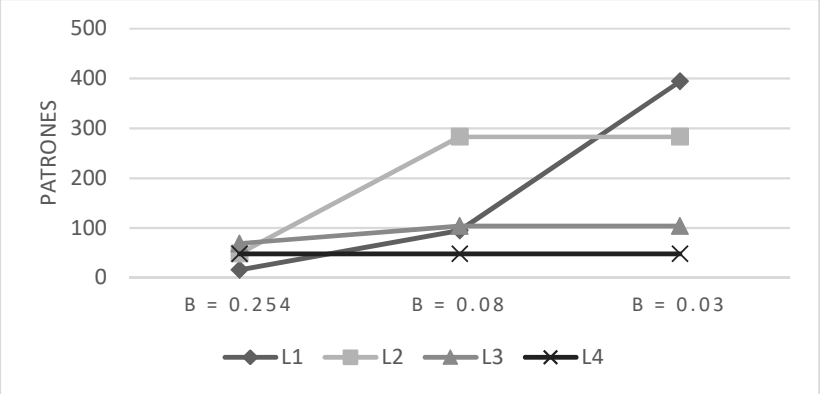

Fig. 1. Cantidad de patrones léxicos extraídos por cada β, de acuerdo con la longitud de la frase.

3.3. Evaluación de patrones léxicos

El algoritmo 1, representa el proceso realizado para poder evaluar cada uno de los patrones léxicos. El conjunto de ${P_{léxicos}}^{l^{train}}$ es obtenido del conjunto $\boldsymbol{D_{train}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3} \ldots \boldsymbol{d_j}\}$ de acuerdo con l^n. Un aspecto importante de considerar es que uno o más ${p_{busqueda}}_w^{\,l\ train}$, pueden identificar la misma $\boldsymbol{kp}_{\boldsymbol{ident}_{\boldsymbol{w\,d_j}}^{\boldsymbol{l}}}{}^{test}$ más de una vez. La evaluación se realiza para las $\boldsymbol{kp}_{\boldsymbol{ident}_{\boldsymbol{w\,d_j}}^{\boldsymbol{l}}}{}^{test}$ de cada ${p_{busqueda}}_w^{\,l\ train}$ mediante las medidas de precisión (P), recuerdo (R) y F-Measure (F-M).

3.4 Selección y evaluación de frases clave

En el algoritmo 2, se muestra el proceso de selección y evaluación de frases clave. La identificación de las frases clave está considerada de tal manera que no dependa de la frecuencia que ${kp_{ident}}_{w\,d_j}^{\,l\ \ test}$ tiene en el documento del conjunto D_{test}, sino por el número de patrones por la que es extraída de acuerdo con el rendimiento del patrón, es por ello que se toman como pesos sus respectivos valores de acuerdo a su evaluación con P, R y F-M. Además de estos valores, se consideró agregar otro pesado, que en este trabajo es nombrado Booleano (B), donde el valor de 1 se asigna siempre que ${kp_{ident}}_{w\,d_j}^{\,l\ \ test}$ sea extraída por ${p_{busqueda}}_m^{\,l\ train}$, en caso contrario el valor será ϕ.

De esta fase, se obtiene ${KP_{candidatas}}_{w\,d_j}^{\,l\ \ test}$, donde se considera el top 5 y 10 para cada documento. Por último, se evalúa ${KP_{candidatas}}_{w\,d_j}^{\,l\ \ test}$ mediante las medidas de Precisión (P) y Recuerdo (R).

Tabla 2. Rendimiento del método propuesto de los mejores β para el Top 5.

Umbral	Top 5			
	NO CONTROLADAS		COMBINADAS	
	P	R	P	R
β = 0.254-P	86.20%	38.45%	91.40%	29.47%
β = 0.08-P	78.60%	35.06%	85.80%	27.66%
β = 0.08-FM	71.40%	31.85%	77.60%	25.02%
β = 0.254-FM	70.60%	31.49%	76.00%	24.50%
β = 0.08-R	69.20%	30.87%	75.20%	24.25%
β = 0.254-R	69.60%	31.04%	74.80%	24.12%
β = 0.254-B	68.80%	30.69%	74.00%	23.86%
β = 0.03-P	61.20%	27.30%	68.80%	22.18%
β = 0.08-B	58.40%	26.05%	64.20%	20.70%
β = 0.03-B	56.60%	25.25%	62.60%	20.18%

Tabla 3. Comparación de resultados del método propuesto con el estado del arte para el Top 5.

Umbral	Top 5			
	NO CONTROLADAS		COMBINADAS	
	P	R	P	R
β = 0.254-P	86.20%	38.45%	91.40%	29.47%
β = 0.08-P	78.60%	35.06%	85.80%	27.66%
β = 0.08-FM	71.40%	31.85%	77.60%	25.02%
Gallegos [29]	17.48%	17.79%	19.44%	13.56%
KEA [35]	9.58%	8.50%	7.82%	5.46%

4. Experimentación

4.1. Datos

El conjunto utilizado para evaluar el método propuesto es Inspect [38], está constituido por 2,000 resúmenes en inglés extraídos de artículos científicos de revistas. Cada resumen tiene dos conjuntos de frases clave de oro (*gold*): un conjunto de términos controlados, los cuales son restringidos al tesauro de la base de datos Inspec; y un conjunto de términos no controlados que puede ser cualquiera de los términos adecuados, asignados por algún especialista en las disciplinas.

Tanto los términos controlados como no controlados pueden o no estar presentes en los resúmenes. El conjunto de resúmenes se dividió en tres subconjuntos: el de entrenamiento (*train*) que consta de 1 000 documentos; validación (*validation*) con 500 documentos, y un conjunto de prueba (*test*) con 500 documentos. Para este trabajo el

rendimiento se evalúa con las medidas de Precisión (P) y Recuerdo (R), además se hizo la unión de las frases clave *gold* controladas y no controladas para evaluar el rendimiento de su combinación (conjunto combinado).

4.2. Resultados

Para Inspect, se evaluaron solo los conjuntos *gold* de frases no controladas y combinadas, para poder comparar el método propuesto, con los resultados reportados en el trabajo de Gallegos [29] y el método KEA [35]. El método se probó con tres β. En la figura 1, se muestra la cantidad de patrones extraídos para cada β y cada longitud de frase.

En esta gráfica se puede visualizar que β determina la cantidad de patrones léxicos encontrados. β considera la frecuencia del patrón, de acuerdo con la longitud de la frase, entre más pequeño sea β, mayor será la cantidad de patrones léxicos encontrados. En la tabla 2, se muestran los mejores diez rendimientos obtenidos de los diferentes β, en cada conjunto de frases *gold* (no controlado y combinado) y cada pesado (P, R, F-M, B), para el top 5.

Los resultados muestran que el rendimiento del método propuesto en $\beta = 0.254$ con pesado P, es el más alto comparado con los demás β. En el conjunto no controlado se tuvo un rendimiento de $P = 86.20\%$ y $R = 38.45\%$, mientras para el conjunto combinado $P = 91.40\%$ y $R = 29.47\%$. Es por ello, por lo que el conjunto de palabras combinadas se considera con mayor énfasis en este trabajo, siendo éste la combinación de palabras que existen y no, dentro del documento. En la tabla 3, se realiza una comparación de resultados obtenidos en este trabajo, con los reportados en el trabajo de Gallegos [29] y el método KEA [35].

Como se muestra en la tabla anterior, los resultados reportados en este trabajo para el top 5 son superiores a los reportados en [29] y [35]. Para el conjunto combinado (unión de los conjuntos controlado y no controlado), en $\beta = 0.254$ se obtuvo un rendimiento de $P = 91.40\%$ y $R = 29.47\%$, quedando por un rango alto a los presentados en el estado del arte. En la tabla 4, se muestra otra comparación de resultados obtenidos para el top 10.

Para este top, el rendimiento obtenido en $\beta = 0.08$, es de $P = 32.60\%$ y $R = 21.02\%$, superando el método de [29] y [35]. Los patrones extraídos por $\beta = 0.254$ y $\beta = 0.08$ fueron de mejor calidad en comparación con otros umbrales generados por el método. En la tabla 5, se muestran algunos ejemplos de estos patrones extraídos por el método propuesto, así como las frases clave extraídas por patrón.ue la fase de pre-procesamiento es importante a la hora de preparar los datos, ya que, si existe una buena limpieza, el método propuesto en este trabajo puede dar mejores resultados, como los mostrados en este trabajo.

Uno de los retos por abordar para este método es la implementación de patrones extraídos por un conjunto de datos, aplicados a otro para poder demostrar la independencia del dominio, ya que los patrones extraídos son del mismo conjunto de datos del conjunto *train*.

Tabla 4. Rendimiento de extracción de frases clave de los mejores β para el top 10.

Umbral	Top 10			
	NO CONTROLADAS		COMBINADAS	
	P	R	P	R
β = 0.08-P	30.70%	27.39%	32.60%	21.02%
β = 0.08-FM	28.60%	25.51%	30.40%	19.60%
β = 0.08-B	27.90%	24.89%	29.60%	19.09%
Gallegos [29]	17.47%	17.79%	19.44%	13.56%
KEA [35]	9.58%	8.50%	7.82%	5.46%

Tabla 5. Patrones extraídos por el método propuesto y frases clave extraídas por patrón.

Longitud	Patrones léxicos	Frases clave extraídas
l^1	OF<KP>,	OF *TRANSFORM*,
l^1	OF<KP>AND	OF *SYSTEM* AND
l^2	FOR<KP><KP>,	FOR *SPECIF APPLIC*,
l^2	, <KP><KP>.	, *THE INTERNET*.
l^3	(<KP><KP><KP>)	(COMPRESS *GRAPHIC IMAG*)
l^3	OF<KP><KP><KP> (	OF *BUILD MANAG SYSTEM* (
l^4	OF<KP><KP><KP><KP> (	OF *ALGORITHM IN COMPUT ELECTROMAGNET* (
l^4	FOR<KP><KP><KP><KP>OF	FOR *A SOCIAL IDENT PERSPECT* OF

Agradecimientos. Trabajo realizado con apoyo del Gobierno Mexicano (COMECYT). Los autores agradecen a la Universidad Autónoma Estado del México.

Referencias

1. Boudin, F.: A comparison of centrality measures for graph-based keyphrase extraction. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 834–838 (2013)
2. Alrehamy, H., Walker, C.: Exploiting extensible background knowledge for clustering-based automatic keyphrase extraction. Soft Computing, vol. 22, pp. 7041–7057 (2018) doi: 10.1007/s00500-018-3414-4
3. Alami-Merrouni, Z., Frikh, B., Ouhbi, B. Automatic Keyphrase extraction: a survey and trends. J Intell Inf Syst, vol. 54, pp. 391–424 (2020) doi: 10.1007/s108 44-019-00558-9
4. Asl, J. R., Banda, J. M.: GLEAKE: Global and local embedding automatic keyphrase extraction. arXiv preprint arXiv:2005.09740 (2020) doi: 10.48550/ar Xiv.2005.09740
5. Turney, P. D.: Learning to extract keyphrases from text. ArXiv Preprint ArXiv, cs/0212013 (1999) doi: 0.48550/arXiv.cs/0212013

6. Nguyen, T. D., Kan, M. Y.: Keyphrase extraction in scientific publications. In: Goh, D. H. L., Cao, T. H., Sølvberg, I. T., Rasmussen, E. (eds.) Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, ICADL 2007, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol. 4822, pp. 317–326 (2007) doi: 10.1007/978-3-540-77094-7_41
7. Kim, S. N., Medelyan, O., Kan, M. Y., Baldwin, T.: Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, pp. 723–742 (2013) doi: 10.1007/s10579-012-9210-3
8. Zhang, Q., Wang, Y., Gong, Y., Huang, X.: Keyphrase extraction using deep recurrent neural networks on twitter. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 836–845 (2016)
9. Ying, Y., Qingping, T., Qinzheng, X., Ping, Z., Panpan, L.: A Graph-based approach of automatic keyphrase extraction. Procedia Computer Science, pp. 248–255 (2017)
10. Shapira, O., Pasunuru, R., Dagan, I., Amsterdamer, Y.: Multi-document keyphrase extraction: A literature review and the first dataset. arXiv preprint arXiv:2110.01073 (2021) doi: 10.48550/arXiv.2110.01073
11. Beliga, S., Martinčić-Ipšić, S.: Network-enabled keyword extraction for under-resourced languages. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) Semantic Keyword-Based Search on Structured Data Sources. IKC 2016, Lecture Notes in Computer Science, Springer, Cham. vol. 10151. pp. 124–135 (2016) doi: 10.1007/978-3-319-53640-8_11
12. Johnny, S., Jaya-Nirmala, S.: Key phrase extraction system for agricultural documents. In: Gani, A., Das, P., Kharb, L., Chahal, D. (eds) Information, Communication and Computing Technology. ICICCT 2019, Communications in Computer and Information Science, Springer, Singapore. vol. 1025, pp. 240–252 (2019) doi: 10.1007/978-981-15-1384-8_20
13. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
14. Zhang, Y., Liu, H., Wang, S., Fan, W., Xiao, C.: Automatic keyphrase extraction using word embeddings, Soft Comput, vol. 24, pp. 5593–5608 (2020) doi: 10.10 07/s00500-019-03963-y
15. Lahiri, S., Mihalcea, R., Lai, P. H.: Keyword extraction from emails. Natural Language Engineering, vol. 23, no. 2, pp. 295–317 (2017) doi: 10.1017/S13513 24916000231
16. Haque, R., Penkale, S. Way, A.: TermFinder: Log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. Lang Resources & Evaluation, vol. 52, pp. 365–400 (2018) doi: 10.10 07/s10579-018-9412-4
17. Perez-Guadarrama, Y., Simón-Cuevas, A., Hojas-Mazo, W., Olivas, J. A., Romero, F. P.: A fuzzy approach to improve an unsupervised automatic Keyphrase extraction process. In: Proceedings of IEEE International Conference on Fuzzy Systems, pp. 1–6 (2018) doi: 10.1109/FUZZ-IEEE.2018.8491487
18. Mothe, J., Ramiandrisoa, F., Rasolomanana, M.: Automatic keyphrase extraction using graph-based methods. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 728–730 (2018) doi: 10.1145/3167132.3167392
19. Saxena, A., Mangal, M., Jain, G.: KeyGames: A game theoretic approach to automatic keyphrase extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 2037–2048 (2020) doi: 10.18653/v 1/2020.coling-main.184
20. Mahata, D., Kuriakose, J., Shah, R., Zimmermann, R.: Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2, pp. 634–639 (2018) doi: 10.18653/v1/N18-2100
21. Bougouin, A., Boudin, F., Daille, B.: TopicRank: Graph-based topic ranking for keyphrase extraction. In: Proceedings of International Joint Conference on Natural Language Processing, pp. 543–551 (2013)

22. Xie, F., Wu, X., Zhu, X.: Document-specific keyphrase extraction using sequential patterns with wildcards. In: Proceedings of IEEE International Conference on Data Mining, pp. 1055–1060 (2014) doi: 10.1109/ICDM.2014.105
23. Awan, M. N., Beg, M. O.: Top-rank: A topical postion rank for extraction and classification of keyphrases in text. Computer Speech & Language, vol. 65, pp. 101116 (2021) doi: 10.1016/j.csl.2020.101116
24. Lim, V. M. H., Wong, S. F., Lim, T. M.: Automatic keyphrase extraction techniques: A review. In: Proceedings of IEEE Symposium on Computers & Informatics, pp. 196–200 (2013) doi: 10.1109/ISCI.2013.6612402
25. Hernández-Casimiro, Y., Ledeneva, Y., García-Hernández, R. A., Ramos-Corchado, M. A.: Lexical patterns based on maximal frequent secuences for automatic keyphrase extraction. Computación y Sistemas, vol. 25, no. 1, pp. 153–163 (2021) doi: 10.13053/cys-25-1-3868
26. Xie, F., Wu, X., Zhu, X.: Efficient sequential pattern mining with wildcards for keyphrase extraction. Knowledge-Based Systems, vol. 115, pp. 27–39 (2017) doi: 10.1016/j.knosys.2016.10.011
27. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management, vol. 24, no. 5, pp. 513–523 (1988) doi: 10.1016/0306-4573(88)90021-0
28. Sidorov, G.: Construcción no lineal de n-gramas en la lingüística computacional. Sociedad Mexicana de Inteligencia Artificial (2013)
29. Gallegos-Camacho, E. M., Ledeneva, Y.: Extracción de frases clave utilizando patrones léxicos a partir de resúmenes de artículos científicos. Tesis, Universidad Autónoma de Estado de México, pp. 1–119 (2016)
30. Wang, Q., Sheng, V. S., Wu, X.: Keyphrase extraction with sequential pattern mining. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 5003–5004 (2017)
31. Marsi, E., Skidar, U., Marco, C., Barik, B., Saetre, R.: NTNU-1@ScienceIE at SemEval-2017 Task 10: Identifying and labelling keyphrases with conditional random fields. In: 11th International Workshop on Semantic Evaluations SemEval17 (2017)
32. Wan, X., Xiao, J.: CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 969–976 (2008)
33. Sterckx, L., Demeester, T., Deleu, J., Develder, C.: Creation and evaluation of large keyphrase extraction collections with multiple opinions, Lang Resources & Evaluation, vol. 52, pp. 503–532 (2018) doi: 10.1007/s10579-017-9395-6
34. Barreiro-Guerrero, M., Simón-Cuevas, A., Pérez-Guadarrama, Y., Romero, F. P., Olivas, J. A.: Applying OWA operator in the semantic processing for automatic keyphrase extraction. In: Nyström, I., Hernández Heredia, Y., Milián Núñez, V. (eds) Progress in pattern recognition, image analysis, computer vision, and applications. In: Proceedings of CIARP 2019, Lecture Notes in Computer Science, Springer, Cham vol. 11896, pp. 62–71 (2019) doi: 10.1007/978-3-030-33904-3_6
35. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G.: Kea: Practical automated keyphrase extraction. Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129–152 (2005)
36. García-Hernández, R. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A.: A new algorithm for fast discovery of maximal sequential patterns in a document collection. Computational Linguistics and Intelligent Text Processing, vol. 3878, pp. 514–523 (2006) doi: 10.1007/11671299_53
37. Porter, M. F.: An algorithm for suffix stripping. Program: electronic library and information systems, vol. 14, no. 3, pp. 130–137 (1980) doi: 10.1108/eb046814

38. Hulth, A.: Enhancing linguistically oriented automatic keyword extraction. In: Proceedings of HLT-NAACL 2004, pp. 17–20 (2004)