Extracción de frases clave utilizando patrones léxicos a partir de resúmenes de artículos científicos

Esther Maritza Gallegos Camacho, Yulia Ledeneva,
René A. García Hernández, José Luis Tapia Fabela
Unidad Académica Profesional Tianguistenco, Universidad Autónoma del Estado de México
Paraje el Tejocote, San Pedro Tlaltizapan, Santiago Tianguistenco, 52640, México
maritza.gallegos.c@gmail.com, yledeneva@yahoo.com, renearnulfo@hotmail.com
Área de participación: Sistemas Computacionales

Resumen

El artículo presenta un método propuesto para la extracción automática de frases clave a partir de resúmenes de artículos científicos. Se implementa un método no supervisado que consiste en la identificación de los patrones léxicos que pueden existir al asignar frases clave de forma manual. Las pruebas se realizan sobre el corpus Inspec que consta de 1000 artículos científicos. Cada resumen pasa por un preprocesamiento, palabras vacías, lematización, posteriormente se construyen los conjuntos de datos para la generación de patrones léxicos, estos se convierten en patrones de búsqueda para las frases candidatas y por último se hace una selección. La evaluación de frases clave se realiza utilizando la herramienta *Rouge* la cual nos permite conocer las medidas de Precisión, Recuerdo y F-Medida. Las evaluaciones se aplican para frases de top-5, top-10, top-15 y top-20. Finalmente, se realiza una comparación con métodos destacados del estado del arte obteniendo para F-Medida un tercer lugar dentro de los métodos no supervisados y para la medida de precisión se obtiene el primer lugar en comparación con los métodos.

Palabras clave: extracción automática de frases clave, patrones léxicos, resúmenes científicos.

Abstract

The paper presents a proposed method for the automatic extraction of keyphrases from abstracts of scientific articles. An unsupervised method is implemented which consists of identifying the lexical patterns that may exist when assigning keyphrases manually. The tests are performed on the Inspec corpus consisting of 1000 scientific articles. Each summary goes through a preprocessing, stopwords, stemming, later the datasets are built for the generation of lexical patterns, these become search patterns for the candidate phrases and finally a selection is made. The evaluation of keyphrases is done using the Rouge tool which allows us to know the measures of Precision, Recall and F-Measure. Evaluations apply to top-5, top-10, top-15 and top-20 phrases. Finally, a comparison is made with state-of-the-art methods obtaining for F-Measure a third place within the unsupervised methods and for the measurement of precision the first place is obtained in comparison with the methods.

Keywords: *extraction of keyphrases automatically, lexical patterns, scientific abstracts.*

Introducción

En la actualidad la generación de información electrónica ha crecido de manera acelerada, por lo que este gran volumen de datos es difícil de leer y organizar. Por esta razón investigadores buscan y desarrollan la mejor forma de obtener un resumen de contenido para aprovechar el conocimiento comprendido en dicha información [Ledeneva & García, 2017].

Diversos autores como [Ortiz, 2010] definen las frases clave, afirmando que es una secuencia de una o más palabras que capturan el tema principal del documento, ya que se espera que las frases clave representen las ideas fundamentales expresadas por el autor, mientras que otros autores como [Hansan & Ng, 2010] dicen que las frases clave se refieren a un grupo de frases que representan el documento. En [Nguyen & Kan, 2007] las frases clave se definen como frases que capturan los temas principales tratados en un documento, ya que ofrece un resumen breve pero preciso del contenido del documento, que se puede utilizar para varias aplicaciones. En

este artículo, las frases clave se definen como un grupo de palabras que muestran la idea principal de un documento, facilitando a los lectores decidir si es o no es relevante para ellos.
La extracción de frases clave ha mostrado su importancia para la mejora de muchas tareas del Procesamiento de Lenguaje Natural (PLN) y la Recuperación de Información (IR, por sus siglas en inglés) [Hansan & Ng, 2014].
Los ejemplos de las aplicaciones son apoyo en la generación de un resumen de documentos [Hernández, 2016], ser la entrada de una colección de documentos [Hansan & Ng, 2010; Hulth, 2003], la indexación que auxilia a la búsqueda de documentos [Nguyen & Kan, 2007], agrupación de los documentos [Nguyen & Kan, 2007]. También se ha experimentado para diferentes dominios como, por ejemplo, en redes sociales [García, 2017], correos electrónicos [Mihalcea et al., 2016], artículos científicos [Hurt, 2010; Hernández, 2016; Padilla, 2016], noticias [Lahiri et al., 2014, Jain et al., 2016] resumes de artículos [Hulth, 2003; Popova et al., 2013; Mihalcea & Tarau, 2004].

La generación de frases clave se clasifica en dos grandes enfoques: extracción y asignación. En [Liu et al., 2009] el objetivo de la extracción de frases clave es extraer un conjunto de frases que se relacionan con los principales temas tratados en un documento dado. Por otra parte, [Nguyen & Kan, 2007] afirman que la extracción de frases clave es seleccionar frases presentes en el documento original.

La extracción de frases clave consiste en dos etapas: identificación de candidatas y selección de frases clave [Liu et al., 2009; Kim et al., 2012]. La Identificación de frases candidatas: se restringe el número de frases candidatas con el fin de limitar la complejidad, se puede limitar por su longitud o por el tipo de frase [Liu et al., 2009]. También se implementa como un proceso que filtra las frases sin importancia [Kim et al., 2012]. La selección de frases clave: en esta etapa se determina si una frase candidata es una frase clave o no [Liu et al., 2009]. Pueden ser seleccionados en función de puntuaciones de importancia, tales como frecuencias de palabras, frecuencias de frases, entre otras [Kim et al., 2012].

Los patrones léxicos son aquellos patrones que trabajan en un nivel léxico sin tomar en cuenta elementos sintácticos o semánticos, y éstos pueden ser obtenidos a partir de Secuencias Frecuentes Maximales (SFMs) [García et al., 2004; García et al., 2006; Camacho, 2015]. En [Kim & Kan, 2009] se considera a un patrón léxico como, “aquel que describe la forma en como sucede y aparece un conjunto palabras diferentes de manera periódica en un documento”. En el trabajo de [Kim & Kan, 2009] de define un patrón como un conjunto de características comunes que son distintivas y originales que al momento de estar unidas muestran propiedades en orden y equilibrio. La Real Académica Española [2014] define léxico como un vocabulario, conjunto de palabras de un idioma, o de las que pertenecen al uso de una región, a una actividad determinada, a un campo semántico dado, etc.

Este articulo está organizado en 5 secciones. En la primera sección se presenta la introducción al tema abordado, así como los conceptos importantes. La sección de trabajos relacionados, se muestra el estado del arte para la tarea de extracción de frases clave. La tercera sección, se trata de la descripción del método, donde se explica cada etapa del método utilizado. En la cuarta sección, experimentos y resultados, se presenta una descripción del corpus sobre el cual se hicieron los experimentos, los resultados obtenidos y se hace una comparación con los métodos del estado del arte. Finalmente, se presentan las conclusiones del articulo y trabajos futuros.

Trabajos Relacionados

Los investigadores hoy en día buscan la mejor forma de extraer frases clave de un documento, que pueda brindar una descripción total de éste, y englobe todo su contenido en una sola idea principal. Pero esta tarea no ha resultado fácil. Para cada objetivo se implementan diferentes métodos, herramientas, métricas, corpus, modelos, etc.

En el trabajo de [Beliga et al., 2015] se presenta un estudio de los métodos y enfoques para la tarea de extracción de palabras clave. Además de la sistematización de los métodos, ofrece una revisión exhaustiva de la investigación existente. Se abordan métodos supervisados y no supervisados, con especial énfasis en los métodos basados en grafos, así como en la extracción de palabras clave en idioma croata. Se propone el método de extracción de palabras clave basado en selectividad como un nuevo método basado en grafos no supervisado que extrae nodos de una red compleja como candidatos de palabras clave.

En [Hernández, 2016] se propone un nuevo método no supervisado independiente del lenguaje y del contexto, utilizando los patrones léxicos propuestos anteriormente en el trabajo de Camacho [2015], posicionándose entre los mejores resultados del taller de SemEval-2010 para la tarea # 5: Extracción de frases clave [Kim et. al, 2010]. Se evalúa con Precisión, Recuerdo y F-Medida, obteniendo como resultados, F-Medida para top-5 12.92%, para top-10 20.12% y para top-15 21.71%.

Camacho [2015] se realiza la detección de fragmentos de texto que puedan ser considerados candidatos para generar hipervínculos, por medio de patrones léxicos. Estos patrones fueron generados utilizando las SFMs [García et al., 2004; García et al., 2006] transformados en patrones de búsqueda por medio de expresiones regulares para localizar fragmentos de texto. Como datos se utiliza Wikipedia en español del año 2008. El proceso consiste en crear tres conjuntos de documentos, de los utilizados de Wikipedia, para la identificación, aplicación y evaluación de patrones léxicos. En la identificación de patrones se crea un proceso de etiquetado, seguido de esto se aplica un proceso para la obtención de las SFMs y así obtener los patrones léxicos que contuvieran contexto derecho e izquierdo. Para la aplicación de los patrones léxicos, se forma un proceso de conversión donde los patrones léxicos se transformaban a patrones de búsqueda y de esta forma se pueden aplicar a un texto plano con la finalidad de obtener los fragmentos de texto candidatos a ser hipervínculos. La evaluación del método se calcula por medio de los hipervínculos generados por el humano contra los extraídos por el método aplicado. Las medidas que se utiliza para conocer el rendimiento de este trabajo fueron Precisión, Recuerdo y F-Medida.

En [Mihalcea & Tarau, 2004] los autores también trabajan con el corpus Inspec, introduce *TextRank*, un modelo de clasificación basado en el grafo para el procesamiento de texto, y demuestran cómo este modelo puede ser utilizado con éxito en aplicaciones de lenguaje natural. Se propone como método no supervisado. Evalúan con Precisión, Recuerdo y F-Medida.

Los autores [Tsatsaronis et al., 2010] presentan *SemanticRank*, un algoritmo de clasificación basado en gráfos para la extracción de palabras clave y oraciones del texto. El algoritmo construye un grafo semántico utilizando enlaces implícitos, que se basan en la semántica relación entre los nodos de texto y consecuentemente clasifica los nodos utilizando diferentes algoritmos de clasificación. La evaluación comparativa frente a los métodos de estado del arte relacionados para la extracción de palabras clave muestra que *SemanticRank* se desempeña favorablemente.

En el trabajo de [Gollapalli & Caragea, 2014] estudian la extracción de frases clave de los trabajos de investigación aprovechando las redes de citas y proponen *CiteTextRank* para la extracción de frase de artículos de investigación, un algoritmo basado en grafos que incorpora evidencia tanto del contenido de un documento como de los contextos en los que se hace referencia al documento dentro de una red de citas.

En el año 2010, Kim junto a otros organizadores se realizó la tarea nombrada "Task 5: *Automatic Keyphrases extraction from Scientific Articles*" que se incluyó en el SemEval-2010. El propósito fue desarrollar sistemas de extracción automática de frases clave de artículos científicos y comparar la lista de frases propuestas por cada sistema participante, con las frases clave que fueron asignadas por seres humanos a cada uno de los artículos científicos, evaluando los resultados de manera automática. Entre los sistemas del estado del arte que obtuvo el mejor resultado en la tarea fue HUMB con un F-Medida de 27.5%. Entre los sistemas comerciales el mejor resultado fue obtenido con Alchemy con un F-Medida de 21.37% [Padilla, 2016].

Otro trabajo que utiliza el corpus de Inspec es [Hansan & Ng, 2010]. Sin embargo, éste implementa un método no supervisado, el estudio consiste en la comparación de 5 diferentes algoritmos de extracción de frases clave: *TF-IDF* (del inglés *Term Frequency – Inverse Document Frequency*) que es frecuencia de término – frecuencia inversa de documento [Lahiri et al., 2014], *TextRank* [Mihalcea & Tarau, 2004], *SingleRank* [Wan and Xiao, 2008] , *ExpandRank* [Wan and Xiao, 2008] *y KeyCluster* [Liu et al., 2009] aplicados cada uno a 4 diferentes corpus: Inspec, DUC-2001, *NUS Keypharases* corpus y ICSI. La evaluación fue con recuerdo, precisión y F-Medida.

En el trabajo de [Nguyen & Kan, 2007] se centran en la extracción frase clave en publicaciones científicas mediante el uso de las nuevas características que capturan fenómenos morfológicos que se encuentra en frases

clave científicas. El método que aplican es supervisado, para la extracción de las frases clave es por medio de máxima entropía y Naives Bayes. Utiliza la precisión para evaluar el método con 10 *fold cross validation*.

Otro trabajo relacionado a la extracción de frases clave es [Hulth, 2003], que presenta el corpus Inspec, dando un giro a lo ya realizado en [Hernández, 2016; Kim et al., 2010; Nguyen & Kan, 2007] debido a que solo utiliza los resúmenes de artículos científicos. Es un enfoque supervisado, agrega conocimientos lingüísticos a la representación, en lugar de basarse únicamente en las estadísticas (frecuencia de los términos y n-gramas). Para la selección de términos aplica tres enfoques: n-gramas (unigramas, bigramas y trigramas), NP-chunks y patrones de etiqueta POS (Part-of-Speech). La evaluación es con precisión, recuerdo y F-Medida.

En el trabajo de [Popova et al., 2013] realizan la extracción de frases clave sobre resúmenes de publicaciones científicas. Los autores se enfocan en conocer si la cantidad aplicada de palabras vacías en el preprocesamiento son un factor importante para la tarea de extracción de frases clave. El trabajo no se compara con ningún método relacionado.

Descripción del método

Etapa 1. Preprocesamiento

Al corpus electo se le aplica una serie de pasos previos a introducirlos al método, los textos se transforman a algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis.

Caracteres especiales

El primer paso consiste en sustituir aquellos caracteres especiales por letras o caracteres válidos. Por ejemplo, la letra Ñ se sustituye por N; o bien (À|Á|Â|Ã|Ä|Å) = A este cambio de caracteres debe aplicar tanto a mayúsculas como a minúsculas, de esta forma se obtiene un texto limpio de caracteres especiales para sus posteriores procesos.

Palabras vacías

También conocido en inglés como *stopwords*. Consiste en el etiquetado de palabras vacías, carentes de significado, como son preposiciones, artículos, conjunciones, etc. Los signos de puntuación y palabras vacías son importantes para la etapa de identificación de frases candidatas, por tanto, se codifican y etiquetan.

Lematización

Conocido en inglés como *stemming*. Se utiliza como parte del preprocesamiento; para este trabajo se aplicaron dos tipos de lematizado (ℓ), ambos derivados del algoritmo de [Porter, 1980]. Estos algoritmos se encuentran escritos en lenguaje Perl. El primer algoritmo aplica el lematizado a todas las frases que se encuentran en el documento, incluyendo palabras vacías. El segundo algoritmo implementado tiene una modificación en la cual, las frases que se encuentran en mayúscula no se les aplica lematización, pero las que se encuentran en minúscula sí se aplica, esto permite tener las palabras vacías escritas correctamente, a comparación del lematizado anterior.

Etapa 2. Construcción y preparación de datos

Se retoma la generación de tres conjuntos de datos, de la etapa de selección, limpieza y trasformación del método de Camacho [2015].

Etapa 3. Extracción de patrones léxicos

En esta etapa, se utiliza la técnica de minería de patrones denominada *DIMASP* [García, 2007]. Es una herramienta que es aplicada a la colección para la generación de SFMs donde los patrones léxicos son extraídos [Hernández, 2016]. En la Minería de Patrones Secuenciales, una secuencia de palabras se considera frecuente si ésta se encuentra en al menos un cierto número de documentos (umbral mínimo de frecuencia) [Camacho, 2015].

Etapa 4. Identificación de frases clave candidatas

Teniendo la colección de patrones léxicos obtenidos en la etapa 3, se transforma a una colección de patrones de búsqueda por medio de expresiones regulares. Posteriormente el conjunto de patrones de búsqueda se aplica a la colección generada en la etapa 2 (documentos en texto plano), con el fin de obtener las frases clave candidatas clasificadas por longitud.

Etapa 5. Comparación y evaluación de los patrones léxicos

En esta etapa, se comparan las frases clave candidatas extraídas a través de los patrones de búsqueda contra el conjunto de frases clave nombradas anteriormente en la colección y se aplica a cada conjunto de datos generado por longitud de frases clave. Para evaluar los resultados generados por el método, se realizó la determinación de las medidas de Precisión, Recuerdo y F-Medida, tomando el cálculo de [Camacho, 2015].

Etapa 6. Selección de frases clave

Se realiza la selección de frases clave a partir del conjunto de frases clave candidatas extraídas por los patrones de búsqueda en la etapa 4. Al conjunto de frases clave candidatas extraídas por cada patrón de búsqueda, se le asignan los valores obtenidos del cálculo de [Camacho, 2015] en la etapa 5. Se realizan diferentes enfoques de pesado: Booleano, Precisión, Recuerdo y F-Medida.

Etapa 7. Evaluación

Para evaluar las frases clave extraídas se implementaron las métricas: Precisión (P), Recuerdo (R), F-Medida por medio de la herramienta de evaluación propuesta por Lin [2004], denominada Rouge *(Recall-Oriented Understudy for Gisting Evaluation)* y elaborada para la evaluación de generación de resúmenes. Se implementa para evaluar la tarea de extracción de frases clave por los autores Mihalcea & Tarau [2004] y los autores Tsatsaronis et al. [2010]. Se encuentra escrito en lenguaje Perl y es ejecutado en el sistema operativo Ubuntu. La evaluación se realiza para frases candidatas top-5, top-10, top-15 y top-20.

Las frases clave extraídas se evalúan con las siguientes medidas utilizadas en el estado del arte como lo son:

$$P = \frac{\#\ frases\ clave\ correctas}{\#\ (frases\ clave\ correctas + \ frases\ clave\ incorrectas)} \quad (1)$$

$$R = \frac{\#\ frases\ clave\ correctas}{\#\ (frases\ clave\ correctas + \ frases\ clave\ no\ extraídas)} \quad (2)$$

$$F - Medida = \frac{2 * P * R}{P + R} \quad (3)$$

Experimentos y Resultados

A continuación, se presenta el corpus que se utilizó en los diversos experimentos y los resultados obtenidos con el modelo descrito en la sección anterior.

Corpus

El corpus utilizado en este trabajo se denomina Inspec, utilizado por primera vez en [Hulth, 2003], el cual está compuesto por 2000 resúmenes en inglés, con su correspondiente título y frases clave de la base de datos Inspec. Los resúmenes fueron extraídos de artículos de revistas de los años 1998 a 2002, comprendiendo las disciplinas computación, control, y tecnologías de la información.

Cada resumen tiene dos conjuntos de frases clave asignadas por un indexador profesional asociado a ellos: un conjunto de términos controlados, es decir, términos restringidos al tesauro de Inspec y un conjunto de términos no controlados que puede ser cualquiera de los términos adecuados, asignados libremente por los indexadores. El conjunto de resúmenes se dividió arbitrariamente en tres grupos: un conjunto de entrenamiento "*Training*" (para construir el modelo) que consta de 1000 documentos, un conjunto de validación "*Validation*" (para evaluar los modelos, y seleccionar el mejor desempeño) que consta de 500 documentos, y un conjunto de prueba "*Test*" (para obtener resultados objetivos) con los 500 resúmenes restantes. Tanto los términos controlados y los no controladas pueden o no estar presentes en los resúmenes

Para los experimentos presentados se ocupó solo el conjunto de términos no controlados, con base en [Hulth, 2003] en el cual alude que, los indexadores tuvieron acceso a los documentos completos cuando asignaron las frases clave, concluyendo así que los términos no controlados están presentes en los resúmenes en mayor medida, con un porcentaje de 76,2% a los términos controladas que solo tiene un 18.1%.

Resultados

Se realizaron diferentes experimentos cada uno con diferentes parámetros para poder determinar cuáles fueron los que dieron mejores resultados. Se utilizó el evaluador *Rouge* el cual nos proporciona la F-Medida. El primer parámetro que se probó fue el umbral de frecuencia, esto para la obtención de patrones. Se dedujo que con umbral de 0.1% se obtienen mejores resultados de acuerdo a [Camacho, 2015]. Posteriormente se aplicó los dos diferentes preprocesamientos mencionados en la etapa 4.

Por otra parte, considerando que el corpus original no contenía un delimitador que indicara la separación del título con el texto, así como el final de cada resumen, se creó preciso la colocación de un punto que nos permitiera considerar el título de cada resumen, de igual forma un punto al final del texto que indicara la terminación del resumen que se va a procesar. Otro procesamiento utilizado fue la reducción de cada resumen, consta de eliminar la última oración de cada resumen con el fin de reducir el espacio de búsqueda de patrones léxicos.

En la tabla 1, se presentan los patrones léxicos que se obtuvieron con los resultados más altos al aplicar el método y extraer las frases clave. Se puede notar que en la mayoría de los patrones aparece el punto (@PUNTO), lo que nos indica que en su mayoría se ocupa para la identificación de una frase clave. En la tabla 2, se muestra de forma concisa los resultados obtenidos y los parámetros utilizados para cada uno de ellos. Cabe señalar que solo se tomaron los mejores resultados de cada experimento realizado y es para un top de 20 frases clave.

Tabla 1. Los 10 mejores patrones léxicos de longitud 1, 2 y 3.

Longitud 1	Longitud 2	Longitud 3
OF @LINK @COMA A	AND THE @LINK @LINK @COMA	@COMA THE @LINK @LINK @LINK AND THE
IN @LINK @PUNTO WE	A @LINK @LINK @PUNTO WE	USING @LINK @LINK @LINK @PUNTO THIS PAPER
@CBR @LINK @CCI @PUNTO	WITH @LINK @LINK @PUNTO IN	WITH A @LINK @LINK @LINK @PUNTO
SUCH AS @LINK AND	USING THE @LINK @LINK OF	A @LINK @LINK @LINK AND A
@COMA AND @LINK @PUNTO	THE @LINK @LINK @PUNTO IT IS	A @LINK @LINK @LINK @PUNTO A
AND THE @LINK @PUNTO	ON @LINK @LINK @PUNTO THIS	AND THE @LINK @LINK @LINK @PUNTO
ON THE @LINK @COMA	OF @LINK @LINK @PUNTO TO	A @LINK @LINK @LINK @PUNTO THE
OF @LINK AS	EFFECT OF @LINK @LINK ON THE	THE @LINK @LINK @LINK @PUNTO THIS
THE @LINK @PUNTO THIS	WITH @LINK @LINK @PUNTO	IN @LINK @LINK @LINK USING
OF @LINK ON	FOR @LINK @LINK @PUNTO THIS PAPER	BY @LINK @LINK @LINK @PUNTO IN

Tabla 2. Resultados para la extracción de frases clave para diferentes parámetros.

Experimento	Parámetros				Evaluación		
	Umbral de frecuencia	Lematizado	Consideración de titulo	Reducción de resumen	Precisión	Recuerdo	F-Medida
Propuesto 1	0.1 %	ℓ =1	No	No	44.78	28.81	33.11
Propuesto 2	0.1%	ℓ =1	Si	No	44.27	29.44	33.40
Propuesto 3	0.1%	ℓ =2	Si	No	43.51	29.97	33.53
Propuesto 4	0.1%	ℓ =1	Si	No	42.92	27.54	31.71
Propuesto 5	0.1%	ℓ =2	Si	No	41.79	30.01	33.21
Propuesto 6	0.1%	ℓ =2	Si	Si	45.17	27.91	32.58
Propuesto 7	1%	ℓ =1	No	No	36.51	11.51	16.58

Comparación de resultados

Los resultados obtenidos se compararon con trabajos del estado del arte, métodos supervisados y métodos no supervisados. En nuestro resultado más alto obtuvimos de F-Medida 33.53 logrando con este posicionarnos en el tercer lugar sobre los métodos no supervisados, y en segundo lugar comparándonos con métodos supervisados. A pesar de que nos encontramos en el segundo lugar comparándonos con métodos supervisados, es importante resaltar que los métodos no supervisados no hacen uso de datos anotados, como diccionarios, tesauros, vocabularios, tal como lo hacen los supervisados, que necesitan una gran cantidad de datos de entrenamiento, lo que los hace ser dependientes del lenguaje y contexto, por esa razón la investigación se inclina sobre los no supervisados, que pueden ser independientes del contexto.

En la tabla 3, se presenta el desempeño de los métodos del estado del arte distinguiendo los no supervisados de los supervisados. Muestra los parámetros utilizados con su evaluación de F-medida, así como los mejores resultados del método propuesto. En la figura 1, se presenta una comparación de manera gráfica de la Tabla 3.

Tabla 3. El desempeño del método propuesto y los métodos del estado del arte rankeado de mayor a menor.

Enfoque	Método	Parámetros	F-Medida
No Supervisado	TextRank 1 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 2	36.20
	TextRank 2 [Mihalcea & Tarau, 2004]	Directo Co-occ. = 2	35.90
	Propuesto 3	Umbral 0.1%, *l* = 2 con titulo	33.53
	Propuesto 2	Umbral 0.1%, *l* = 1, con titulo	33.40
	Propuesto 5	Umbral 0.1%, *l* = 2, con titulo	33.21
	Propuesto 1	Umbral 0.1%, *l* = 1, sin titulo	33.11
	TextRank 3 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 3	32.60
	Propuesto 6	Umbral 0.1%, *l* = 2, con titulo	32.58
	TextRank 4 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 5	32.20
	TextRank 5 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 10	32.20
	Propuesto 4	Umbral 0.1%, *l* = 1, con titulo	31.71
Supervisado	N-gramas 1 [Hulth, 2003]	Con etiquetado POS	33.90
	NP-Chunking 1 [Hulth, 2003]	Con etiquetado POS	33.00
	Patrones 1 [Hulth, 2003]	Con etiquetado POS	28.10
	Patrones 2 [Hulth, 2003]	Sin etiquetado POS	25.60
	NP-Chunking 2 [Hulth, 2003]	Sin etiquetado POS	22.70
	N-gramas 2 [Hulth, 2003]	Sin etiquetado POS	17.60

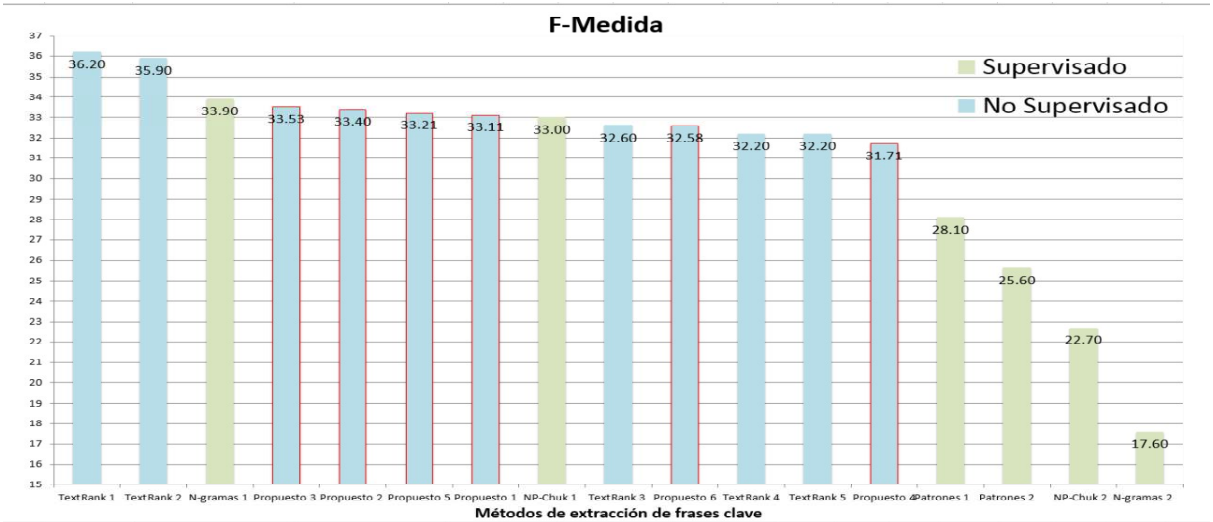

Figura 1. El desempeño del método propuesto y los métodos del estado del arte.

Con los resultados de precisión, de todos los propuestos, logramos posicionarnos en primer lugar sobre todos los métodos de [Mihalcea & Tarau, 2004] y [Hulth, 2003], logrando así la principal aportación del trabajo. Con estos resultados afirmamos que el método de extracción de frases clave utilizando patrones léxicos tiene una alta presión sobre los demás métodos. También se reafirma que es un método independiente del contexto, ya que en los propuestos 4 y 5 se utilizaron patrones léxicos del trabajo de [Hernández, 2016] y se aplicaron a este trabajo, dando buenos resultados.

En la tabla 4, se presenta la precisión de los métodos supervisados y no supervisados, así como los mejores resultados del método propuesto. Nos permite observar que nos encontramos por encima de los demás métodos, con lo que podemos determinar de una forma concreta la aportación de este método. En la figura 2, se presenta una comparación de manera gráfica de la Tabla 4.

Tabla 4. La precisión del método propuesto y los métodos del estado del arte rankeado de mayor a menor.

Enfoque	Método	Parámetros	Precisión
No Supervisado	Propuesto 6	Umbral 0.1%, ℓ = 2, con titulo	45.17
	Propuesto 1	Umbral 0.1%, ℓ = 1, sin titulo	44.78
	Propuesto 2	Umbral 0.1%, ℓ = 1 con titulo	44.27
	Propuesto 3	Umbral 0.1%, ℓ = 2, con titulo	43.51
	Propuesto 4	Umbral 0.1%, ℓ = 1, con titulo	42.92
	Propuesto 5	Umbral 0.1%, ℓ = 2, con titulo	41.79
	TextRank 1 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 2	31.20
	TextRank 2 [Mihalcea & Tarau, 2004]	Directo Co-occ. = 2	31.20
	TextRank 3 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 3	28.20
	TextRank 4 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 5	28.20
	TextRank 5 [Mihalcea & Tarau, 2004]	Indirecto Co-occ. = 10	28.10
Supervisado	NP-Chunking 1 [Hulth, 2003]	Con etiquetado POS	29.70
	N-gramas 1 [Hulth, 2003]	Con etiquetado POS	25.20
	Patrones 1 [Hulth, 2003]	Con etiquetado POS	21.70
	NP-Chunking 2 [Hulth, 2003]	Sin etiquetado POS	16.70
	Patrones 2 [Hulth, 2003]	Sin etiquetado POS	15.90
	N-gramas 2 [Hulth, 2003]	Sin etiquetado POS	10.40

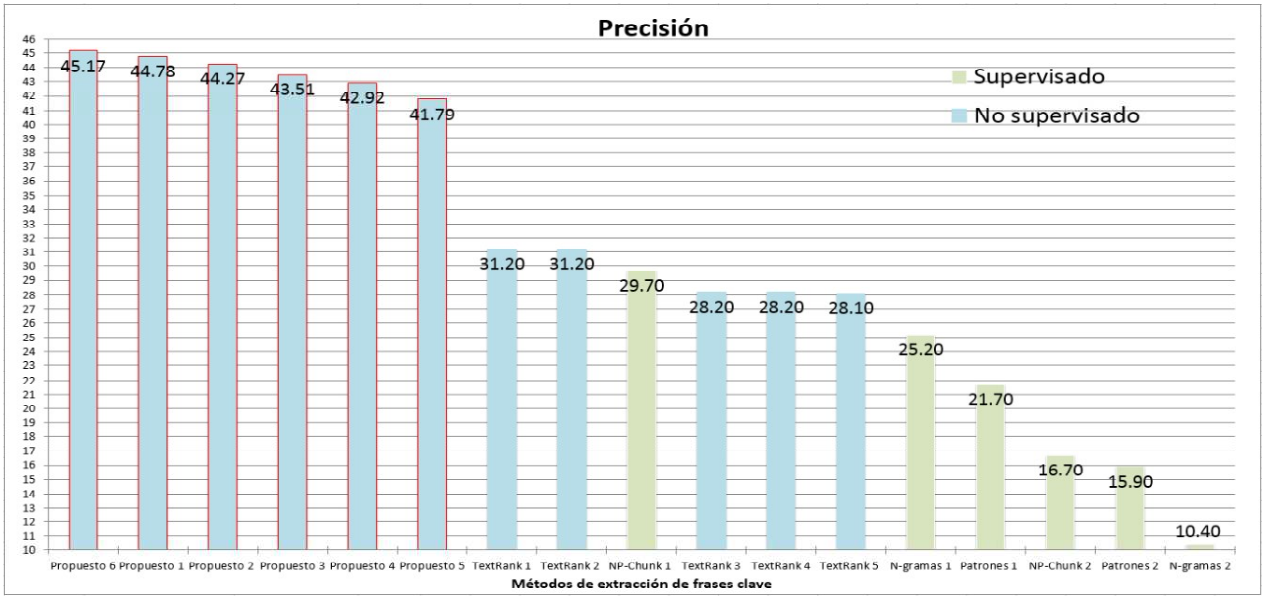

Figura 2. La precisión del método propuesto en comparación los métodos del estado del arte.

Conclusiones

Una de las principales conclusiones fue que se pudo comprobar que los patrones léxicos extraídos de resúmenes de artículos científicos son suficientes para extraer las mejores frases clave de acuerdo a las evaluaciones realizadas.
En la aplicación del método se observó que los experimentos presentan los mejores resultados para la evaluación de precisión, se nota mayor tendencia en combinación 2, que consta de palabras de longitud 1 y longitud 2. Los resultados de precisión nos llevaron a posicionarnos por encima de los métodos del estado del arte presentados, el mejor resultado de Precisión fue 45.17 mientras que el método que más se aproxima a nosotros cuenta con una Precisión de 31.20 [Mihalcea, 2004].
Otro parámetro importante es el umbral de frecuencia para la búsqueda de patrones. Con un umbral de 0.1% se pudieron obtener patrones suficientes, el corpus Inspec mostró ser útil para la tarea de extracción de frases clave, a pesar de tener información corta ya que solo esta constituido por resúmenes de artículos científicos. A diferencia del umbral de frecuencia de 1% que la información corta no es suficiente. Esto se puede afirmar en base al resultado de F-Medida del propuesto 7 que obtuvo 16.58 y el propuesto 1 que obtuvo 33.11.
Otra conclusión importante se deriva de la consideración del título de cada resumen, así como un delimitador que nos permitió conocer la finalización del texto. En el propuesto 2, donde se aplicaron estas dos consideraciones mencionadas anteriormente, se obtiene una F-Medida de 33.40 en comparación con el propuesto 1 donde no se considera el título ni la finalización de texto se obtiene 33.11; cabe señalar que los demás parámetros eran los mismos. De esta manera se observa una mejora en los resultados confirmando que la consideración del título es fundamental para una mejor obtención de frases clave.
Cabe resaltar que los mejores resultados obtenidos a partir de diversas pruebas fueron para top-20 frases candidatas con un F-Medida de 33.53, y obteniendo para precisión 45.17 donde el umbral de frecuencia fue 0.1%, aplicando en el preprocesamiento el segundo algoritmo de Porter, considerando el título y la finalización del resumen.
Por último, pero no menos importante, se corroboro que el método de patrones léxicos es independiente del contexto, esta conclusión se toma de los propuestos 4 y 5 que obtuvieron buenos resultados de F-Medida 31.71 y 33.21 respectivamente. Los patrones aplicados en estos propuestos fueron obtenidos del corpus semEval 2010, el cual está constituido por artículos científicos completos [Hernández, 2016]. Estos patrones se aplicaron al corpus Inspec que como se menciona anteriormente está formado solo de resúmenes de artículos científicos.

Trabajo a Futuro

Como trabajo futuro se probará el método para el idioma español, para corroborar que es independiente del lenguaje, así como se comprobó que es independiente del contexto, como hasta el momento no se ha encontrado corpus en idioma español, se desea crear un corpus en este idioma y ser probado con otros métodos como TextRank que muestra el mejor desempeño en el estado del arte.

Referencias

1. Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences, 39(1), 1-20.
2. Camacho, A. Marcela. (2015). Detección de Fragmentos de Texto como Candidato a Hipervínculo. UAEM. Tesis de Maestría. Tianguistenco, Edo de México. p. 73.
3. García-Hernández, R.A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. (2004). A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text, In: Sanfeliu, A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. (eds.) CIARP 2004. LNCS (3287) 478-486. Springer-Verlag.
4. García-Hernández, R.A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. (2006). A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. In: Gelbukh, A. (ed.) CICLing 2006, LNCS (3878) 514–523, Springer-Verlag.
5. García Hernández Rene Arnulfo. (2007). Desarrollo de Algoritmos para el Descubrimiento de Patrones Secuenciales Maximales. INAOE. Puebla, México 2007.
6. García Martínez Detzani.(2017), Extracción de palabras clave en twitter utilizando patrones léxicos., Tesis para obtener el título de Ingeniería de Software. Universidad Autónoma del Estado de México.

7. Gollapalli, S. D., & Caragea, C. (2014). Extracting Keyphrases from Research Papers Using Citation Networks. In AAAI (pp. 1629-1635).
8. Hasan, K. S., and Ng, V. (2010). Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. Coling 2010: Poster Volume. p. p. 365- 373.
9. Hasan, K. S., and Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. p. p. 1262–1273.
10. Hernández Casimiro, Y. (2016). Extracción de Frases Clave Usando Patrones Léxicos en Artículos Científicos. Tesis Maestría en Ciencias de la Computación. UAEM. p.124.
11. Hulth, A. (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Department of Computer and Systems Sciences Stockholm University Swedez. p. 8.
12. Hurt, C. D., (2010) Automatically Generated Keywords: A Comparison to AuthorGenerated Keywords in the Sciences. Journal of Information and Organizational Sciences, 34(1):81-88.
13. Jain, N., Gupta, S., & Patel, D. (2016). E3: Keyphrase based News Event Exploration Engine. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (pp. 327-329). ACM.
14. Kim, S. N. and Kan, M-Y. (2009). Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles. Proceedings of the Workshop on Multiword Expressions Identification Interpretation Disambiguation and Applications. p. 8.
15. Kim, S. N., Medelyan, O., Kan, M-Y. And Baldwin, T. (2010). SemEval-2010 TAsk 5: Automatic Keypharases Extraction from Scientific. Proceedings of the 5th International Workshop on Semantic Evaluation, ACL. p. p. 21–26.
16. Kim, S. N., Medelyan, O., Kan, M-Y. And Baldwin, T. (2012). Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation. p. 22.
17. Lahiri, S; Choudhury, S. R; Caragea, C. (2014) Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks, Cornell University Library, arXiv preprint arXiv:1401.6571.
18. Ledeneva Nikolaevna Yulia, García Hernández Rene Arnulfo. (2017). Generación automática de resúmenes. Retos, propuestas y experimentos. Editorial CIGOME S.A. de C.V. Primera edición
19. Liu, Z., Li, P., Zheng, Y. and Sun, M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. Department of Computer Science and Technology State Key Lab on Intelligent Technology and Systems National Lab for Information Science and Technology Tsinghua University. Beijing. p. p. 257-266.
20. Mihalcea, Rada and Tarau, Paul. (2004). TextRank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. p. p. 404–411.
21. Mihalcea Rada, Lahiri Shibamouli, Lai Po-Hsiang, (2016). Keyword Extraction from Emails, Journal of Natural Language Engineering (JNLE).
22. Nguyen, T. D. and Kan, M-Y. (2007). Keyphrase Extraction in Scientific Publications. Department of Computer Science, School of Computing. p. 10.
23. Ortiz Méndez, R. C. (2010), Detección Automática de Temas Importantes. Benemérita Universidad Autónoma de Puebla Facultad de Ciencias de la Computación, Puebla, Tesis de Licenciatura.
24. Padilla Camacho, Jesús Ernesto (2016). Evaluación de sistemas de extracción de frases clave. Tesis de Licenciatura. UAEM.
25. Popova, S., Kovriguina, L., Mouromtsev, D., & Khodyrev, I. (2013). Stop-words in keyphrase extraction problem. In Open Innovations Association, November 2013 14th Conference of IEEE. p.p.113-121.
26. Porter, M.F. (1980). An algorithm for suffix stripping. Computer Laboratory. (14) No.3, pp. 130-137.
27. Real Academia Española y Asociación de Academias de la Lengua Española (2014). «diccionario». Diccionario de la lengua española (23.ª edición). Madrid: España. ISBN 978-84-670-4189-7.
28. Tsatsaronis, George. Varlamis, Iraklis & Nørvåg, Kjetil (2010). SemanticRank: Ranking Keywords and Sentences Using Semantic Graph. Proceedings of the 23rd International Conference on Computational Linguistics Beijing, August 2010. p. p. 1074–1082.
29. Wan, Xiaojun and Jianguo Xiao. (2008). Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pages 855–860.