SISTEMAS PARA LA EXTRACCIÓN DE FRASES CLAVE EN DOCUMENTOS CIENTÍFICOS

Gerardo Flores Petlacalco
Benemérita Universidad Autónoma de Puebla
gerardo.florespe@alumno.buap.mx

Mireya Tovar Vidal
Benemérita Universidad Autónoma de Puebla
mtovar@cs.buap.mx

Hilda Castillo Zacatelco
Benemérita Universidad Autónoma de Puebla
hilda@cs.buap.mx

José A. Reyes-Ortiz
Universidad Autónoma Metropolitana
jaro@correo.azc.uam.mx

Resumen

En este documento se describen dos sistemas para la extracción de frases clave en textos científicos. El primer sistema usa la generación de *n-gramas* y posteriormente se realiza la discriminación de términos candidatos usando reglas empíricas. El segundo sistema se basa en la construcción de patrones para la eliminación de frases candidatas. Además, se hace una comparación de estos sistemas con otros propuestos que realizan la misma tarea y se muestran los resultados obtenidos en la evaluación.

Palabras Claves: Frases clave, *n-gramas*, patrones.

Abstract

In this document, we describe two systems for keyphrase extraction on scientific texts. The first system use n-gram generation and candidate term discrimination

using empirical rules. The second system is based in the patterns construction for candidate phrases elimination. Further, we do a systems comparison with other approaches that perform the same task and we show the evaluation results.
Keywords**:** *keyphrases, n-grams, patterns.*

1. Introducción

Las palabras clave tienen como funcionalidad capturar la información importante del contenido de un texto con el objetivo de ayudar a los lectores al momento de estudiar o resumir dando una idea general del tema que aborda [Siqueira, 2015]. Por otro lado, en los sistemas computacionales, su funcionalidad radica en los sistemas de indexación. Estos recurren muchas veces a las palabras para agruparlas en tópicos y hacer sencilla una recuperación de información en caso de alguna consulta.

La selección de palabras clave no es una tarea sencilla, se tienen que evitar términos muy generales puesto que se corre el riesgo de que sean intrascendentes o demasiado objetivos, que provoquen que los lectores no los encuentren por desconocimiento de ellos. Una adecuada combinación de ambas características y técnicas de selección como pueden ser la frecuencia de ciertos términos en el documento o la identificación de conceptos fundamentales que describan nombres, acciones o características del trabajo nos darán métodos para identificar las palabras clave adecuadas que contengan lo que el autor quiere explicar en su trabajo. Este proceso se hace de forma manual por parte del creador del documento o expertos, sin embargo, es una tarea compleja y pesada que requiere mucho tiempo, además de que no está exenta de fallos. Por esta razón es que áreas del Procesamiento de Lenguaje Natural ven este problema como una oportunidad para implementar modelos que extraigan de forma automática palabras clave que ayuden a realizar la selección correspondiente.

SemEval es una serie de evaluaciones para sistemas computacionales de análisis semántico, desde el 2012 publican problemas donde se involucra el área de Procesamiento de Lenguaje Natural (PLN) y todos sus campos de estudio. En el año 2017 se propusieron un total de 12 tareas, cada una con sus respectivos

objetivos, evaluaciones y recursos que implican áreas de estudio específicas. Para objeto de estudio en este trabajo se dará solución a la tarea número diez que tiene como objetivo crear un Sistema de extracción de palabras clave y relaciones semánticas aplicado a textos científicos [Augenstein, 2017]. La tarea consta de tres subtareas:

- Extracción de palabras clave en textos.
- Clasificación de palabras clave identificadas.
- Extracción de relaciones de sinonimia o hiperonimia en las frases clave identificadas.

En esta investigación se aborda la subtarea 1, que tiene como objetivo la creación de un sistema que realice de forma automática una extracción de frases clave en publicaciones científicas, los sistemas propuestos deben recibir como entrada un extracto del texto y devolver las frases clave del mismo.

Con el objetivo de resolver esta subtarea, nuestro sistema es un enfoque basado en la extracción y discriminación de *n-gramas* sobre los textos procesados, la discriminación se realiza siguiendo reglas empíricas y posteriormente se adiciona una regla más al método para obtener mejores resultados.

Durante el desarrollo de esta investigación se realizó un estudio del estado del arte en materia de extracción de frases clave, entre los trabajos consultados se encuentran los siguientes:

- [Matsuo, 2010] presentaron una propuesta de algoritmo para extraer palabras clave pertenecientes a un corpus, usan un modelo probabilístico que evalúa la ocurrencia de los términos dentro de los documentos y después miden su rendimiento con base a una métrica llamada χ^2-medida.
- [Park, 2010] usan un método basado en Naïve Bayes donde extraen las palabras clave en documentos científicos bien estructurados, consideran la posición de un candidato en el título, encabezado o cuerpo del texto para contar sus ocurrencias dentro de las distintas secciones y asignarle un puntaje, los más altos son seleccionados como palabras clave.

- [Thuy Dung, 2010] presentan un enfoque basado en la estructura lógica del texto para determinar si un candidato podría ser o no frase clave basado en su ocurrencia dentro de las secciones que lo conforman, esto limita mucho la cantidad de candidatos que son considerados y hace más sencilla la evaluación. Además, se puede combinar con otras técnicas de extracción para mejorar el rendimiento.
- [Stuart, 2010] describen a RAKE, un sistema para la extracción de palabras clave en textos individuales sin necesidad de que pertenezcan a un corpus. El enfoque usa la premisa de que las palabras clave raramente contienen puntuación y palabras vacías, bajo este enfoque se extraen candidatos que son divididos en términos individuales, luego a cada término se le asigna un puntaje basado en la frecuencia dentro del texto y a cada candidato se le calcula una evaluación usando la suma de sus términos. Se ordenan los resultados y los mejores puntajes son los elegidos.
- [Ouyang, 2010] proponen PolyU, un sistema para la obtención de frases usando la identificación y posterior expansión de palabras núcleo. Éstas serán conseguidas por medio de la frecuencia dentro del cuerpo siguiendo la premisa de que palabras muy repetidas serán consideradas como importantes, después de la identificación cada palabra será expandida por medio de una combinación de la frecuencia de su contexto y un patrón *PoS tagger*.
- [Ortiz, 2010] crean un sistema que combina dos técnicas para el descubrimiento de las palabras claves en textos científicos. Dichas técnicas son las secuencias de frecuencia máxima y el algoritmo de *PageRanking*. La secuencia de frecuencia máxima se realiza usando *n-gramas*, seleccionando aquellas que tengan alta ocurrencia para hacer un ordenamiento por *PageRanking* y determinar las palabras clave.

En la investigación proponemos un enfoque basado en el uso de *n-gramas* que se extraen desde el cuerpo del texto y forman nuestra colección de términos candidatos, usamos reglas de discriminación empíricas para disminuir la cantidad

y finalmente obtenemos los más relevantes por medio de patrones morfológicos obtenidos sobre los términos clave de un conjunto de datos de entrenamiento proporcionado por SemEval 2017 Tarea 10. El documento está organizado de la siguiente manera, iniciamos en la sección 2 donde presentamos los sistemas para la extracción de frases clave, en la sección 3 se presentan los resultados obtenidos y finalmente una discusión del trabajo en la sección 4.

2. Métodos

En esta sección mostramos los sistemas propuestos para dar solución a la subtarea 1 de SemEval 2017 Tarea 10 que tiene como objetivo la extracción de frases clave en textos científicos. Se explican dos sistemas, el primero se basa en un enfoque de extracción de *n-gramas* y discriminación de términos usando reglas empíricas obtenidas por observación. El segundo, continúa con el enfoque del primero, pero se le añade una mejora al pre-procesado y una nueva regla de discriminación usando patrones obtenidos después de un etiquetado *PoS Tagger*, es decir su etiqueta gramatical, sobre un conjunto de entrenamiento.

Primer Sistema GMBUAP

El primer sistema, nombrado como GMBUAP, es un enfoque que inicia con una extracción de *n-gramas* de un texto, después se hace una discriminación usando reglas empíricas conseguidas por observación sobre el conjunto de datos de entrenamiento proporcionado por SemEval 2017 Tarea 10 [Augenstein, 2017].
Este sistema consta de las siguientes fases, que se ilustran en el algoritmo de la figura 1:

- Pre-procesamiento de los documentos. Al texto de los documentos se eliminaron caracteres extraños como corchetes o no imprimibles. Dejando únicamente los símbolos: paréntesis, comas, puntos, guion medio, guion bajo, comillas, punto y coma, diagonal, llaves. Posteriormente, el documento fue dividido en oraciones.
- Formación de *n-gramas*. Cada oración del texto fue dividida en términos y se encontraron todas las posibles combinaciones de *n* palabras. En esta

aproximación decidimos formar gramas que van de 1 hasta una longitud máxima de 5 (*1-grama, … ,5-gramas*).

```
Entrada: Texto_científico
Salida: Conjunto de palabras_clave
Inicio
      Gramas <- []
      TextoPreprocesado <- Preprocesamiento(Texto_científico)
      Oraciones <- PartirEnOraciones(TextoPreprocesado)
      Para cada oracion en Oraciones hacer:
            GramasDeLaOracion <- CrearGramas(oracion)
            Gramas <- Gramas + GramasDeLaOracion
      Finpara
      //Parte de la discriminación de candidatos
      NuevosCandidatos <- Regla_A(Gramas)
      NuevosCandidatos <- Regla_B(NuevosCandidatos)
      PalabrasClaveDelTexto <- Regla_C(NuevosCandidatos)
      Return PalabrasClaveDelTexto
Fin
```

Figura 1 Algoritmo del Primer sistema (GMBUAP).

- Discriminación de candidatos. Usando la observación sobre los datos de la tarea, fueron propuestas tres reglas empíricas para la reducción de candidatos:
  a. Eliminar candidatos con palabras vacías (*stopwords*) al inicio y al final. Esta regla se creó bajo la premisa de que las frases clave identificadas en las anotaciones no contenían estas palabras al inicio. Por lo tanto, se decidió eliminar las que estaban al final puesto que dan por entendido que la idea sigue, pero se vio truncada en el momento de generar los gramas (ver resultado de regla A en la tabla 1). Además, en este paso eliminamos aquellas frases candidatas de longitud un carácter.
  b. Eliminar candidatos que no formen parte del texto. El pre-procesado ocasiona que algunos caracteres considerados como no imprimibles fueran eliminados y con esas deficiencias se formaron los *n-gramas*. Al tener este problema las frases clave candidatas formadas pueden no estar completas en el texto al hacer el mapeo y por consiguiente son eliminadas (ver resultado regla B en la tabla 1).
  c. Eliminar candidatos que no tengan ambos paréntesis. Las palabras clave pueden tener paréntesis dentro de ellas pero es necesario que posean ambos, el paréntesis de inicio y el paréntesis de cierre (ver resultado regla C en la tabla 1).

Tabla 1 Ejemplo del funcionamiento de las reglas.

Texto	Frases candidatas	Regla	Resultado
…such as X-ray absorption spectroscopy (XAS) and X-ray emission spectroscopy (XES) at…	**and** X-ray emission spectroscopy	A	Rechazada
…such as X-ray absorption spectroscopy (XAS) and X-ray emission spectroscopy (XES) at…	X-ray emission spectroscopy (XES)	B	Aceptada
Frase clave: X-ray emission spectroscopy (XES)	emission spectroscopy (XES	C	Rechazada

En la tabla 1 se muestra un ejemplo de la aplicación de las reglas mencionadas anteriormente, en la columna uno se muestra un extracto de un documento, en la columna dos un ejemplo de frase candidata, es decir, un *5-grama*, en la columna 3 la regla aplicada y el resultado de esta en la columna 4.

Segundo Sistema

El segundo sistema continua con el enfoque de creación de candidatos con base a *n-gramas* y siguiendo las técnicas de discriminación antes descritas. Sin embargo, en esta aproximación se hacen operaciones adicionales al pre-procesado y se añade una regla de discriminación que utiliza patrones.

La finalidad de añadir otras operaciones al pre-procesado responde a una deficiencia localizada en el paso de la división de un texto en oraciones debido a abreviaciones encontradas en el cuerpo. Estas abreviaciones interfieren al momento de realizar el corte del texto puesto que nuestro criterio es hacerlo al encontrar el carácter punto '.'. Para evitar estos conflictos reemplazamos las abreviaciones con el carácter punto en su cuerpo de tal forma que el proceso pudiera ser revertido y no comprometieran la integridad de los candidatos. En la tabla 2, se muestran las abreviaciones con su respectivo reemplazo que fueron aplicadas a los textos.

Con el fin de encontrar un complemento a las tres primeras reglas, se recurrió al etiquetado gramatical o *Part-of-Speech Tagging* (*PoS* en inglés) [Toutanova, 2003]. Su objetivo es agregar una etiqueta a cada término de la frase clave. El etiquetado gramatical, en base a su categoría léxica, brinda información sobre una

palabra y su contexto [Rodriguez, 2013]. En este caso, lo usamos para formar patrones.

Tabla 2 Abreviaciones localizadas y su reemplazo.

Abreviación	Reemplazo
e.g.	e-g
Fig.	Fig-
Eq.	Eq-
Ref.	Ref-
al.	al-

Para ello, las palabras clave encontradas en los archivos de anotaciones fueron etiquetadas mediante *PoS tagger*, después de realizar este paso se obtuvieron un total de 1420 patrones que fueron ordenados de acuerdo a la cantidad de repeticiones que tenían dejando solamente aquellos que tuvieran diez o más para motivos prácticos. Los patrones seleccionados se muestran en la tabla 3, junto con sus frecuencias. La finalidad de esta regla fue validar y discriminar candidatos respetando aquellos que sean lo más consistentes con una estructura de frase clave.

Por lo tanto, el funcionamiento de la segunda aproximación se describe a continuación y se ejemplifica en el algoritmo de la figura 2:

- Obtención de patrones de palabras clave para la validación. Se emplea un etiquetado *PoS Tagger* sobre las frases clave de los archivos de anotaciones del corpus de entrenamiento y se realiza la extracción de patrones considerando aquellos con frecuencia de aparición mayor a diez (ver tabla 3).
- Pre-Procesado de los documentos. Se realiza un nuevo pre-procesado sobre el texto, en esta ocasión se usó la nueva técnica de reemplazo en abreviaciones conflictivas (ver tabla 2).
- Formación de *1-gramas, … ,5-gramas*.
- Discriminación de candidatos. Las tres primeras reglas se describen en el algoritmo de la figura 1.
    a. Eliminar candidatos con palabras vacías al inicio y al final.

b. Eliminar candidatos que no formen parte del texto.

c. Eliminar candidatos que no tengan ambos paréntesis.

Tabla 3 Patrones para el segundo sistema y sus frecuencias.

Patrón	Frecuencia	Patrón	Frecuencia
NN	990	NNP NNP NNP	17
NN NN	492	NN IN DT NN	17
JJ NN	390	JJ NN NN NNS	16
NN NNS	265	RB NN	15
JJ NNS	252	NNP NN NNS	15
NNS	231	NNP	15
JJ NN NN	191	NN VBG	15
JJ NN NNS	135	NN JJ NN	15
NNP NN	129	NNP NNP	14
NN NN NN	89	NN JJ NNS	14
NNP NNS	82	JJ NNP NNS	13
VBG	51	JJ CC JJ NNS	13
NN NN NNS	51	DT NN	13
JJ JJ NN	50	VBN JJ NN	12
JJ	43	VBG NN NNS	12
VBG NN	39	VBG DT JJ NN	12
JJ JJ NNS	31	NN IN NNS	12
VBN NN	30	NN CC NN NNS	12
VBN	24	JJ JJ NN NN	12
VBG NNS	23	CD NNS	12
JJ NN NN NN	23	CD NN	12
NNP NN NN	21	NNP JJ NN	11
NN IN NN	21	NN IN NN NNS	11
VBN NNS	20	VB DT JJ NN	11
VBG NN NN	18	NNP NNP NN	10
JJ NNP	18	NN VBG NN	10
NNS NN	17	NN IN NN NN	10

```
Entrada: Texto_científico
Salida: Conjunto de palabras_clave
Inicio
      Gramas <- []
      Patrones <- CargaPatrones()
      TextoPreprocesado <- NuevoPreprocesamiento(Texto_científico)
      Oraciones <- PartirEnOraciones(TextoPreprocesado)
      Para cada oracion en Oraciones hacer:
            GramasDeLaOracion <- CrearGramas(oracion)
            Gramas <- Gramas + GramasDeLaOracion
      Finpara
      //Parte de la discriminación de candidatos
      NuevosCandidatos <- Regla_A(Gramas)
      NuevosCandidatos <- Regla_B(NuevosCandidatos)
      NuevosCandidatos <- Regla_C(NuevosCandidatos)
      PalabrasClave <- ValidacionPorPatrones(NuevosCandidatos,Patrones)
      Return PalabrasClave
Fin
```

Figura 2 Algoritmo del Segundo sistema.

- Validación de candidatos usando patrones. Los candidatos se recibieron, etiquetaron y posteriormente mapearon con los patrones encontrados en la tabla 3 para validarlos mediante su estructura y definir si es una frase clave candidata.

3. Resultados

Los sistemas fueron evaluados usando las medidas clásicas de *Precisión* (ecuación 1)*, Exhaustividad* (ecuación 2) y *Medida-F_1* (ecuación 3) que dan como resultado el rendimiento general del sistema [Tolosa, 2008]. Ambos se calificaron usando un script proporcionado por los organizadores de la Tarea 10 de SemEval 2017, junto con un *gold* estándar [Augenstein, 2017].

$$Precisión(S) = \frac{Cantidad\ de\ términos\ relevantes\ recuperados}{Cantidad\ de\ términos\ recuperados} \quad (1)$$

$$Exhaustividad(S) = \frac{Cantidad\ de\ términos\ relevantes\ recuperados}{Cantidad\ de\ términos\ relevantes} \quad (2)$$

$$Medida - F_1(S) = \frac{2}{\frac{1}{Precisión(S)} + \frac{1}{Exhaustividad(S)}} \quad (3)$$

Conjunto de Datos

Los datos proporcionados por los organizadores de SemEval 2017 tarea 10, constan de 500 artículos científicos del área de Ciencias de la Computación, Ciencias de Materiales y Física. Estos fueron divididos en tres grupos, 350 como conjunto de entrenamiento, 50 como conjunto de desarrollo y 100 para realizar las pruebas de nuestros sistemas. El total de frases clave del *gold* estándar es de 2051.

Cada uno de los 500 artículos fue extraído de la página de *ScienceDirect* y consta de tres partes, un archivo XML con todo el texto, TXT con un párrafo del texto y un archivo de anotaciones, el archivo contiene las palabras clave con un identificador, su clasificación, la posición dentro del texto y la frase clave. Todo el conjunto de datos se encuentra en el idioma inglés.

Evaluación

El primer sistema (GMBUAP) es un algoritmo reportado en la Tarea 10 de SemEval [Augenstein, 2017], el número de frases claves candidatas se presentan en la tabla 4 y su disminución a nivel de renglones al aplicar cada regla del algoritmo, ver columna 2 de la tabla 4. El número de frases claves candidatas del segundo sistema se presenta en la tercera columna de la tabla 4. Como puede observarse la aplicación de la regla correspondiente a patrones del segundo sistema consiguió un decremento notable de términos candidatos para la evaluación.

Tabla 4 Decremento de frases clave candidatas en el primer y segundo sistema.

Filtro	**Frases clave candidatas**	
	Primer sistema	**Segundo sistema**
Original	71804	77339
Primera regla	27994	27254
Segunda regla	21553	24611
Tercera regla	20871	22398
Patrones	----	11909

En la tabla 5 se muestran los resultados experimentales de los dos sistemas comparándolos con los resultados de otros equipos que participaron en SemEval 2017 Tarea 10. Se observa que el primer sistema GMBUAP obtiene bajos resultados, mientras que el segundo sistema logra superar los resultados del primero, consiguiendo un 0.22 de *Medida-F_1*.

Tabla 5 Evaluación general de las aproximaciones.

Equipo	***Medida-F_1***
SciX	0.42
IHS-RD-BELARUS	0.41
Know-Center	0.39
LIPN	0.38
SZTE-NLP	0.35
Segundo sistema	**0.22**
GMBUAP	**0.08**

La tabla 6 presenta los resultados experimentales de los dos sistemas. Se observa un incremento en los resultados de *Precisión*, *Exhaustividad* y *Medida-F_1* del

Segundo sistema con respecto a GMBUAP. Sin embargo, la gran cantidad de candidatos generados por el procedimiento de *n-gramas* comprometieron la precisión y con ello la calificación general. El siguiente objetivo de nuestra investigación consiste en disminuir la cantidad de frases candidatas con la intención de mejorar los resultados.

Tabla 6 Resultados de precisión, recuerdo y medida-F_1 de ambos sistemas.

	Precisión	***Exhaustividad***	***Medida-F_1***
Segunda aproximación	0.18	0.63	0.22
Primera aproximación	0.04	0.53	0.08

4. Discusión

Se presentan dos sistemas para la extracción de frases clave en textos científicos, el primer sistema utiliza *n-gramas* y realiza discriminación usando reglas empíricas. Posteriormente, presentamos una ligera mejora como segundo sistema que utiliza patrones obtenidos a través de un etiquetado *PoS tagger*.

El segundo sistema logra un 0.22 de *Medida-F_1* superando al primer sistema que consiguió un 0.08 de *Medida-F_1*. Asimismo, la *Exhaustividad* y la *Precisión* del segundo sistema también aumento de manera evidente con respecto al primer sistema, lo cual indica que el segundo sistema recupera más de la mitad de las frases clave en el conjunto de prueba y disminuye la cantidad de términos que no son consideradas frases clave. Partimos de la hipótesis de que las frases clave mantiene una estructura morfológica en el texto, es decir, están formadas normalmente por adjetivos y/o sustantivos, lo cual permitió disminuir la cantidad de frases clave candidatas y validar la hipótesis.

La puntuación en *Medida-F_1* de ambos sistemas es competitivo con respecto a los otros equipos participantes en la tarea de SemEval, según lo reportado por [Augenstein, 2017] en el documento de descripción de la tarea. En ambos sistemas solo se utilizaron los datos proporcionados por los organizadores, sin embargo, los otros equipos utilizaron recursos externos y enfoques supervisados que les permitió mejorar los resultados de sus propuestas.

5. Conclusiones

En esta investigación se presentan dos sistemas para la extracción de frases clave en textos científicos, el primer sistema usa una extracción de *n-gramas* para la extracción de frases clave candidatas y después se discriminan usando reglas empíricas que se obtuvieron por observación sobre el conjunto de datos de entrenamiento de la tarea. Posteriormente, se realizó una mejora al primer sistema y a las reglas empíricas a través del uso de patrones obtenidos después de un etiquetado *Part-of-Speech (PoS)* sobre frases clave identificadas del conjunto de entrenamiento y usando aquellos patrones con una frecuencia de aparición mayor a 10.

Los resultados obtenidos en la evaluación de los sistemas en *Medida-F_1* son de 0.08 para el primer sistema y 0.22 para el segundo sistema. Lo anterior, muestra que el enfoque propuesto basado en la hipótesis de que una frase clave tiene una estructura morfológica obtiene resultados competitivos comparado a otros enfoques que realizan su detección de frases clave usando aprendizaje supervisado y recursos externos para mejorar sus resultados.

Para futuras investigaciones, se pretende disminuir el número de términos extraídos y mejorar la puntuación de nuestros sistemas. Por lo cual, se están estudiando otras formas de extracción que puedan usarse de forma independiente o combinarlas con lo propuesto, entre ellas está la intención de aplicar similitud semántica, recurrir a técnicas más tradicionales como es el pesado de términos, es decir, *TF-IDF* o incluso modelos de aprendizaje supervisado, por ejemplo, Naïve Bayes, Máquinas de Soporte Vectorial y Árboles de decisión.

6. Bibliografía y Referencias

[1] Augenstein, I., Riedel, S., Vikraman, L., McCallum, A., & Das, M., SemEval-2017 task 10: Extracting keyphrases and relations from scientific publications. The 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada: Association for Computational Linguistics, 2017.

[2] Matsuo, Y., & Ishizuka, Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. FLAIRS, pp. 392-396, 2003.

[3] Ortiz, R., David, P., Tovar, M., & Jiménez-Salazar, H., BUAP: An Unsupervised Approach to Automatic Keyphrase Extraction from Scientific Articles. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 174-177, 2003.

[4] Ouyang, Y., Li, W., & Zhang, R., 273. Task 5. Keyphrase Extraction Based on Core Word Identification and Word Expansion. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 142–145, 2010.

[5] Park, J., Gun Lee, J., & Daille, B., UNPMC: Naïve Approach to Extract Keyphrases from Scientific Articles. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 178–181, 2010.

[6] Rodriguez , F. J., Nuevas fuentes de información para entrenamiento de etiquetados gramaticales. Buenos Aires: Universidad de Buenos Aires, 2013.

[7] Siqueira, C., ¿Cómo encontrar las palabras clave en un texto?, 22 de Diciembre de 2005. Obtenido de Universia.net: https://goo.gl/q1JgPy

[8] Stuart, R., Dave, E., Nick Cramer, & Wendy Cowley, Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, pp. 1-20, 2010.

[9] Thuy Dung, N., & Minh-Thang, L., WINGNUS: Keyphrase Extraction Utilizing Document Logical Structure. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 166-169, 2010.

[10] Tolosa, G. H., & Bordignon, F. R., Introducción a la Recuperación de Información. Buenos Aires : Tolosa y Bordignon, 2008.

[11] Toutanova, K., Klein, D., Manning, C., & Singer, Y., Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259, 2003.