

Usando la clasificación para una extracción de frases clave en textos científicos

FLORES-PETLACALCO, Gerardo†*, TOVAR-VIDAL, Mireya` y REYES-ORTIZ, José A. ``

†Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación

``Universidad Autónoma Metropolitana, Unidad Azcapotzalco

Recibido 2 de Enero, 2017; Aceptado 8 de Abril, 2017

Resumen

Una frase clave recopila los conceptos importantes de un texto proporcionando al lector una idea general que el autor aborda en su trabajo, de ahí la importancia de elegir aquellas que tengan gran relación con el contenido. Su utilidad varía desde la implementación de técnicas de estudio como resúmenes hasta su uso en el área de Ciencias de la Computación para la clasificación de textos en tópicos. Años atrás, la elección era realizada por expertos del dominio o los autores, sin embargo, la gran cantidad de información creada diariamente hizo esta tarea tediosa y motivaron a la investigación en la creación de sistemas de extracción automáticos. Nuestra investigación tiene como objetivo crear un sistema de extracción de frases clave en textos científicos, el enfoque expuesto inicia con una extracción de candidatos usando n-gramas con una primera discriminación por medio de reglas heurísticas para posteriormente clasificarlas haciendo uso de clasificadores de aprendizaje automático como Naïve Bayes y Máquinas de Soporte Vectorial entrenados con un conjunto de características extraídas de los documentos donde existe la frase clave, por ejemplo: el etiquetado de Partes de la Oración, la posición de la frase clave dentro del texto, su contexto y medidas de pesado de términos como el Coeficiente de Dice. Se realizó una validación cruzada sobre el conjunto de datos de entrenamiento logrando un rendimiento del 78% de precisión usando Naïve Bayes y 77% de precisión con Máquinas de Soporte Vectorial.

Extracción automática, SVM, Naïve Bayes, n-gramas

Abstract

A keyphrase represents the significant concepts of a scientific text provided to reader as a general idea that the author expresses in his work, hence the importance of choosing those that have great relation with the content. The usefulness is varied from the implementation in techniques of study to the use in the area of Computer Science for a text classification based on topics. Years ago, the choice of them was made by domain experts or authors, however, the large amount of information created daily made this task tedious, it motivates research into the creation of automatic extraction systems. Our research aims to create a system of keyphrases extraction from scientific texts. The approach presented begins with a candidate extraction using n-grams as a first discrimination by using heuristic rules, later, they are classified using machine learning methods as Naïve Bayes and Vector Support Machines trained with a set of features extracted from the documents where the keyphrase are contained, for example: The Part of Speech Tagging, the position of the keyphrase within the text, the context in a window size and weighing measures of terms such as the Dice's Coefficient. We carried out a cross-validation on the training data set achieving 78% accuracy, using Naïve Bayes and 77% accuracy using Vector Support Machines.

Automatic extraction, SVM, Naïve Bayes, n-grams

Citación: ORES-PETLACALCO, Gerardo, TOVAR-VIDAL, Mireya y REYES-ORTIZ, José A. Usando la clasificación para una extracción de frases clave en textos científicos. Revista de Ciencias de la Educación 2017. 1-1:1-10

† Investigador contribuyendo como primer autor.

*Correspondencia al Autor Correo Electrónico: gerardo.florespe@alumno.buap.mx

Introducción

A diario, una gran cantidad de información se crea con la intención de ponerse a disposición al público para su consulta. Esta información debe pasar por un proceso de clasificación en tópicos de acuerdo con su contenido con la intención de hacer más ágil el proceso de consulta para los usuarios.

El trabajo de clasificación se hace a través de sistemas de indexación que usan elementos del texto conocidos como frases clave. Una frase clave es una sentencia que captura la información sobresaliente del texto y dan una idea general del trabajo de investigación.

De ahí, que la elección de la frase clave sea un proceso importante puesto que deben ser sentencias que tengan gran relevancia dentro del trabajo para evitar aquellas que sean muy generales, ya que esto puede ocasionar una pérdida de conocimiento por una clasificación errónea. Los autores del trabajo o expertos en el tema son los encargados de realizar la elección, lamentablemente la tarea es tardada y la información creada ha provocado que la selección sea imposible para un pequeño grupo de personas. Por lo anterior, existe la necesidad de crear sistemas de extracción automática de frases clave donde áreas de investigación dentro del Procesamiento de Lenguaje Natural (PLN) han puesto un notable interés en los últimos años.

SemEval 2017 es un congreso anual de análisis semántico que propone una serie de tareas con relación al PLN. En la edición de este año se propuso la tarea 10: “Extracción de frases clave y relaciones semánticas en textos científicos” (Augenstein, Das, Riedel, Vikraman, & McCallum, 2017), la cual incluye tres subtareas:

- Identificación de frases clave.
- Clasificación de frases clave identificadas.
- Extracción de relaciones semánticas entre dos frases clave identificadas.

En este trabajo de investigación, se propone un sistema que da solución específicamente a la subtarea 1. La subtarea tiene como objetivo la creación de un sistema de identificación y posterior extracción de frases clave en textos científicos del idioma inglés.

El sistema propuesto combina la extracción de candidatos usando n-gramas para realizar una discriminación usando heurísticas y finalmente se determina su validez como frases clave usando un clasificador entrenado con un conjunto de características extraídas de un conjunto de entrenamiento. El modelo de clasificación propuesto se evaluó sobre el conjunto de datos de entrenamiento haciendo una validación cruzada en 10 pasos obteniendo resultados satisfactorios con dos clasificadores, Naïve Bayes y Máquinas de Soporte Vectorial (Support Vector Machines, SVM).

El contenido del trabajo se organiza de la siguiente manera, en la siguiente sección se exponen trabajos relacionados con la extracción de frases clave. Posteriormente se da una introducción a la teoría de clasificadores y conceptos de validación cruzada, explicamos el enfoque propuesto y su funcionamiento de manera general. En la siguiente sección se muestran los resultados y el conjunto de datos que se usarán para calcular el rendimiento del sistema. Finalmente, las conclusiones y el trabajo a futuro son presentados.

Trabajos relacionados

El trabajo de extracción automática de frases clave fue abordado por diversos investigadores en el área del Procesamiento de Lenguaje Natural. A continuación, se listan algunos trabajos relacionados y sus contribuciones.

(H. Witten, W. Paynter, Frank, Gutwin, & G. Nevill-Manning, 1999) presentan KEA un trabajo de extracción de frases clave que usa el clasificador Naïve Bayes, los candidatos son obtenidos mediante un preprocesado del texto de entrada y el clasificador es entrenado usando TF-IDF (Term Frequency – Inverse Document Frequency) como atributo de discriminación entre las frases clave y aquellas que no lo son.

(Barker & Cornacchia, 2000) describen un sistema para elegir frases clave en textos usando sentencias conocidas como noun-phrases. Las noun-phrases son oraciones que contiene un sujeto y modificadores, para extraer estas oraciones usan una estructura definida con anterioridad y un diccionario on-line. Su sistema elige los candidatos basándose en la longitud, la frecuencia de la oración y la frecuencia con que se repite el sujeto de la oración. Su enfoque presenta resultados satisfactorios al hacer una comparación con frases clave elegidas por expertos del tema.

(Turney, 2003) presenta una mejora al algoritmo KEA con la intención de aumentar la coherencia del contenido de las frases clave extraídas y el cuerpo del texto. Para ello, usa medidas estadísticas que miden el grado de asociación entre las frases clave candidatas por medio de su relación semántica. La medida estadística que calcula la asociación entre candidatos es provista usando minería web.

(Liu, Pennell, Liu, & Liu, 2009) en su investigación exploran la extracción de palabras clave usando medidas de pesado de términos TD-IDF e incorporan un etiquetado PoS (Part-of-Speech), la clasificación de palabras y una medida que calcula la importancia de una palabra con base a las conexiones que tenga con otras palabras u oraciones según un grafo previamente hecho. El sistema compara las palabras extraídas con anotaciones hechas por expertos del tema para medir su rendimiento. Los resultados muestran que este enfoque funciona bien si es combinado con el etiquetado PoS y la medida de pesado de términos por medio del grafo.

(Ouyang, Li, & Zhang, 2010) proponen PolyU, un sistema para la obtención de frases usando la identificación y posterior expansión de palabras núcleo. Las palabras núcleo son aquellas que presentan alta relevancia para el texto, estas son elegidas por medio de la frecuencia dentro del cuerpo siguiendo la premisa de que palabras muy repetidas serán consideradas como importantes. Después de la identificación, las palabras son usadas como semilla para crear oraciones eligiendo las palabras a la izquierda o derecha de la palabra núcleo. Cada candidato se válida midiendo su frecuencia en el texto y un etiquetado PoS.

(Park, Lee, & Daille, 2010) emplean un sistema para la extracción de frases clave que combina reglas empíricas y características de las frases clave obtenidas por observación sobre el conjunto de datos para crear un modelo estadístico basado en Naïve Bayes. Entre las características consideradas se encuentra su posición y ocurrencia dentro de diferentes partes del texto y la longitud de las frases clave.

Considerando estas características, los candidatos son extraídos del texto, se descartan aquellos que contengan stopwords y se les busca en las secciones del texto para asignarles un puntaje, posteriormente se ordenan y se seleccionan los más altos.

(Segura-Bedmar, Colón-Ruiz, & Martínez, 2017) usan una herramienta conocida como BANNER tool, que es un sistema de reconocimiento de entidades nombradas basado en Conditional Random Fields (CRF) con la finalidad de extraer frases clave de textos científicos. Abordaron la tarea como un problema de etiquetado para encontrar los candidatos a frases clave, primero partieron cada texto en palabras, posteriormente formaron n-gramas y finalmente lo representaron como un conjunto de características que determinan si es una frase clave o no. Su sistema obtuvo una alta precisión en la recuperación de frases clave, pero una baja exhaustividad.

(Animesh & Min-Yen, 2017) usan Conditional Random Field para la extracción de frases clave en textos científicos. El modelo se entrena con características de frases clave confirmadas de un conjunto de entrenamiento, estas características toman en cuenta su etiquetado PoS, la capitalización de las letras que la conforman, los símbolos que están alrededor de la misma, si existe operadores matemáticos y la ocurrencia en el título. El modelo obtiene buenos resultados en la detección de frases clave en los textos y puede usarse para clasificación o incluso para la extracción de relaciones semánticas.

El enfoque propuesto en esta investigación combina la creación de n-gramas con una discriminación usando heurísticas para conseguir un conjunto de candidatos.

Posteriormente, entrenamos dos modelos de clasificación con características de frases clave identificadas de un conjunto de entrenamiento para evaluar los candidatos extraídos y determinar si pertenecen o no al conjunto de frases clave. La capacidad de predicción de los modelos creados fue evaluada con validación cruzada sobre los datos de entrenamiento de la tarea alcanzando un 78% de éxito en las predicciones esperadas, ver sección de resultados.

Clasificadores automáticos

Los clasificadores tienen la función de agrupar objetos de acuerdo con las características que tienen en común. Estas características definen los límites entre las clases que pueden existir y funcionan como medida de similitud cuando un nuevo objeto sin clase llega a la clasificación.

Según (Garrido Satué, 2013) los clasificadores necesitan ser entrenados con el objetivo de definir las fronteras entre las clases, para el entrenamiento usan características de instancias definidas. De manera general, un proceso de clasificación se realiza en cuatro pasos:

1. Reunir muestras de objetos con las clases definidas. Se usan para obtener las características propias de las clases, generalmente estas son condensadas en vectores.
2. Entrenar el clasificador. Los vectores con las características se usan para calcular las fronteras entre las clases, dependiendo de las muestras de objetos se definirán las clases.
3. Extraer las mismas características que se usaron para entrenar el clasificador de instancias sin clase definida.

4. Medir la distancia entre el objeto sin clasificar y las clases obtenidas, para determinar si este se encuentra dentro de la frontera de una clase y determinar a donde pertenece.

De acuerdo con (Ramirez García, Carrillo Ruiz, & Sánchez López, 2015) los tres clasificadores más usados son los siguientes:

- Naïve Bayes. Clasificador probabilístico que usa el teorema de Bayes para estimar la probabilidad de que un objeto x pertenezca a la clase y . Por esto, se considera un clasificador generativo.
- Máquinas de Soporte Vectorial (SVM). Clasificador binario que separa los valores de una variable de acuerdo con el espacio característico de ellas. Trata de encontrar el hiperplano óptimo que usa como frontera de separación entre las instancias positivas y negativas de un conjunto de datos de entrenamiento que contiene N observaciones.
- Árboles de decisión. Los árboles describen un conjunto de reglas que definen las clases. Se compone de hojas y nodos, cada hoja define una clase y cada nodo las condiciones de la clase que le corresponde el valor único de la hoja. Se construye de forma jerárquica y recursiva usando los datos de entrenamiento.
- Validación holdout. El conjunto de datos se divide aleatoriamente en dos conjuntos, uno para entrenamiento y otro para evaluar la calidad. No se considera una validación cruzada como tal puesto que los datos no se cruzan, solo se dividen.
- Validación cruzada en k -pasos. En este caso el conjunto de datos se divide en k subconjuntos. El proceso de pasos de validación se repite un número determinado de veces, donde un conjunto de los k se usa para pruebas y los $k-1$ restantes para entrenamiento.
- Validación cruzada leave-one-out. Es una validación de k -pasos donde k es el número de elementos en el conjunto de datos. Es el caso más extremo pues en cada iteración el modelo se entrena usando todo el conjunto de datos menos uno para predecir la clase del elemento restante. Es muy tardado y tiene un alto grado de coste en cálculos.

En este caso, el enfoque propuesto, usa la Validación Cruzada en $k-1$ -pasos, con $k = 10$.

Enfoque propuesto

El sistema propuesto consta de dos partes, la primera es la extracción de candidatos a frases clave que se realiza por n -gramas, la segunda parte es la creación del modelo de clasificación entrenado con características de las frases clave detectadas en un conjunto de datos de entrenamiento. A continuación, se describe cada parte.

Validación cruzada

(Fernández, 2007) define la validación cruzada como una técnica para evaluar modelos de predicción. Existen tres tipos de validación, los cuales se explican a continuación:

Extracción de candidatos

La extracción de frases clave candidatas para entrenar el modelo de clasificación se realizó mediante el siguiente proceso, tal como se muestra en (Flores Petlascalco, Tovar Vidal, Reyes-Ortiz, & Castillo Zacatelco, 2017) usando la propuesta del segundo sistema y consta de los siguientes pasos:

1. Preprocesado. Se reemplazan abreviaciones con conflicto en el texto.

2. Formación de n-gramas. El texto se parte en sentencias y cada sentencia en términos, con los términos se forman grupos de palabras que van desde una palabra (1-grama) hasta cinco palabras (5-gramas).

3. Discriminación de candidatos.

Cada n-grama creado en la sección anterior se discrimina usando alguna de las tres reglas heurísticas siguientes.

a. Eliminar candidatos con stopwords al inicio o al final del n-grama.

b. Eliminar candidatos que no formen parte del texto.

c. Eliminar candidatos que no tengan ambos paréntesis.

4. Validación de candidatos por patrones. Los candidatos restantes del paso anterior se válida por medio de patrones morfológicos obtenidos por un etiquetado PoS-Tagger, para validar la estructura de los candidatos y determinar si pueden o no ser frases clave.

5. Los candidatos obtenidos se colocan en un conjunto y después se extraen sus características desde el corpus, y se entrena el clasificador.

Creación del modelo de clasificación

El modelo de clasificación se crea con base a las características de las frases clave confirmadas en los archivos del conjunto de datos de entrenamiento proporcionados por los organizadores de SemEval 2017 Tarea 10 y se enriquecen con las características de los candidatos extraídos en el paso anterior.

Por cada frase clave confirmada del conjunto de datos de entrenamiento y de las extraídas en el paso anterior, se obtiene el siguiente conjunto de características, estas se dividieron en tres conjuntos: características de las frases clave, características del etiquetado PoS Tagger y una medida estadística como el Coeficiente de Dice. La lista de las características se muestra a continuación y se resumen en la Tabla 1.

Características de la frase clave

- Longitud de la frase clave. Cada frase clave se divide en sus términos y se calcula la longitud y es un valor que se coloca como característica entera.
- Frecuencia. Se calcula la frecuencia de la frase clave en todo el conjunto y el resultado es una característica entera.

Características del etiquetado PoS Tagger. Cada frase clave se etiquetó usando el etiquetador PoS Tagger del paquete NLTK de Python. Estos valores son características booleanas, verdadero o falso.

- La frase clave contiene un sujeto. Si después del etiquetado PoS en la frase clave existe un sujeto.
- La frase clave contiene adjetivos. Si después del etiquetado existe evidencia de un adjetivo dentro de la frase clave.
- La frase clave contiene verbos. Si hay evidencia de un verbo después del etiquetado.
- Medida estadística Dice. Esta medida es conocida como Coeficiente de Dice Generalizado (Generalized Dice Coeficient, GDC) definida en (Lopez & Romary, 2010) y se muestra en la ecuación (1)

$$GDC(T) = \frac{|T| \log_{10}(freq(T)) Freq(T)}{\sum_{w_i \in T} freq(w_i)} \quad (1)$$

Donde, dada una frase T , $|T|$ indica el número de palabras en T , $freq(T)$ es la frecuencia de ocurrencia de T y $freq(w_i)$ es la frecuencia de cada palabra dentro de T

El conjunto de características obtenido fue puesto en un archivo de tipo *arff* para entrenar el modelo de clasificación usando el programa *WEKA* (Frank, Hall, & Witten, 2016), los clasificadores usados en el entrenamiento fueron SVM¹ y Naïve-Bayes².

| Características | |
|---|--|
| Características de las frases clave | Longitud de la frase clave |
| | Frecuencia de la frase clave |
| Características del etiquetado PoS Tagger | Contiene sujeto |
| | Contiene adjetivo |
| | Contiene verbos |
| Medida estadística Dice | Coeficiente de Dice Generalizado (GDC) |

Tabla 1 Características de las frases clave

Resultados

El conjunto de entrenamiento creado en el paso anterior se creó usando la herramienta *WEKA* (Frank, Hall, & Witten, 2016) que se usa para la creación de modelos de clasificación. La evaluación de la efectividad del modelo se realiza por medio de validación cruzada con k -pasos, donde $k = 10$. Las métricas para medir el rendimiento fueron Precisión (2), Exhaustividad (3) y Medida-F1 (4) mostradas en (Tolosa & Bordignon, 2008).

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

$$E(S) = \frac{\text{Cantidad de términos relevantes recuperados}}{\text{Cantidad de términos relevantes}} \quad (3)$$

$$\text{Medida} - F_1(S) = \frac{2}{\frac{1}{P(S)} + \frac{1}{E(S)}} \quad (4)$$

El conjunto de datos usado fue creado por los organizadores de SemEval 2017 Tarea 10 y consta de 500 artículos del acervo libre de la página web ScienceDirect en temas de Ciencias de la Computación, Ciencias de Materiales y Física. El corpus se dividió en tres grupos 350 para entrenamiento de los sistemas, 50 para desarrollo y 100 para pruebas.

Cada publicación del corpus contiene un archivo con un párrafo del artículo original en texto plano, un archivo XML con todos los datos del artículo original directo de la página web. Los artículos del conjunto de entrenamiento y desarrollo cuentan con un archivo extra de anotaciones con frases clave identificadas (subtarea 1), la clasificación (subtarea 2) y las relaciones semánticas entre frases clave (subtarea 3).

El número de instancias creadas por el sistema para entrenar el modelo y hacer la validación cruzada tiene un total de 10225 muestras para el entrenamiento. Fueron creados dos modelos, uno basado en SVM y el otro en Naïve-Bayes que fueron evaluados haciendo una validación cruzada sobre los mismos datos de entrenamiento. En la Tabla 2, se muestra los elementos correctamente clasificados y los elementos incorrectamente clasificados de los dos modelos creados.

| | SVM | Naïve-Bayes |
|---|------|-------------|
| Instancias correctamente clasificadas | 7846 | 7988 |
| Instancias incorrectamente clasificadas | 2379 | 2237 |

Tabla 2 Instancias clasificadas de acuerdo al modelo

De lo anterior, se calcula el rendimiento del sistema y el clasificador únicamente sobre los datos del conjunto de entrenamiento. Las medidas utilizadas son: Precisión, Exhaustividad y Medida-F1 y los resultados son mostrados en la Tabla 3.

| Modelo | Precisión | Exhaustividad | Medida-F ₁ |
|-------------|-----------|---------------|-----------------------|
| Naïve-Bayes | 0.783 | 0.781 | 0.782 |
| SVM | 0.767 | 0.767 | 0.766 |

Tabla 3 Medidas de rendimiento de los modelos

Donde se muestra que el modelo creado por Naïve Bayes muestra un rendimiento ligeramente superior al SVM en la tarea de detección de frases clave.

Los resultados del sistema propuesto en esta investigación son muy prometedores para la detección de frases clave en textos científicos, cabe resaltar que fue desarrollado sobre el conjunto de datos de entrenamiento proporcionado por los organizadores de SemEval 2017 Tarea 10.

Agradecimientos

Esta investigación es apoyada por el Fondo Sectorial de Investigación para la Educación bajo el proyecto CONACYT CB-2015/257357, beca CONACYT 25547, por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854 y por el proyecto ID 00478 VIEP-BUAP.

Conclusiones

En esta investigación se muestra un sistema de extracción de frases clave que usa clasificadores para determinar si un candidato extraído previamente por medio de n-gramas pertenece a un conjunto de frases clave.

El sistema propuesto fue probado sobre los datos de un conjunto de entrenamiento proporcionado por los organizadores de la tarea. Las frases clave candidatas se obtienen por medio de un método de extracción que usa n-gramas y reglas heurísticas. El modelo es entrenado con las características asociadas a cada frase clave candidata. Posteriormente, se hizo una validación cruzada con $k = 10$ pasos para determinar su rendimiento, sobre el conjunto de datos construido. Los resultados obtenidos son prometedores, es decir, se logra una Precisión del 78% con el clasificador Naïve-Bayes y un 76% con el clasificador SVM.

Los resultados arrojados por la validación cruzada son altos tomando en cuenta la cantidad de características consideradas por el clasificador y que no se recurre a recursos externos como diccionarios o corpus adicional, por mencionar algunos. Como trabajo a futuro se realizará las pruebas sobre el conjunto de prueba de la tarea y se medirá el comportamiento del enfoque propuesto.

Referencias

- Animesh, P., & Min-Yen, K. (2017). WING-NUS at SemEval-2017 Task 10: Keyphrase Extraction and Classification as Joint Sequence Labeling. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 973 - 977.
- Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 546-555.
- Barker, K., & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. *Advances in Artificial Intelligence*, 40-52.
- Fernández, A. (2007). Memoria presentada por Antonio Fernández Álvarez para obtener el Diploma de Estudios Avanzados. Universidad de Almería.
- Flores Petlascalco, G., Tovar Vidal, M., Reyes-Ortiz, J. A., & Castillo Zacatelco, H. (2017). Sistemas para la extracción de frases clave en textos científicos. *Pistas Educativas No. 123 SENIE 2017*, 1626-1642.
- Frank, E., Hall, M., & Witten, I. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition.
- Garrido Satué, M. (2013). Teoría de clasificadores. En M. Garrido Satué, Reconocimiento de señales de tráfico para un sistema de ayuda a la conducción (págs. 13-24). Sevilla: Departamento de Ingeniería de Sistemas y Automática.
- H. Witten, I., W. Paynter, G., Frank, E., Gutwin, C., & G. Nevill-Manning, C. (1999). KEA: Practical automatic keyphrase extraction. Proceedings of the fourth ACM conference on Digital libraries, 254-255.

Liu, F., Pennell, D., Liu, F., & Liu, Y. (2009). Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 620-628.

Lopez, P., & Romary, L. (2010). HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. Proceedings of the 5th International Workshop on Semantic Evaluation, 248-251.

Ouyang, Y., Li, W., & Zhang, R. (2010). 273. Task 5. keyphrase extraction based on core word identification and word expansion. Proceedings of the 5th international workshop on semantic evaluation, 142-145.

Park, J., Lee, J., & Daille, B. (2010). UNPMC: Naïve Approach to Extract Keyphrases from Scientific Articles. Proceedings of the 5th International Workshop on Semantic Evaluation, 178-181.

Ramirez García, M., Carrillo Ruiz, M., & Sánchez López, A. (2015). Combinación de clasificadores para el análisis de sentimientos. Research in Computing Science 94 (2015), 193–206.

Segura-Bedmar, I., Colón-Ruiz, C., & Martínez, P. (2017). LABDA at SemEval-2017 Task 10: Extracting Keyphrases from Scientific Publications by combining the BANNER tool and the UMLS Semantic Network. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 947–950.

Tolosa, G. H., & Bordignon, F. R. (2008). Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos. Buenos Aires: Universidad Nacional de Luján.

Turney, P. D. (2003). Coherent keyphrase extraction via web mining. arXiv preprint cs/0308033.