UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

EXTRACCIÓN DE FRASES CLAVE UTILIZANDO PATRONES LÉXICOS A PARTIR DE RESÚMENES DE ARTÍCULOS CIENTÍFICOS

TESIS

PARA OBTENER EL TÍTULO DE

INGENIERA EN SOFTWARE

PRESENTA:

ESTHER MARITZA GALLEGOS CAMACHO

ASESORA:

DRA. YULIA NIKOLAEVNA LEDENEVA

TIANGUISTENCO, ESTADO DE MÉXICO JULIO 2016

UAEM | Universidad Autónoma del Estado de México

UAP TIANGUISTENCO
Unidad Académica Profesional Tianguistenco

El comité revisor designado por la Subdirección Académica de la Unidad Académica Profesional Tianguistenco de la Universidad Autónoma del Estado de México, aprobó la tesis: **EXTRACCIÓN DE FRASES CLAVE UTILIZANDO PATRONES LÉXICOS A PARTIR DE RESÚMENES DE ARTÍCULOS CIENTÍFICOS** y autorizó la impresión de la misma del C. **ESTHER MARITZA GALLEGOS CAMACHO** el día **11** de **JULIO** de **2016**.

ATENTAMENTE
PATRIA, CIENCIA Y TRABAJO
"2016, Año del 60 Aniversario de la Universidad Autónoma del Estado de México"

Revisor
M. en C.C. Marcela Camacho Avila

Revisor
M. Gilda González Villaseñor

Asesor
Dra. Yulia Nikolaevna Ledeneva

M. en Ing. Gloria Ortega Santillán
Subdirectora Académica de la UAP Tianguistenco
Vo.Bo.

Subdirección Académica
Tianguistenco

www.uaemex.mx
Paraje el tejocote S/N, San Pedro Tlaltizapan, Tianguistenco Edo. de México
Tel. y fax: (01 722) 481 08 00 | E-mail: uapsantiago@uaemex.mx

Resumen

Hoy en día personas buscan la mejor forma de adquirir conocimiento y aprendizaje, sin embargo, este conocimiento se encuentra dentro de documentos con cantidades enormes de información, es por eso que las personas requieren un acercamiento al contenido, para poder determinar si es o no relevante para ellas.

Las frases clave captan la idea principal de un documento y dan al lector una descripción del mismo. No obstante su asignación manual resulta costosa y se tiene que invertir mucho tiempo. Este inconveniente ha llevado a que investigadores busquen métodos que permitan extraer frases clave de manera automática y que éstas contengan la información principal de un documento. Es ahí donde inicia la tarea de extracción de frases clave, la que consiste en dos etapas: la identificación de frases clave candidatas y la selección de frases clave.

En el presente trabajo de tesis, se extraen frases clave a partir de patrones léxicos para los resúmenes de artículos científicos en el idioma inglés.

Se realizan los experimentos con el corpus de Inspec, el cual está constituido por 2000 resúmenes de artículos científicos en el idioma inglés. Cada resumen cuenta con 2 conjuntos de frases clave asignadas manualmente por un experto.

Se comparan los resultados obtenidos con otros métodos del estado del arte.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Anexos

CAPÍTULO 1

Introducción

En este capítulo, se realiza un preámbulo a la tarea de extracción de frases clave, se identifica el problema y se construye una hipótesis.

Posteriormente, se plantean los objetivos generales, seguidos de los objetivos específicos, se muestra una delimitación al problema y se da a conocer al lector la estructura del contenido de dicho trabajo de tesis.

1.1 Antecedentes

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, es por esta razón, que investigadores buscan y desarrollan la mejor forma de obtener un resumen de contenido para aprovechar el conocimiento comprendido en dicha información.

Por otra parte, las personas se apoyan de buscadores que les permita realizar consultas en la web, pero no siempre sus búsquedas son exitosas, en ocasiones la información no es encontrada, debido a que las frases que identifican un documento no son las correctas. Es por eso que el proceso de búsqueda de información es tardado, si se desea realizar una mejor búsqueda, es necesario contar con las frases clave que identifiquen correctamente el documento y para ello existen métodos que apoyan a la generación de dichas frases. Estos problemas dieron inicio a toda una ciencia denominada recuperación de información (Gelbukh & Sidorov, 2006).

La Recuperación de Información (RI ó IR por sus siglas en inglés) no es un área nueva sino que se ha desarrollado desde finales de la década de 1950. Sin embargo, debido al valor que tiene la información digital adquiere un papel más importante (Tolosa & Bordignon, 2008).

Por otro lado, cualquier disciplina que utiliza documentos digitales, para realizar sus labores, podría verse beneficiado a través del uso de RI.

Dentro de la RI podemos encontrar la tarea de extracción de frases clave que ha mostrado su importancia para la mejora de muchas tareas del Procesamiento de Lenguaje Natural (PLN)(Hasan & Ng, 2014).

Las frases clave pueden servir como un breve resumen de un documento, auxiliar en la búsqueda de información o utilizarse para agrupar documentos. Sin embargo, a pesar de las ventajas y aplicaciones que las frases clave poseen, es necesaria la generación

y estudio de métodos que permitan extraer frases clave para describir documentos ya que no todos los documentos cuentan con éstas.

Por esta razón es importante identificar de forma automática que frases de un texto resultan más adecuadas para caracterizarlo, es decir, que conjunto de palabras y su correcta combinación deben elegirse como frases clave.

Por otra parte, para probar métodos, que nos permitan extraer frases clave se requiere de un conjunto de datos llamado corpus.

Los corpus son fuente de información de todos los fenómenos del lenguaje, y se usan para varios tipos de investigación lingüística, los corpus son recursos léxicos muy valiosos y ayudan a los lexicógrafos a completar su percepción (Gelbukh & Sidorov, 2006).

La autora Hernández (Hernández, 2016) también los define como un conjunto de documentos específicos, los cuales contienen documento original, conjunto de frases clave generadas por autor y en algunos casos frases clave generadas por lector.

Dentro de los corpus más comunes utilizados en la tarea de extracción de frases clave podemos encontrar, Inspec (corpus de 2000 resúmenes de artículos científicos en inglés) (Hulth, 2003), SemEval-2010 (corpus de 284 artículos científicos en inglés) (Kim et al., 2010), NUS (corpus de 211 artículos científicos en inglés) (Nguyen & Kan, 2007), DUC-2001 (corpus de 308 artículos en inglés) (Over, 2001) y (Hasan & Ng, 2010).

1.2 Problema

Hoy en día las frases clave son útiles para diversos propósitos como generación de resumen, el agrupamiento, la indexación, etcétera, ya que capturan la idea principal de un documento, y permiten al lector conocer a grandes rasgos el contenido de éste. Sin embargo, pocos documentos tienen asignadas frases clave de forma manual debido a que es costoso y consume mucho tiempo. Es por esta razón que se busca

generar frases clave de forma automática que se asemejen a las de un humano y cumplan con el objetivo de capturar la idea principal de un documento.

Otra razón consiste en que muchas revistan están disponibles solamente el resumen del documento de lo planteado anteriormente surge el problema de la presente tesis:

¿Cómo obtener frases clave que se asemejen a las generadas por el ser humano, utilizando el resumen de artículos científicos en inglés?

1.3 Objetivos

El objetivo principal de esta tesis es extraer frases clave a partir de patrones léxicos para los resúmenes de artículos científicos.

1.3.1 Objetivos específicos

- Identificar técnicas o métodos que se utilizan para la extracción de frases clave.
- Analizar el método del estado del arte, que se aplicará en la tarea de extracción de frases clave.
- Buscar y seleccionar un corpus en inglés grande y disponible para descargar.
- Investigar, analizar y probar el corpus.
- Comparar si las frases clave extraídas a partir de los resúmenes son mejores que las frases clave dadas por un humano.
- Comparar los resultados obtenidos con otros métodos del estado del arte.
- Analizar los resultados obtenidos.

1.4 Hipótesis

Si se aplica un método de extracción de frases clave utilizando patrones léxicos, se podría saber si estos patrones son suficientes para extraer frases clave de los resúmenes de artículos científicos que se asemejen a las de un lector humano.

1.5 Delimitación del problema

El corpus solo cuenta con los resúmenes de los artículos científicos, no contiene documento completo.

Se selecciona solamente un corpus para el idioma inglés.

1.6 Estructura de la tesis

A continuación, se detalla el contenido de este presente trabajo de tesis.
En el capítulo 1, se da una introducción a la tarea de extracción de frases clave, se identifica el problema y se construye una hipótesis. Posteriormente, se plantean los objetivos generales, seguidos de los objetivos específicos, se muestra una delimitación al problema.

En el capítulo 2, se presentan las bases teóricas del tema abordado en este presente trabajo de tesis, que permitan al lector una mayor compresión, como términos y conceptos básicos.

En el capítulo 3, se realiza una descripción del estado del arte, en el cual se encuentran trabajos que comprenden la tarea de extracción de frases clave, trabajos que realizan esta tarea pero utilizan el corpus, antes mencionado, Inspec y por último un trabajo previo.

En el capítulo 4, se muestra la metodología propuesta y se describen los pasos del método implementado para la solución del problema presentado en este trabajo de tesis.

En el capítulo 5, se da una introducción y análisis al corpus a utilizar, se presentan experimentos con diferentes parámetros realizados a este corpus, se obtienen resultados para frases candidatas 5, 10, 15 y 20. Por último, se presenta una discusión de los resultados obtenidos con resultados del estado del arte.

Capítulo 6, se dan a conocer las conclusiones y trabajo futuro que se obtienen al finalizar este trabajo de tesis.

CAPÍTULO 2

Estado del Conocimiento

En este capítulo, se presentan las bases teóricas del problema que llevan al lector a una compresión mayor del presente trabajo de tesis, mostrando los términos y conceptos básicos.

Primeramente, en la sección 2.1, se habla sobre conceptos de las frases clave, su aportación hacia otras tareas del procesamiento del lenguaje natural. Posteriormente, se mencionan dos enfoques de métodos que se utilizan para la tarea de extracción de frases clave, en la sección 2.2, se abordan métodos supervisados y, en la sección 2.3, se abordan métodos no supervisados.

Por último, en la sección 2.4, se presentan conceptos sobre patrones léxicos y la sección 2.5, se aborda las métricas a utilizar en el presente trabajo de tesis.

2.1 Frases clave

Las frases clave son definidas por varios autores, para Ortiz (Ortiz, 2010) 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.

Frank et al. (Frank et al., 1999) afirma que las frases clave dan una descripción de alto nivel de un documento.

Para Nguyen & Kan (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.

Otros autores como Hasan & Ng (Hasan & Ng, 2010) dicen que las frases clave se refieren a un grupo de frases que representan el documento.

Las frases clave tienen aplicaciones como:

Para la autora Hulth (Hulth, 2003) producen un resumen denso para un documento, conducen a la mejora de la recuperación de información, o ser la entrada de una colección de documentos.

Los autores Nguyen & Kan (Nguyen & Kan, 2007), expresan que se pueden ocupar de forma proactiva en el IR, en la indexación que asista a los usuarios en la búsqueda de documentos. También se pueden utilizar para agrupar los documentos.

Hasan & Ng (Hasan & Ng, 2010) reafirman que las frases clave son medios excelentes para proporcionar un resumen conciso de un documento.

Pero además de proporcionar resúmenes concisos, Frank et al. (Frank et al., 1999) afirma que se pueden utilizar como una medida para la similitud de documentos, con lo que es posible una agrupación. Otra aplicación relacionada es la búsqueda de

temas: al introducir una frase clave en un motor de búsqueda, todos los documentos con esta frase clave en particular adjunta se devuelven al usuario.

Por otro lado, los autores Modal & Maji (Mondal & Maji, 2013) utilizan estas frases clave extraídas para la generación de títulos de un texto.

De acuerdo con los autores se puede concluir que las frases clave muestran la idea principal de un documento, facilitando a los lectores decidir si es o no es relevante para ellos.

El trabajo sobre la generación de frases clave se pueden clasificar en dos grandes enfoques: extracción y asignación.

2.1.1 Extracción de frases clave

La extracción de frases clave es también un paso esencial en diversas tareas de procesamiento del lenguaje natural.

Para Liu et al. (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.

Los autores Nguyen & Kan (Nguyen & Kan, 2007) afirman que la extracción de frases clave es seleccionar frases presentes en el documento original.

Y este enfoque consiste en dos etapas: identificación de candidatas y selección de frases clave (Nguyen & Kan, 2007) y (Kim et al., 2012).

1. 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 (Nguyen & Kan, 2007). También se implementa como un proceso que filtra las frases sin importancia (Kim et al., 2012).

2. Selección de frases clave: para Nguyen & Kan, (Nguyen & Kan, 2007) en esta etapa se determina si una frase candidata es una frase clave o no. Los autores

Kim et al. (Kim et al., 2012) complementan que pueden ser seleccionados en función de puntuaciones de importancia, tales como frecuencias de palabras, frecuencias de frases, entre otras.

2.1.2 Asignación de frases clave

En contraste con la extracción, la asignación de frases clave se utiliza cuando el conjunto de posibles frases clave se limita a un conjunto conocido por lo general deriva de un vocabulario controlado (Nguyen & Kan, 2007).

Para el presente trabajo de tesis nos enfocamos en la extracción de frases clave. Por otra parte, para realizar esta tarea los investigadores han explorado dos técnicas: método supervisado y método no supervisado.

2.2 Método supervisado

El método supervisado se propuso por primera vez en el año 1999 por Turney (Turney, 1999) para la extracción automática de frases clave.

Frank et al. (Frank et al., 1999), Turney (Turney, 2000), Hulth (Hulth, 2003) y Medelyan et al. (Medelyan et al., 2009) lo utilizan como tarea de clasificación binaria donde un modelo es entrenado en los datos anotados para determinar si una frase dada es una frase clave o no.

Una desventaja es que requiere una gran cantidad de datos de entrenamiento y sin embargo muestra sesgo hacia el dominio en el que se han entrenado (Hasan y Ng, 2010).

A continuación se presentan algunos métodos supervisado, utilizados previamente para la tarea de extracción de frases clave.

2.2.1 N-grams

Los n-gramas tradicionales son secuencias de elementos tal como aparecen en un documento. En este caso la letra *n* indica cuántos elementos contiene esta secuencia (Sidorov, 2013).

Para la longitud de la secuencia o de n-grama existen bigramas (2-gramas), trigramas (3-gramas), 4-gramas, 5-gramas, etc. Un unigrama, constituido de un solo elemento, es una palabra.

2.2.2 NP-chunks

Conocidos como *Noun Phrase* (NP) *chunks*, o frases nominales, frases sustantivas. Los sustantivos pueden ser propios para la descripción del contenido de un documento. Autores como Nguyen & Kan (Nguyen & Kan, 2007) y Hulth (Hulth, 2003) concuerdan que al analizar frases asignadas manualmente, la mayoría resultan ser sustantivos o frases nominales.

La autora Hulth (Hulth, 2003) lo implementó como su primer enfoque el cual consistió en extraer todas las frases nominales en los documentos, según lo evaluado por un *NP-chunker*.

2.2.3 Patrones de etiqueta POS

Para Sidorov (Sidorov, 2013) las etiquetas POS son etiquetas de clases gramaticales (en inglés, *POS tags, part of speech tags*), tales como sustantivos, verbos, etc. Posiblemente las etiquetas pueden incluir características gramaticales más detalladas, por ejemplo, una etiqueta como "VIP1S", podría significar "verbo, indicativo, presente, primera persona, singular".

Una etiqueta POS puede formar parte de un n-grama (Sidorov, 2013).

2.3 Método no supervisado

El enfoque no supervisado puede ser una alternativa a la desventaja del supervisado.

Los enfoques no supervisados para la extracción de frases clave propuestos hasta la fecha incluyen varias técnicas, entre ellas el modelado de lenguaje, clasificación basada en grafos y la agrupación. Estos métodos han demostrado que funcionan bien en un dominio particular de texto, como los resúmenes de trabajos cortos y artículos de noticias. Su eficacia y portabilidad a través de diferentes dominios sigue siendo un problema sin explorar (Hasan & Ng, 2010).

A continuación se presentan algunos métodos no supervisado, utilizados previamente para la tarea de extracción de frases clave.

2.3.1 Tf-idf

Tf-idf es conocido como frecuencia de términos. En (Sidorov, 2013) se afirma que entre más frecuente es la palabra, más importante es esta palabra para el documento.

La frecuencia de una palabra se denomina *tf* (*term frequency*), y representa cuantas veces la palabra ocurre en un documento.

En el trabajo de (Sidorov, 2013) se denomina tf_{ij}, es decir, cuantas veces la palabra *i* aparece en el documento *j*. Por esta razón, es un valor que puede ser diferente en cada documento de la colección. Para el método de Kim & Kan (Kim & Kan, 2009) mide *tf* (términos frecuentes) en subcadenas, así como en coincidencias exactas.

Normalmente la frecuencia de una palabra se combina con otra medida, que se llama *idf* (en inglés, *inverse document frequency*) conocida como frecuencia inversa de documento.

Sidorov (Sidorov, 2013) interpreta el *idf* de la siguiente marera si una palabra se encuentra en todos los documentos de nuestra colección, entonces esta palabra es

incapaz de distinguir entre los documentos, y por lo tanto no nos sirve. Y al contrario, si una palabra se encuentra exactamente en un documento en nuestra colección, es una palabra muy útil para cálculos de similitud.

Otros autores como Hasan & Ng (Hasan & Ng, 2010) utilizan *tf-idf* asignando una puntuación a cada término *t* en un documento *d* basado en la frecuencia de *t* en *d* (tf) y cuantos otros documentos que incluyen *t* (idf) y lo definen como:

$$tf - idf\boldsymbol{t} = tf\boldsymbol{t} * \log D/D\boldsymbol{t}$$

Donde D = número de documento

D**t** = número de documento que contiene ***t***

Sin embargo, una desventaja de la función es en que requiere un gran corpus a procesar para que *idf* sea útil (Kim & Kan, 2009).

2.3.2 Basados en grafos

2.3.2.1 TextRank

Los autores Mihalcea & Tarau en su trabajo (Mihalcea & Tarau, 2004) proponen el modelo TextRank para clasificar frases clave basadas en los enlaces de coocurrencia entre las frases.

Este modelo de clasificación basado en grafos para el procesamiento de texto, demuestra cómo puede ser utilizado con éxito en aplicaciones del procesamiento del lenguaje natural, como la extracción de frases clave y sentencias.

En el trabajo de Mihalcea & Tarau dan una descripción de cómo funciona: define que son una forma de decidir la importancia de un vértice dentro de un grafo, basado en la información global de forma recursiva extraído de todo el grafo. La idea básica

implementada por un modelo de clasificación basado en el grafo es el de "voto" o "recomendación". Cuando une un vértice a otro, que es básicamente la emisión del voto para ese otro vértice. Cuanto mayor sea el número de votos que se emitan por un vértice, mayor será la importancia del vértice. Además, la importancia del vértice que emite el voto determina la importancia de la votación en sí, y esta información también se tiene en cuenta en el modelo de clasificación. Por lo tanto, la puntuación asociada con un vértice se determina con base en los votos que son depositados a su favor (Mihalcea & Tarau, 2004).

Los autores Hasan & Ng (Hasan & Ng, 2010) utilizan TextRank para representar un texto por medio de un grafo. Cada vértice corresponde a una palabra.

2.3.2.2 SingleRank

SingleRank (Wan y Xiao, 2008) es esencialmente un enfoque de TextRank con tres diferencias principales. En primer lugar, mientras que cada borde en un grafo de TextRank (Mihalcea & Tarau, 2004) tiene el mismo peso en SingleRank tiene un peso igual al número de veces de ocurrencias de los dos tipos de frases correspondientes. Segundo, mientras que en TextRank sólo la palabra correspondiente a los vértices de alta clasificación se puede utilizar para formar frases clave, en SingleRank, no se filtra ningún vértice de baja puntuación. Más bien, se marca cada frase clave candidata en el texto que se examina, y se suman las puntuaciones de los tipos de frases clave obtenidos a partir del grafo SingleRank, y se da como salida las puntuaciones más altas de las *N* candidatas como frases clave para el texto. Finalmente, SingleRank emplea una ventana tamaño de 10 en lugar de 2 (Wan y Xiao, 2008) y (Hasan & Ng, 2010).

2.3.2.3 ExpandRank

ExpandRank (Wan y Xiao, 2008) es una extensión de TextRank que explota vecinos cercanos para la extracción de frases clave. Para un documento dado *D*, el enfoque

se encuentra por primera vez más cercano a *k* documentos de la colección vecina de documentos utilizando una medida de similitud (por ejemplo, la similitud del coseno).

El grafo de *D* se construye a partir de las estadísticas de coocurrencia de las frases candidatas recogidas del propio documento y sus k vecinos más próximos (Wan y Xiao, 2008) y (Hasan & Ng, 2010).

2.4 Patrones léxicos

Patrones: Es un conjunto de características comunes que son distintivas y originales que al momento de estar unidas muestran propiedades en orden y equilibrio (Hernández, 2016).

Léxico: 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. (Real Academia Española, 2014).

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 (García, 2004), (García, 2006), (Camacho, 2015).

Hernández (Hernández, 2016) 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".

2.5 Métricas

En los trabajos del estado del arte, el desempeño de un sistema se evalúa de acuerdo a tres medidas: Precisión (P), Recuerdo (R) y F-Measure (F). Medidas utilizadas en varios

trabajos como (Ledeneva, 2008), (Ledeneva, 2008a), (Montiel, 2009), (Ledeneva, 2010), (Ledeneva, 2014), (Hernández, 2016), (Padilla, 2016).

2.5.1 Precisión (P)

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

2.5.2 Recuerdo (R)

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

2.5.3 F-Measure (F)

$$F = \frac{2 * P * R}{P\ +\ \ R}$$

CAPÍTULO 3

Estado del arte

En este capítulo, se presentan trabajos del estado del arte relacionados con el presente trabajo de tesis. Cuenta con tres secciones, la sección 3.1, se muestran trabajos relacionados a la tarea de extracción de frases clave. En la sección 3.2, se describen trabajos relacionados a la tarea de frases clave donde todos ocupan un corpus en común.

Por último, en la sección 3.3, se encuentra el trabajo que identifica, aplica y evalúa los patrones léxicos y es fundamental para este trabajo de tesis.

Los investigadores hoy en día buscan la mejor forma de extraer frases clave de un documento, que pueda brindar una descripción total de este, 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 entre otras características.

3.1 Trabajos dedicados a la tarea de extracción de frases clave

En este punto se presentan diversos trabajos orientados al tema de extracción de frases clave.

3.1.1 KEA: Practical Automatic Keyphrase Extraction

En el trabajo de Witten et al. (Witten et al., 1999) se creó un algoritmo para la extracción de frases clave: *KEA Automatic Keyphrase Extraction* y la implementación del algoritmo *KEA* en JAVA por Jones (Jones, 2001). El trabajo se realiza sobre el enfoque supervisado, utiliza la técnica Naive Bayes, el cual a partir de datos de entrenamiento crea un modelo de formación que puede extraer las frases clave de un texto completo. El conjunto de todas las frases seleccionadas en un documento se identifican utilizando procesamiento léxico primitivo. Utiliza técnicas de aprendizaje automático, como *tf-idf*, para generar un clasificador que determina que frases candidatas deben ser asignadas como frases clave. Esta herramienta puede ser utilizada en forma local y se necesita una fase previa de entrenamiento.

3.1.2 Learning Algorithms for Keyphrase Extraction

Turney en el año 2000 (Turney, 2000) presenta los resultados de una comparación entre un modelo de extracción basado en un algoritmo genético y una implementación de árboles de decisión C4.5, donde indica que el algoritmo genético extrae mejores frases

clave que los árboles de decisión. Turney reporta la precisión para cinco y quince palabras clave por documento, pero no informa en sus estudios la métrica de recuerdo. Sin embargo, utilizó otra posibilidad de evaluación con evaluadores humanos. Una desventaja de este enfoque es que el número de tokens en una frase clave está limitado a tres, y por otro lado, el usuario debe indicar el número de palabras clave para extraer de cada documento.

3.1.3 Thesaurus based Automatic Keyphrase Indexing KEA++

Medelyan & Witten (Medelyan & Witten, 2006) proponen un nuevo algoritmo para la indexación de frases clave KEA ++, que es una mejora al algoritmo original de extracción de frases clave KEA ((Witten, et al, 1999), (Frank, et al, 1999)). Mediante el uso de información semántica en términos y frases extraídas de un diccionario específico del dominio.

Se basa en el aprendizaje automático y trabaja en dos etapas principales: 1) la identificación de candidatos, que identifica los términos del tesauro que se relacionan con el contenido, 2) filtrado del documento, que utiliza un modelo de aprendizaje para identificar los términos más significativos en base a ciertas propiedades o "características". KEA ++ utiliza la técnica de Naïve Bayes porque es simple y presenta mejores resultados.

3.1.4 Keyphrase Extraction in Scientific Publications

En el trabajo Nguyen & Kan, del año 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. Realizan la extracción de frases clave para un corpus de 211 artículos científicos, aportado previamente en la tesis de Nguyen (Nguyen, 2007). Los autores observaron que las publicaciones científicas tienen un lenguaje técnico, así como una estructura similar, y

sacaron provecho de estas cualidades. El corpus nombrado como NUS es de dominio específico, cuenta con dos conjuntos de frases clave unas proporcionadas por autor y otras por lector. El método que aplican es supervisado, para la extracción de las frases clave es por medio de máxima entropía y Naives Bayes, utiliza la métrica de precisión y evalúan este método con 10 *fold cross validation*. Igualmente que Turney, 2000, aplica una evaluación humana. Sin olvidar que comparan resultados con el sistema KEA (Frank, et al, 1999); (Witten, et al, 1999).

3.1.5 SemEval-2010 Task 5: Automatic Keypharases Extraction from Scientific

En el año 2010, Kim (Kim et al., 2010) junto a otros organizadores 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-measure de 27.5%. Entre los sistemas comerciales el mejor resultado fue obtenido con Alchemy con un F-Measure de 21.37% (Padilla, 2016).

3.1.6 Extracción de Frases Clave usando Patrones Léxicos en Artículos Científicos

En este trabajo se propone un nuevo método no supervisado independiente del lenguaje y del contexto, utilizando los patrones léxicos propuestos anteriormente en el trabajo de Camacho (Camacho, 2015), posicionándose entre los mejores resultados del taller de SemEval-2010 para la tarea # 5: Extracción de frases clave, el corpus utilizado fue, con el mismo nombre, SemEval-2010, que está compuesto de 248 artículos científicos con frases elegidas por autor, lector y configuración combinada, y

se encuentra dividido en tres conjuntos más pequeños (test, train y trial). Con una clasificación de C (Sistemas distribuidos), H (Búsqueda y recuperación de información), I (Inteligencia artificial y sistemas multi-agentes) y J (Ciencias sociales y del comportamiento económico). El método consta de 7 etapas:

1. Pre-procesamiento.
2. Construcción y preparación de datos.
3. Extracción de patrones léxicos.
4. Identificación de frases clave candidatas.
5. Comparación y evaluación de los patrones léxicos.
6. Selección de frases clave.
7. Evaluación.

Las frases candidatas fueron evaluadas con el performance.pl, creado en lenguaje Perl, propuesto por Kim et al. (Kim et al., 2010). Tomando como métricas Precisión Recuerdo y F-measure, obteniendo como resultados F-measure, para top 5 12.92%, para top 10 20.12% y para top 15 21.71%, rebasando el *baseline* propuesto y colocándose en lugares intermedios.

3.1.7 Evaluación de sistemas de extracción de frases clave

En el trabajo de Padilla (Padilla, 2016) se comparan los sistemas de extracción automática de frases clave sobre un conjunto de artículos científicos utilizados en la tarea 5 del SemEval-2010, con el objetivo de conocer que sistemas pueden encontrar las mejores frases clave que se asemejen a las propuestas por un ser humano. En la experimentación se presentan los resultados de la comparación entre los sistemas instalables y en línea los cuales fueron Alchemy, Extractor, Fivefilters, Genia, Kea,

Skyttle, Tree tagger, Texlexan, Translatedlab y Wordstat. Por último, los resultados de la evaluación se comparan con los de la tarea 5 del SemEval-2010.

3.2 Trabajos que ocupan Inspec

A diferencia de Turney (Turney, 2000), Frank et al. (Frank et al., 1999), y Nguyen (Nguyen, 2007) que experimentan con la extracción de frases clave a partir de textos completos, en este trabajo nos enfocamos a la extracción de frases clave a partir de los resúmenes. Ya que Hulth (Hulth, 2003) afirma que muchos artículos de revistas no están disponibles como textos completos, solamente como resúmenes, como es el caso del Internet.

Se presentan trabajos donde se ocupó el corpus de Inspec.

3.2.1 Improved Automatic Keyword Extraction given more Linguistic Knowledge

En el trabajo de Hulth en el 2003, se realiza por primera vez la extracción de frases clave para el corpus Inspec, utiliza un algoritmo de aprendizaje automático supervisado, agrega conocimientos lingüísticos a la representación (tales como las características sintácticas) 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. Como métricas ocupa F-measure, resultante de la combinación de precisión y recuerdo.

$$F\beta = \frac{(\beta^2 + 1) * Precición * Recuerdo}{\beta^2 * Precisión * Recuerdo}$$

Donde β se le asigna el valor de 1.

En este trabajo se lograron los siguientes resultados, obteniendo 33.9 en F-Score para el método de n-gramas con etiquetas debido a que la información lingüística ayuda al proceso de extracción de frases clave como se muestra en la tabla 3.1 con la opción w. tag (Etiqueta).

Method	Assign. tot.	Assign. mean	Corr. tot.	Corr. mean	Prec.	Recall	F-score
n-gram	21 104	42.21	2 187	4.37	10.4	57.3	17.6
n-gram w. tag	7 815	15.63	1 973	3.95	25.2	51.7	**33.9**
Chunking	8 189	16.38	1 364	2.73	16.7	35.7	22.7
Chunking w. tag	4 788	9.58	1 421	2.84	**29.7**	37.2	33.0
Pattern	15 882	31.76	2 519	5.04	15.9	**66.0**	25.6
Pattern w. tag	7 012	14.02	1 523	3.05	21.7	39.9	28.1

Tabla 3.1 Resultados tomados del trabajo de Hulth (Hulth, 2003).

3.2.2 Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art

Otro trabajo que utiliza el corpus de Inspec es el de Hasan & Ng (Hasan & Ng, 2010). Sin embargo, este implementa un método no supervisado, el estudio consiste en la comparación de 5 diferentes algoritmos de extracción de frases clave: tf-idf, TextRank, SingleRank, ExpandRank y KeyCluster aplicados cada uno a 4 diferentes corpus: Inspec, DUC-2001, NUS Keypharases corpus y ICSI, su métrica de evaluación fue por recuerdo, precisión y F-score. Cabe señalar que cada algoritmo se le ajusta parámetros diferentes.

De este trabajo se pudo concluir y seleccionar que nuestra mejor opción de corpus para aplicar al método es Inspec, debido a que presenta los mejores resultados entre los demás corpus, para todos los algoritmos de extracción de frases clave que fue probado.

3.2.3 TextRank: Bringing Order into Texts

Los autores Mihalcea & Tarau, en el 2004 realizan la tarea de extracción de frases clave también para el corpus Inspec. En este trabajo, 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-measure, y demuestran que los resultados obtenidos se comparan favorablemente con los resultados publicados anteriormente en sus puntos de referencia establecidos Figura 3.2.

Method	Assigned		Correct		Precision	Recall	F-measure
	Total	Mean	Total	Mean			
TextRank							
Undirected, Co-occ.window=2	6,784	13.7	2,116	4.2	**31.2**	43.1	**36.2**
Undirected, Co-occ.window=3	6,715	13.4	1,897	3.8	28.2	38.6	32.6
Undirected, Co-occ.window=5	6,558	13.1	1,851	3.7	28.2	37.7	32.2
Undirected, Co-occ.window=10	6,570	13.1	1,846	3.7	28.1	37.6	32.2
Directed, forward, Co-occ.window=2	6,662	13.3	2,081	4.1	31.2	42.3	35.9
Directed, backward, Co-occ.window=2	6,636	13.3	2,082	4.1	31.2	42.3	35.9
Hulth (2003)							
Ngram with tag	7,815	15.6	1,973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4,788	9.6	1,421	2.8	29.7	37.2	33.0
Pattern with tag	7,012	14.0	1,523	3.1	21.7	39.9	28.1

Figura 3.2 Resultados tomados del trabajo de (Mihalcea & Tarau, 2004).

3.2.4 Approximate Matching for Evaluating Keyphrase Extraction

En este trabajo de Zesch & Gurevych (Zesch & Gurevych 2009) se propone una nueva estrategia de evaluación para la tarea de extracción de frases clave basada en coincidencias aproximadas. Por primera vez, se comparan los resultados de los enfoques de extracción de frases clave no supervisados y supervisados del estado de del arte en tres conjuntos de datos de evaluación, Inspec, Duc dataset, SP dataset. Demuestran que el comportamiento relativo de los enfoques depende en gran medida de la métrica de evaluación, así como en las propiedades de los conjunto de

datos de evaluación. El rendimiento de la mayoría de los algoritmos de extracción de frase clave se evalúa mediante la comparación y la coincidencia exacta de las frases clave extraídas con las del ser humano (frases clave asignadas, Gold standard).

Proponen como medida de evaluación solo para precisión denominándola R-p, y se calcula como:

$$R - p = \frac{Mc}{M}$$

Dónde: *M* es la lista generada por el humano (Gold Standard)

Mc es la lista de las coincidencias correctas en *M*.

Como principales métodos de extracción de frases clave utilizan para el enfoque supervisado el sistema KEA y para el no supervisado TextRank, cabe señalar que la métrica de R-p la muestra de dos formas 1) para una coincidencia aproximada ($R\text{-}p_{ap}$) 2) para una coincidencia exacta ($R\text{-}p_{ex}$).

3.2.5 SemanticRank: Ranking Keywords and Sentences Using Semantic Graph

En este trabajo los autores Tsatsaronis et al. (Tsatsaronis et al., 2010) implementan SemanticRank, un algoritmo no supervisado y de clasificación basado en el grafo de la palabra clave y la sentencia de extracción de texto. El algoritmo construye un grafo semántico utilizando enlaces implícitos, que se basan en relación semántica entre los nodos de texto y, en consecuencia filas nodos utilizando diferentes algoritmos de clasificación.

En evaluaciones comparativas se demuestran que SemanticRank obtiene resultados favorables en el conjunto de datos Inspec utilizado anteriormente por Hulth (Hulth, 2003) y Mihalcea & Tarau (Mihalcea & Tarau, 2004).

Evaluaron SemanticRank (SEM), utilizando diferentes valores de k (5, 10, 15 y 20), donde k es el número de frases clave para ser extraídos de cada resumen. Utilizan como métricas precisión, recuerdo y F-Measure. La precisión para cada resumen es el número de frases clave extraídos correctamente, dividido por el número de frases clave extraídos, y el recuerdo se diferencia sólo en el denominador (número de frases clave sugeridas por los indexadores).

3.2.6 Stop-words in Keyphrase Extraction Problem

El problema fundamental de este trabajo presentado por Popova et al. (Popova et al., 2013) es la extracción de frases clave de resúmenes de publicaciones científicas. Ya que está tarea apoya a las sub-tareas de la recuperación de información como: clasificación y agrupación, la minería de datos, extracción de conocimiento y representación, el resumen de texto, la indexación de datos. Los autores dividen esta tarea en dos partes: 1) la extracción de frases clave candidatos; 2) la clasificación de frases clave candidatos para su posterior selección.

Los autores realizan un estudio para saber si la cantidad de Stopwords aplicadas a las pruebas son un factor para la mejora de resultados, evalúan con F-score, que es una combinación de dos características: Precisión y Recuerdo de las frases clave extraídas automáticamente con respecto a la frases clave extraídas de forma manual, el cálculo de estas medidas se realizó de la siguiente forma.

$$Precisión = \frac{C \cap G}{G}$$

$$Recuerdo = \frac{C \cap G}{C}$$

$$F - Score = \frac{2 * Precición * Recuerdo}{Precisión * Recuerdo}$$

Dónde:

C ∩ G: es el número de "verdaderos positivos" - frases clave que se han extraído adecuadamente en todos los documentos considerados.

C**:** Número total de frases clave extraídas manualmente en todos los documentos.

G: Número total de frases clave extraídas de forma automática en todos los documentos.

3.2.7 Improved Algorithms for Keyword Extraction and Headline Generation from Unstructured Text

En este trabajo presentado por Modal & Maji (Modal & Maji, 2013), el problema no se centra a la tarea de frases clave, pero si se ve como una base principal para la solución que desean dar.

Utilizan algoritmos de extracción de frases clave para la generación de títulos para un texto no estructurado. Algunos enfoques utilizados son Naive Bayesiano, *tf-idf* (término de frecuencia * frecuencia inversa de documento) Ranking, Modelo de texto. En el enfoque Naive bayesiano, se trata de captar la correlación entre las frases en el documento y las frases en el título. Cada enfoque con un pre-procesamiento de etiquetado, normalizado, segmentado y aplicación de stemming.

Sin embargo, para el problema de generación de títulos, el corpus Inspec no arroja buenos resultados, debido a que se necesita más información para cada algoritmo aplicado.

3.3 Otros trabajos

3.3.1 Detección de fragmentos de texto como candidato a hipervínculo

Un trabajo base para esta investigación es el elaborado por Camacho (Camacho, 2015), quien realizó 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 Secuencias Frecuentes Maximales (García, 2007) y 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 SFM y así obtener los patrones léxicos que contuvieran contexto derecho e izquierdo. Para la aplicación de los patrones léxicos, se forma un procesos 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 hipervincularse. 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-measure.

Precisión (P), está definida como la fracción de casos recuperados que son relevantes:

$$P = \frac{|\{hipervínculos\ relevantes\} \cap \{fragmentos\ de\ texto\ recuperados\}|}{|\{fragmentos\ de\ texto\ recuperados\}|}$$

Recuerdo (R), está definido como la fracción de casos recuperados con la consulta del sistema:

$$R = \frac{|\{hipervínculos\ relevantes\} \cap \{fragmentos\ de\ texto\ recuperados\}|}{|\{hipervínculos\ relevantes\}|}$$

F-Measure (F), es la métrica armónica basada en la precisión y el recuerdo definida como:

$$F = \frac{2\ P\ R}{P\ +\ R}$$

CAPÍTULO 4

Metodología

En este capítulo, se describe la metodología a implementar que nos permita llegar a la solución del problema planteado sobre la tarea de extracción de frases clave.

Dicha metodología está compuesta de 4 etapas, donde cada etapa conlleva una serie de pasos que se detalla a lo largo de este capítulo.

En la sección 4.1, se presenta el pre-procesamiento, con sus correspondientes etapas internas, siguiendo se encuentra la aplicación del método en la sección 4.2, y todos sus pasos que la integran. Posteriormente, en la sección 4.3, se realiza una evaluación y para finalizar en la sección 4.4, se obtienen resultados y se llegan a conclusiones.

4.1 Metodología propuesta

Existen diferentes pasos para las tareas de recuperación de información, cada uno enfocado al problema que desea resolver, es por ello que no hay una metodología formal, para la presente investigación se propone la siguiente metodología para la tarea de extracción de frases clave.

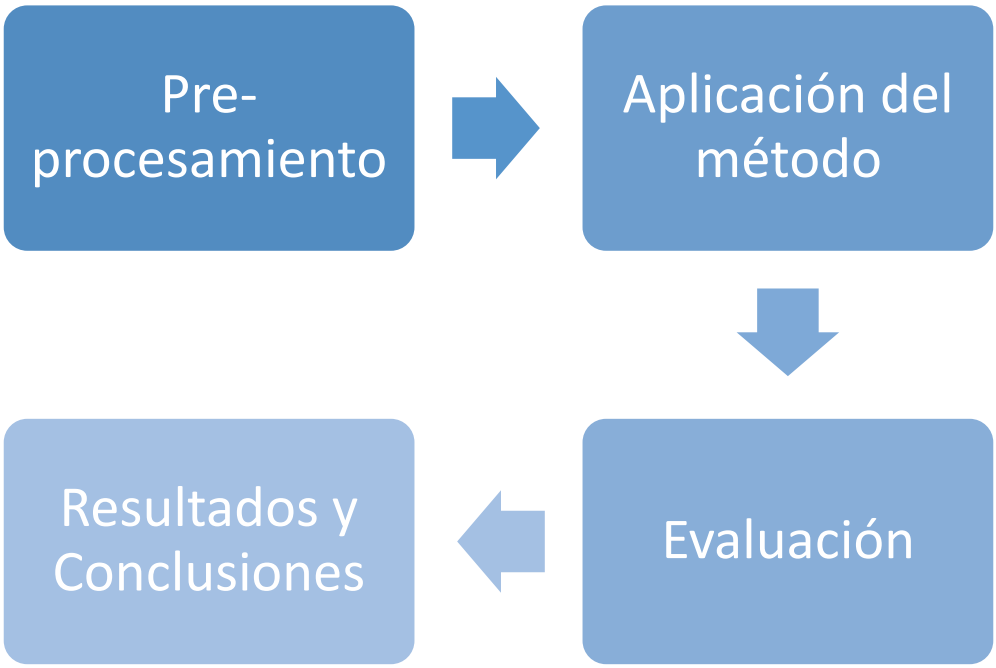

Figura 4.1 Esquema principal de la metodología.

4.2 Etapa pre-procesamiento

Al corpus electo se le aplica una serie de pasos previos a introducirlos al método a evaluar del estado del arte.

Pre-procesamiento de los documentos. Los textos se transforman a algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis, (Montes, 2003). Esta técnica consiste en extraer las frases utilizadas en un documento,

o segmentar el texto en distintas formas gráficas. Una forma gráfica se define como una secuencia de caracteres no delimitadores (en general, letras), comprendida entre dos caracteres delimitadores (espacios o signos de puntuación) (Eíto & Senso, 2004).

4.2.1 Eliminación de caracteres especiales

El primer paso consiste en eliminar aquellos caracteres especiales, sin embargo no se realiza una eliminación como tal dentro del texto, si no se sustituyen por letras o caracteres válidos. 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 (è|é|ê|ë) = e (Ver lista en anexo 2.a.).

De esta forma se obtiene un texto limpio de caracteres especiales para sus posteriores procesos.

4.2.2 Stopwords

También conocido como palabras vacías. Las palabras vacías son palabras que no tienen un significado importante, por lo tanto no son indispensables en varias tareas de PLN (Hernández, 2016).

Consiste en la eliminación de palabras vacías, carentes de significado, como son preposiciones, artículos, conjunciones, etc. Sin embargo, no todos los autores coinciden en los beneficios de eliminar las palabras vacías.

En (Hernández, 2016) y en el presente trabajo, los signos de puntuación y *stopwords* son importantes para la etapa de identificación de frases candidatas, por tanto se codifican y etiquetan (Ver lista en anexo 2.b.) De igual forma se aplica para números, fechas y correos electrónicos, que es texto no importante y no afecta en la identificación de frases candidatas (ver lista de números en anexo 2.c).

En la figura 4.2, se muestra un fragmento del texto original. Posteriormente, en la figura 4.3, se muestra la codificación de signos de puntuación y caracteres (excepto '@'). Finalmente, en la figura 4.4, se muestra el etiquetado de *stopwords*, utilizando una lista de 183 palabras (ver anexo 1.a).

When using speaker-dependent tfpc filters, our results show a relative improvement of approximately 20% compared to the use of the classical cepstral coefficients augmented by their delta -coefficients, which is sianificantly better with a 90% confidence leve

Figura 4.2 Fragmento original de un *abstract*.

When using speaker **@GM** dependent tfpc filters **@COMA** our results show a relative improvement of approximately **@NUM** compared to the use of the classical cepstral coefficients augmented by their delta **@GM** coefficients **@COMA** which is significantly better with a **@NUM** confidence level

Figura 4.3 Codificación de signos de puntuación y otros caracteres (números).

@WHEN @USING speaker @GM dependent tfpc filters @COMA **@OUR** results **@SHOW @A** relative improvement **@OF @APPROXIMATELY** @NUM compared **@TO @THE @USE @OF @THE** classical cepstral coefficients augmented **@BY @THEIR** delta @GM coefficients @COMA **@WHICH @IS @SIGNIFICANTLY** better **@WITH @A** @NUM confidence level

Figura 4.4 Etiquetado de *stopwords*.

4.2.3 Stemming

Finalmente, como parte del pre-procesamiento se suele realizar la normalización de las palabras extraídas del documento conocido como *stemming*. Esta normalización — también llamada lematización — consiste en dividir cada palabra en los lemas que la forman.

Por ejemplo, las palabras alumno, alumna, alumnado, alumnos, etc., comparten una misma raíz léxima (alumn-) que les da el mismo significado semántico (Eíto & Senso, 2004).

Para el presente trabajo de tesis, se aplicaron dos tipos de *stemming*, ambos derivados del algoritmo de Porter (Porter, 1980). Estos algoritmos se encuentran escritos en lenguaje Perl. El primer algoritmo empelado aplica el *stemming* a todas las frases que se encuentran en el documento, incluyendo *stopwords*, y transforma todo el documento en minúsculas, que posteriormente se tiene que transformar a mayúsculas para etapas siguientes. Un ejemplo del *stemming* para *stopwords* queda de la siguiente forma (ver anexo 1.b.).

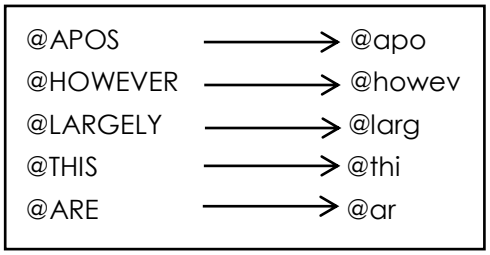

Figura 4.5 Ejemplo del primer algoritmo de Porter aplicado a *Stopwords*.

En la figura 4.6, se muestra el primer algoritmo de Porter utilizado en este presente trabajo de tesis, aplicado a un *abtract* completo del conjunto de test del corpus Inspec.

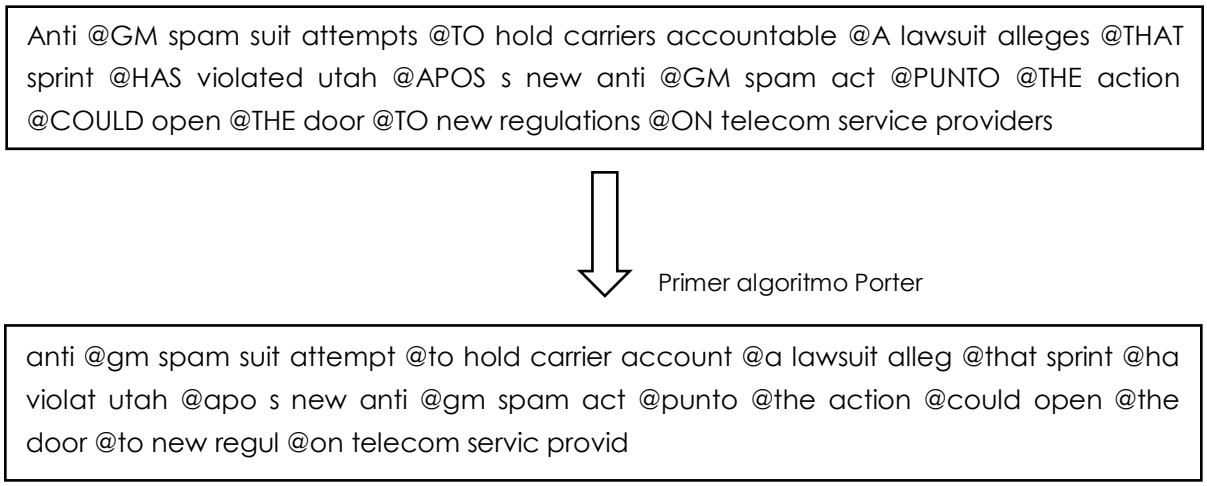

Figura 4.6 Aplicación del primer algoritmo de Porter a un abstract.

El segundo algoritmo implementado tiene una modificación en la cual, las frases que se encuentran en mayúscula no se les aplica *stemming*, pero las que se encuentran en minúscula sí se aplica stemming. Obteniendo *stopwords* escritas correctamente, a comparación del *stemming* anterior.

En la figura 4.7, se muestra la aplicación del segundo algoritmo de Porter sobre el mismo abtract completo, utilizado en la figura 4.6 del primer algoritmo.

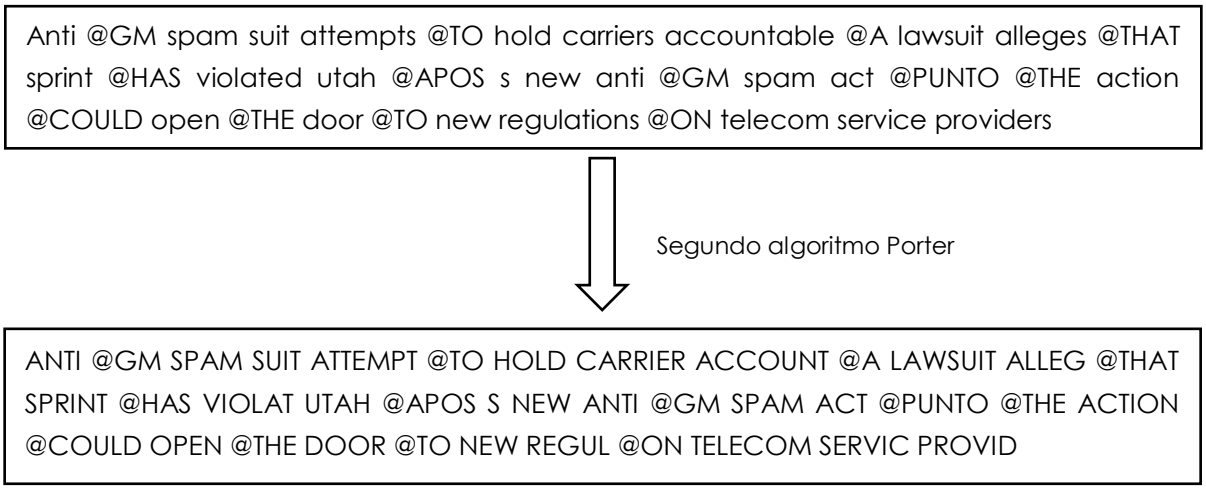

Figura 4.7 Aplicación del segundo algoritmo de Porter a un abstract.

4.3 Etapa aplicación del método

Para este trabajo, se va aplicar el método del estado del arte de Hernández (Hernández, 2016).

El método a ocupar se presenta en el trabajo de "Extracción de Frases Clave Usando Patrones Léxicos En Artículos Científicos" realizado por Hernández en el 2016 (Hernández, 2016), pero se implementará para diferente corpus. El corpus Inspec fue aplicado por primera vez en el año 2003 por Anette Hulth en su trabajo "Improved Automatic Keyword Extraction Given More Linguistic Knowledge" (Hulth, 2003).

4.3.1 Construcción y preparación de datos

El método de Hernández (Hernández, 2016) retoma la generación de tres conjuntos de datos, de la etapa de selección, limpieza y trasformación del método de Camacho (Camacho, 2015). Es por esta razón que para el presente trabajo, se aplica la misma construcción de 3 conjuntos de datos y se aplican 3 pasos del método de Hernández (Hernández, 2016) para el conjunto *test* (prueba) de nuestro corpus seleccionado para esta tarea de extracción de frases clave.

1. Colección 1

La primera colección se obtiene del conjunto de frases clave no controladas, ya incluido en el corpus seleccionado y que fue realizado por indexadores profesionales.

PASO 1 (Hernández, 2016). En el conjunto de frases no controlado, se hace una clasificación de frases clave para cada documento, de acuerdo a Hernández (Hernández, 2016) donde afirma que por cada documento, difiere la cantidad de

frases y longitud, y no todos los documento tienen frases de la misma longitud y cantidad.

La clasificación realizada es con base en la longitud de las frases, tomando en cuenta el tipo de clasificación de longitud que presenta Hernández (Hernández, 2016).

En la figura 4.8, se muestra un ejemplo de la clasificación de longitud, tomada de Hernández (Hernández, 2016), y aplicada a frases del conjunto no controlado del corpus seleccionado para este trabajo. Nótese que para la clasificación **L4** el guión medio se toma como una palabra, ya que en el paso de pre-procesamiento fue codificado.

L1 = lawsuit

L2 = sbc commun

L3= servic oper strategi

L4= anti - spam act

Figura 4.8 Ejemplo de la clasificación de longitud.

PASO 2 (Hernández, 2016). Las frases clave del conjunto no controlado fueron transformadas a frases clave de búsqueda, por medio de expresiones regulares, para más detalles ver (Hernández, 2016). Permitiendo que entre cada una de las palabras que componen las frases, existiera un intervalo de búsqueda de tres stopwords, este intervalo se aplica a partir de las frases de longitud 2.

L1 = LAWSUIT

L2 = SBC (@3stopword) COMMUN

L3= servic (@3stopword) oper (@3stopword) strategi

L4=anti (@3stopword)@GM (@3stopword) spam (@3stopword) act

Figura 4.9 Frases de búsqueda para corpus seleccionado.

PASO 3 (Hernández, 2016). Las frases de búsqueda, obtenidas en el paso anterior, son marcadas en el texto por "{ }", obteniendo un conjunto de documentos por cada longitud de cada palabra.

ANTI @GM SPAM SUIT ATTEMPT @TO HOLD CARRIER ACCOUNT @A LAWSUIT ALLEG @THAT SPRINT @HA VIOLAT UTAH @APO S NEW **{ ANTI @GM SPAM ACT }** @PUNTO @THE ACTION @COULD OPEN @THE DOOR @TO NEW REGUL @ON TELECOM SERVIC PROVID

Figura 4.10 Frases de búsqueda marcadas con llaves, sobre un *abstract* del corpus seleccionado para este trabajo.

2. Colección 2

Es la colección de textos (para este trabajo, los abstract del conjunto de test) resultante de aplicar cada paso del pre-procesamiento, más un paso de recodificación, donde todos los signos de puntuación y caracteres codificados regresan a su estado normal. La lista de re-codificación se encuentra en el Anexo 3.

Por ejemplo:

- @PUNTO regresa a ser " **.** "
- @PUNTOCOMA regresa a ser " **;** "
- @COMA regresa a ser " , "
- @APO o @APOS regresa a ser " ' " (varía dependiendo al stemming aplicado ver paso 4.2.3)

3. Colección 3

En esta colección se aplica el algoritmo DIMASP (García, 2007), y se obtiene a partir de la "Colección 1", donde se clasificaron y marcaron las frases clave de búsqueda en los documentos y se obtuvo un conjunto de documentos separados en colecciones por longitud de frases. Más detalles ver en (Hernández, 2016).

4.3.2 Extracción de patrones léxicos

En esta etapa, se utiliza la técnica de minería de patrones a la "Colección 3", que se llama DIMASP (García, 2007). Es una herramienta que es aplicada a la colección para la generación de secuencias frecuentes maximales donde los patrones léxicos son extraídos (Hernández, 2016). En la Minería de Patrones Secuenciales, una secuencia de palabras se considera frecuente si esta se encuentra en al menos un cierto número de documentos (umbral mínimo de frecuencia) (Camacho, 2015).

Del conjunto de secuencias, es necesario seleccionar las que cumplan con las siguientes características:

- Deben contener una etiqueta que indica que en esa secuencia existe una frase clave "*@LINK*" (etiqueta predefinida por el método de Camacho (Camacho, 2015)).
- Deben contar con contexto antes y después de que aparezca la etiqueta "*@LINK*".

En la figura 4.11 se muestra un ejemplo de la selección de secuencias frecuentes maximales tomada de Hernández (Hernández, 2016).

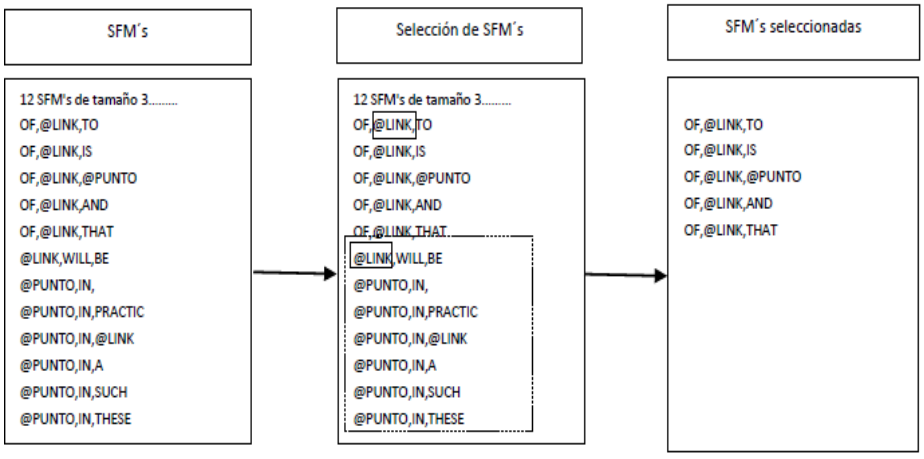

Figura 4.11 Selección de secuencias frecuentes maximales de (Hernández, 2016).

4.3.3 Identificación de frases clave candidatas

El autor Hernández (Hernández, 2016) divide esta etapa en dos procesos que ayudan a identificar frases clave candidatas obtenidas por los conjuntos de patrones.

Etapa 1 (Hernández, 2016). Teniendo la colección de patrones léxicos obtenidos en la sección 4.3.2, se transforma a una colección de patrones de búsqueda por medio de expresiones regulares. En la figura 4.12 se ilustra el proceso.

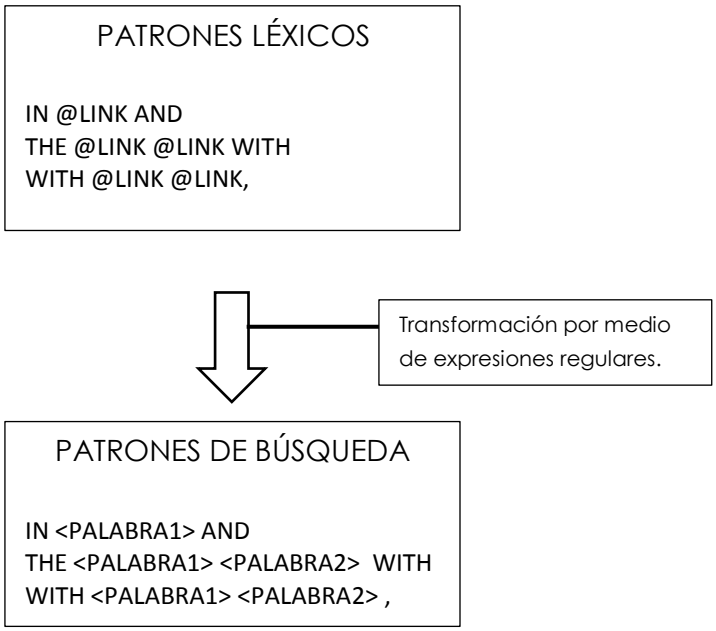

Figura 4.12 Trasformación de patrones léxicos a patrones de búsqueda.

Etapa 2 (Hernández, 2016). Se aplica el conjunto de patrones de búsqueda a la colección 2 generada en la sección 4.3.1 (documentos en texto plano), con el fin de obtener las frases clave candidatas clasificadas por longitud. Siguiendo el método de Hernández (Hernández, 2016), cada palabra que aparezca de manera repetida por un mismo patrón se toma solo una vez.

En la figura 4.13, se muestra el proceso de identificación de frases candidatas, aplicando los patrones de búsqueda a un texto (abstract) del corpus electo para el presente trabajo de tesis.

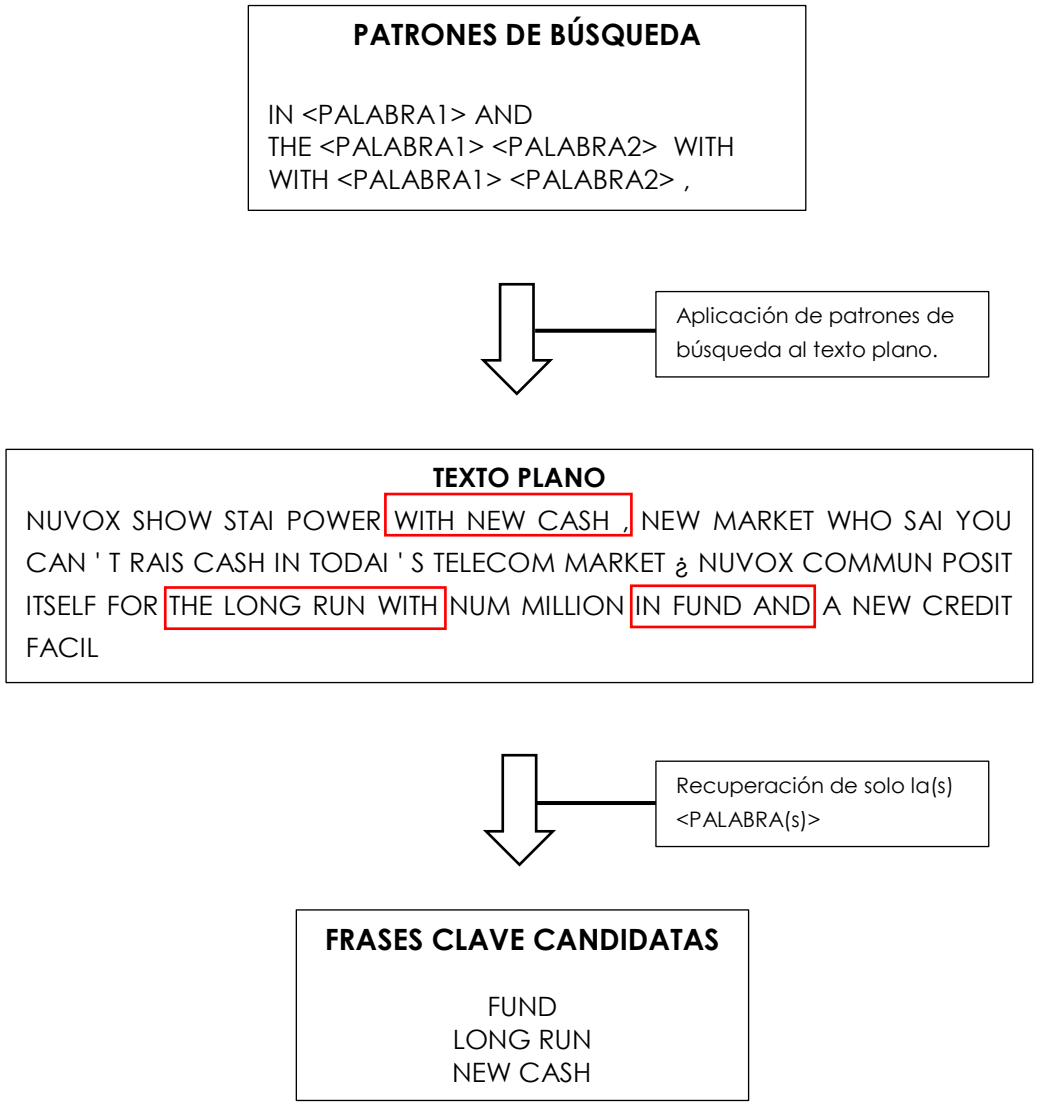

Figura 4.13 Proceso de obtención de frases clave candidatas aplicando patrones de búsqueda.

4.3.4 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 como "Colección 1". Esto 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-Measure, tomando el cálculo de Camacho (Camacho, 2015) y se implementó para frases clave donde:

Precisión

$$P = \frac{|\{hipervínculos\ relevantes\} \cap \{fragmentos\ de\ texto\ recuperados\}|}{|\{fragmentos\ de\ texto\ recuperados\}|}$$

Recuerdo

$$R = \frac{|\{hipervínculos\ relevantes\} \cap \{fragmentos\ de\ texto\ recuperados\}|}{|\{hipervínculos\ relevantes\}|}$$

F-Measure

$$F = \frac{2\ P\ R}{P\ +\ R}$$

4.3.5 Selección de frases clave

En esta etapa se realiza la selección de frases clave a partir del conjunto de frases clave candidatas extraídas por los patrones de búsqueda en la sección 4.3.3. Al conjunto de frases clave candidatas extraídas por cada patrón de búsqueda, se le asignan los valores obtenidos del cálculo de Camacho (Camacho, 2015) en la sección 4.3.4.

De acuerdo al método de Hernández (Hernández, 2016) se realizan diferentes enfoques de pesado; Precisión, Recuerdo, Booleano, F-Measure. Cabe señalar, que el valor Booleano no es más que el valor de "1" si una frase clave ocurre más de una vez en el conjunto de documentos de frases clave candidatas o "0" si esta palabra no es encontrada en los documentos (Hernández, 2016).

Este proceso se realiza por cada conjunto de frases clave candidatas extraída por cada uno de los patrones y por cada clasificación de longitudes (L1, L2, L3). Se toma solo longitudes de L1, L2 y L3 en base al análisis realizado por Hernández (Hernández, 2016).

El pesado consta de 6 pasos:

1. Formatear los archivos de evaluación correspondientes a cada patrón, respecto a la cantidad de frases candidatas extraídas.
2. Formateo de frases extraídas de los patrones de búsqueda.
3. Asignación de la medida del patrón, a las frases extraídas por él.
4. Unir todas las frases extraídas por diferentes patrones de una longitud en particular de un documento.
5. Obtener el peso de las frases.
6. Ordenar de mayor a menor las frases candidatas.

Para una explicación más profunda sobre esta etapa revisar (Hernández, 2016).

4.4 Etapa evaluación

Para evaluar las frases claves generadas por método del estado del arte aplicado al presente trabajo de tesis se implementaron las métricas: Precisión, Recuerdo, F-Measure tomadas del trabajo de Hernández (Hernández, 2016).

Precisión:

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

Recuerdo:

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

F-Measure:

$$F = \frac{2 * P * R}{P\ +\ \ R}$$

En esta etapa de la metodología se implementaron dos sistemas de evaluación:

1. Rouge (Lin, 2004)
2. Performance (Kim et al., 2010)

4.4.1 Rouge

Sistema de evaluación realizado por Lin (Lin, 2004), denominado Recall-Oriented Understudy for Gisting Evaluation (Por sus siglas ROUGE), elaborado para la evaluación de generación de resúmenes, pero implementado para evaluar la tarea de extracción de frases clave por los autores Mihalcea & Tarau (Mihalcea & Tarau, 2004) y los autores Tsatsaronis et al. (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 5, 10, 15 y 20.

4.4.2 Performance

Evaluador propuesto por Kim et al. (Kim et al., 2010) para la tarea 5 de SemEval- 2010, (tarea de extracción de frases clave) el cual está escrito en lenguaje Perl y es ejecutado en sistema operativo Windows.

Permite la evaluación de frases candidatas 5, 10 y 15.

4.5 Etapa resultados y conclusiones

En este paso se presentan todos los experimentos realizados con sus diferentes parámetros. Se realiza un análisis y discusión de los resultados para poder obtener las conclusiones del presente trabajo de tesis.

CAPÍTULO 5

Experimentos y resultados

En este capítulo, se encuentran los experimentos realizados de acuerdo a la metodología propuesta en el capítulo anterior. Como primera parte, en la sección 5.1, se puede encontrar una descripción del corpus implementado así como un análisis que se realizó sobre el mismo.

Posteriormente, en la sección 5.2, se muestra una prueba realizada sobre un sistema, del estado del arte, que se ocupan de la tarea de extracción de frases clave.

Para finalizar, se tiene la sección 5.3, donde se puede observar pruebas realizadas con el método propuesto del estado del arte para realizar esta tarea de extracción de frases clave.

5.1 Corpus

El corpus a emplear para la presente tesis es ***Inspec***, utilizado por primera vez en artículo, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge", de Anette Hulth (Hulth, 2003), el cual está compuesto por 2 000 resúmenes en Inglés, con su correspondiente título y frases clave de la base de datos Inspec.

Los resúmenes fueron extraidos de artículos de revistas de los años 1998 a 2002, comprendiendo las disciplinas *Computers and Control,* y *Information Technology*.

Cada resumen tiene dos conjuntos de frases clave asignadas por un indexador profesional asociada a ellos: un conjunto de términos controlados, es decir, términos restringidos al tesauro de Inspec (ver anexo 5.c); y un conjunto de términos no controlados que puede ser cualquiera de los términos adecuados, asignados libremente por los indexadores (ver anexo 5.b).

El conjunto de resúmenes se dividió arbitrariamente en tres grupos: un conjunto de entrenamiento "*Training*" (para construir el modelo) que consta de 1 000 documentos (ver anexo 4.), 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 (ver anexo 5). Tanto los términos controlados y los no controladas pueden o no estar presentes en los resúmenes.

5.1.1 Análisis

Para la presente tesis se ocupó solo el conjunto de términos no controlados, con base en el estudio de Hulth (Hulth, 2003) en el cual alude que, los indexadores tuvieron acceso a los documentos completos cuando asignaron las frases clave y concluyeron que los conjuntos no controlados están presentes en los resúmenes en mayor medida, con un porcentaje de 76,2% a las controladas que solo tiene un 18.1%.

En las figura 5.1 y 5.2, se muestra el porcentaje de frases clave presentes en los resúmenes (Abstract) del corpus, para el conjunto no controlado y controlado respectivamente. Cabe señalar que estas gráficas se generaron por medio de los datos proporcionados por Hulth (Hulth, 2003), donde demuestran que el conjunto de frases no controladas está presente en mayor medida en los resúmenes que el conjunto de frases controlado.

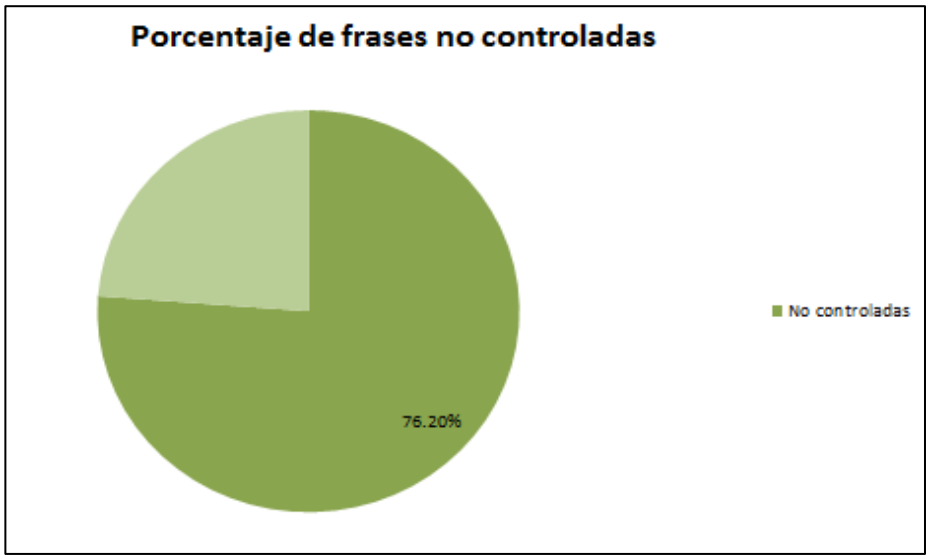

Figura 5.1 Gráfica de porcentaje de las frases clave no controladas presentes en los resúmenes.

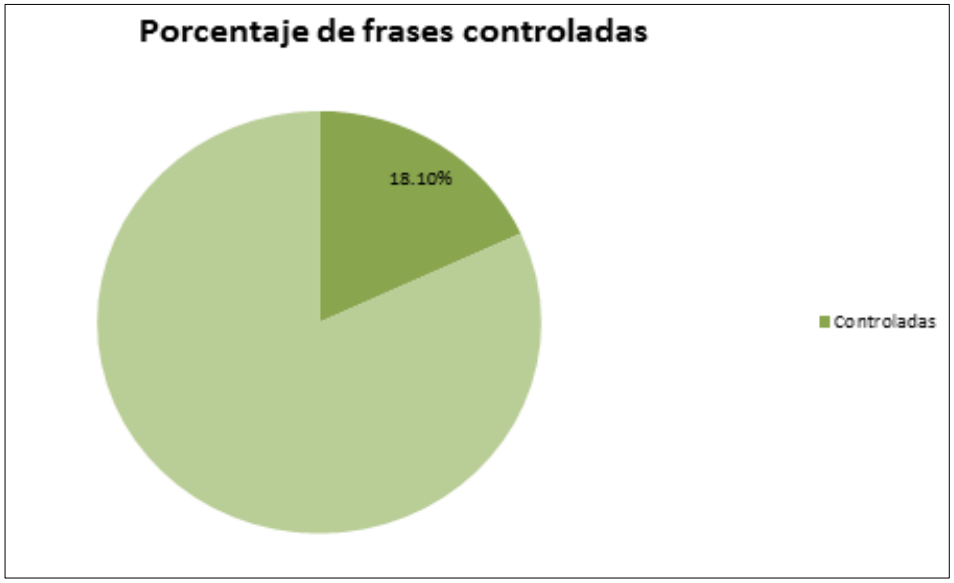

Figura 5.2 Gráfica de porcentaje de las frases clave controladas presentes en los resúmenes.

Para poder determinar los valores de umbral para el número de frases clave seleccionadas, se realizó un análisis donde se obtuvo el número de frases clave que contiene cada resumen. En su lista de no controladas, obteniendo un numero diferente de frases clave para cada documento; dentro del rango de 1 a 30.

En la figura 5.3, se muestra la cantidad de documentos que existen con un número dado de frases clave, esto para el conjunto *Traing*, donde se aprecia que existen más documentos con 6, 7 y 8 frases clave.

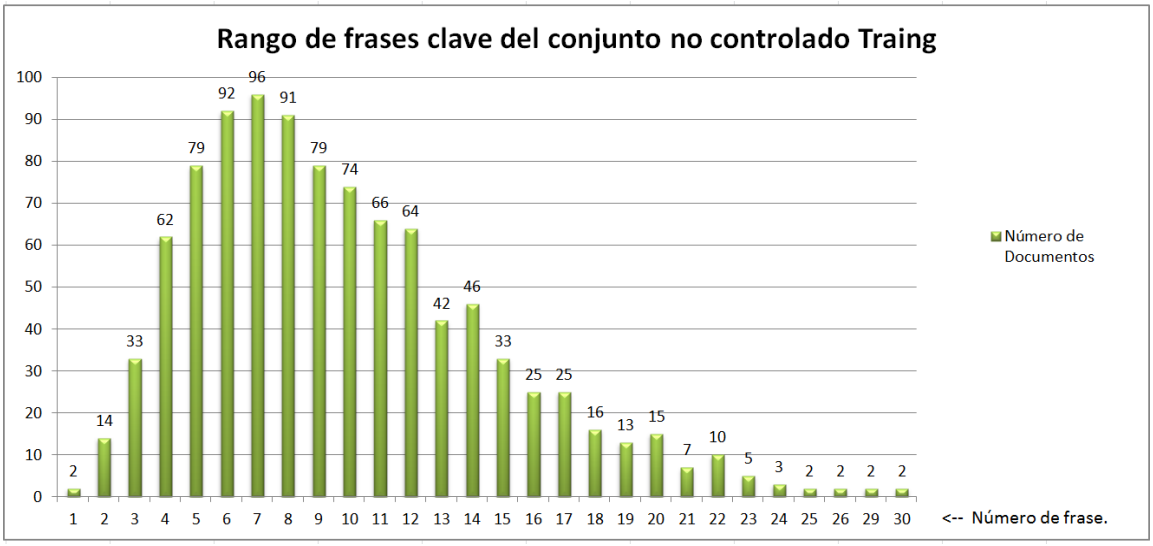

Figura 5.3 Relación del número de documentos para cada rango de número de frases clave, Traing.

De la misma manera, en la figura 5.4, se muestra la cantidad de documentos que existen con un número dado de frases clave, para el conjunto *Test*, donde se aprecia que existen más documentos con 6 y 7 frases clave, mientras que con 5 y 6 la diferencia es de un solo documento para 8 frases.

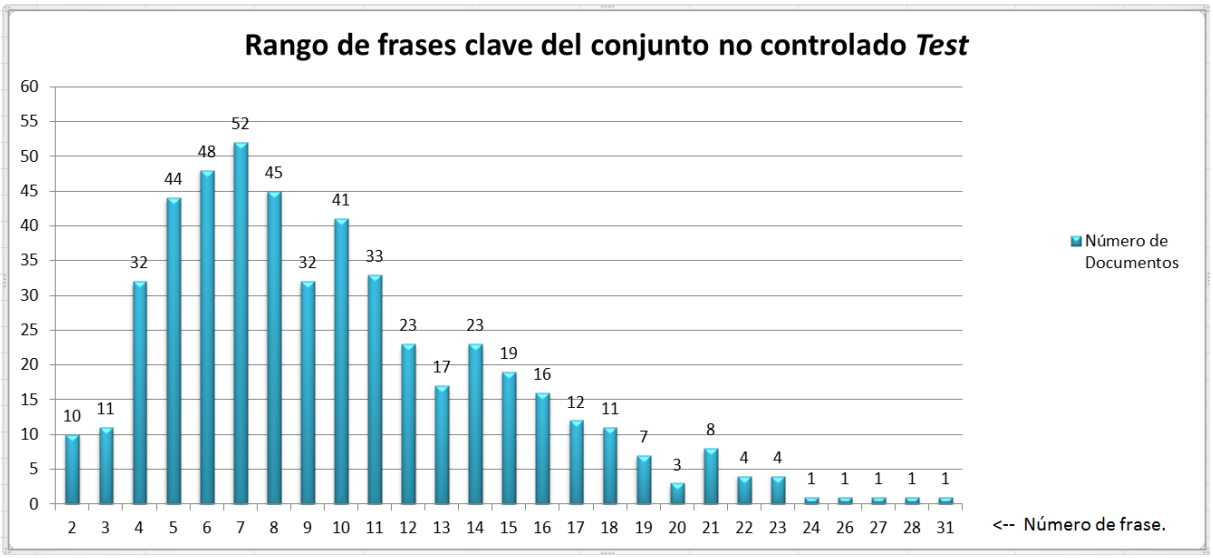

Figura 5.4 Relación del número de documentos para cada rango de número de frases clave, Test.

5.2 Prueba con sistema

5.2.1 Sistema KEA

Como prueba preliminar se ocupó el Sistema KEA (Witten et al., 1999), seleccionado y aplicado en su versión 5.0 como se muestra en el trabajo realizado de Padilla (Padilla, 2016). Obteniendo los siguientes resultados.

En la tabla 5.1, se pueden apreciar los resultados para el conjunto de frases no controladas comparadas con las frases extraídas por patrones, con una puntuación de 33.598. Posteriormente en la tabla 5.2, se muestra los datos para el conjunto controlado obteniendo una puntuación 33.34%. Finalmente en la tabla 5.3, se muestran resultados para los dos conjunto presentados, No controlado + Controlado. Obtenido como puntuación 15.374%, un resultado bajo en comparación con los no controlados.

Los conjuntos no controlado y controlado muestran su mayor puntuación en el top 5 donde evalúa solo para 5 candidatos mientras que el conjunto combinado lo muestra en el top 15.

Top	Precisión	Recuerdo	F-Measure
5	43.674	34.196	**33.598**
10	56.566	26.573	32.03
15	61.267	23.545	30.539
20	62.947	22.29	29.596

Tabla 5.1 Resultados conjunto no controlado, utilizando en sistema KEA de Witten et al., 1999.

Top	Precisión	Recuerdo	F-Measure
5	43.771	33.794	**33.34**
10	57.419	24.659	30.615
15	62.735	20.328	27.884
20	64.938	18.255	26.087

Tabla 5.2 Resultados conjunto controlado, utilizando en sistema KEA de Witten et al., 1999.

Top	Precisión	Recuerdo	F-Measure
5	17.924	15.725	14.458
10	20.769	16.276	15.332
15	21.18	16.284	**15.374**
20	21.18	16.283	15.372

Tabla 5.3 Resultados conjunto combinado, utilizando en sistema KEA de Witten et al., 1999.

En la figura 5.5, se muestran los mejores resultados de cada conjunto evaluado, y se observan resultados cercanos para el conjunto no controlados y controlado; Por otro lado, se esperaba que el conjunto combinado superara o se igualara al conjunto no controlado, sin embargo, no se presentó de esta forma.

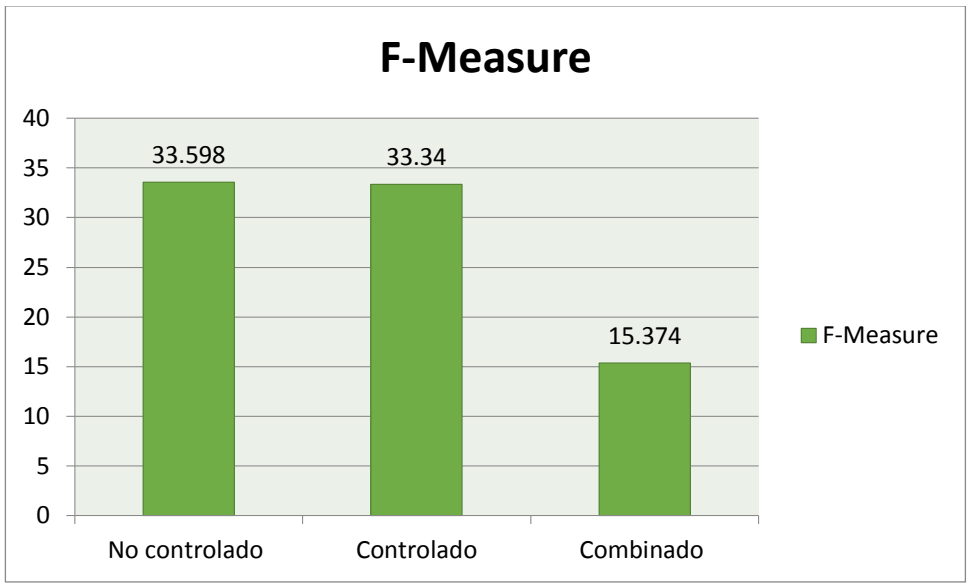

Figura 5.5 Gráfica de los resultados obtenidos con KEA para los tres conjuntos.

Los resultados preliminares nos permiten determinar si nuestro trabajo tiene buenos resultados y podemos llegar a una culminación, como propuesta se considera evaluar también para conjunto de frases controladas así como para un conjunto combinado de frases controladas con no controladas.

5.3 Pruebas

Las pruebas se realizaron con diferentes parámetros como:

1. *Umbral:* para extraer diferentes patrones léxicos.
2. *Stopwords:* lista de stopwords que se ocupó (ver anexo 1)
3. Stemming: dar a conocer que algoritmo de stemming fue aplicado
4. *Conjunto*: para que conjunto del corpus fue aplicado, no controlado, controlado o combinado.
5. *Formato*: también se aplicó un formato a los *Abstract* originales del corpus Inspec, que consistió en la colocación de punto (.) al final del título y al final del texto, debido a que los archivos no los contenían.

Los resultados que se muestran en las tablas presentan una evaluación para diferentes pesados de término, como: Booleano (B), Precisión (P), Recuerdo (R) Y F-Measure (F). Con las métricas de Precisión (P), Recuerdo (R) y F-Measure (F-M).

De igual forma se presentan los resultados para frases clave candidatas, 5, 10, 15 y 20 para evaluador ROUGE.

Cabe mencionar que para los experimentos 7 y 8 se aplicaron los patrones obtenidos en el trabajo de Hernández (Hernández, 2016) y se aplicaron a este presente trabajo de tesis.

5.3.1 Experimento 1

Parámetros

- Umbral: 1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: No

Total de patrones: Longitud1 = 2; Longitud2 = 7; Longitud3 = 7.

Longitud 1	Longitud 2	Longitud 3
THE @LINK OF	THE @LINK @LINK @PUNTO	THE @LINK @LINK @LINK @PUNTO
@PBR @LINK @PCI	AND @LINK @LINK @PUNTO	THE @LINK @LINK @LINK @PBR
	THE @LINK @LINK @COMA	THE @LINK @LINK @LINK OF
	@COMA @LINK @LINK @COMA	OF @LINK @LINK @LINK @PUNTO
	OF @LINK @LINK @PUNTO	OF @LINK @LINK @LINK IN
	THE @LINK @LINK AND	THE @LINK @LINK @LINK @COMA
	THE @LINK @LINK OF THE	A @LINK @LINK @LINK @PUNTO

Tabla 5.4 Mejores patrones léxicos para el conjunto no controladas con umbral de 1%.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com. 1	14.643	8.051	9.532	25.026	7.783	11.193	29.274	7.465	11.312	32.353	7.453	11.54
	Com. 2	16.581	12.17	**12.965**	28.095	12.183	15.885	33.088	11.558	16.206	36.422	11.465	16.519
	Long 1	8.114	2.341	3.377	14.378	2.162	3.607	16.393	2.023	3.464	17.993	2.012	3.474
	Long 2	14.67	6.301	8.225	26.168	6.113	9.359	31.271	5.916	9.45	35.009	5.904	9.599
	Long 3	11.084	5.801	7.313	18.043	5.311	7.798	21.746	4.965	7.671	23.851	4.874	7.662
P	Com. 1	14.676	8.103	9.573	25.089	7.818	11.236	29.319	7.481	11.335	32.42	7.472	11.57
	Com. 2	16.526	12.099	12.904	28.111	12.213	**15.915**	33.097	11.568	**16.223**	36.511	11.511	**16.585**
	Long 1	8.114	2.341	3.377	14.378	2.162	3.607	16.393	2.023	3.464	17.993	2.012	3.474
	Long 2	14.714	6.353	8.273	26.168	6.113	9.359	31.271	5.916	9.45	35.009	5.904	9.599
	Long 3	11.232	5.939	7.455	18.043	5.311	7.798	21.746	4.965	7.671	23.851	4.874	7.662
R	Com. 1	14.625	8.026	9.511	25.009	7.773	11.18	29.272	7.467	11.314	32.35	7.454	11.541
	Com. 2	16.415	11.917	12.763	28.031	12.13	15.83	33.026	11.511	16.155	36.354	11.421	16.47
	Long 1	8.114	2.341	3.377	14.378	2.162	3.607	16.393	2.023	3.464	17.993	2.012	3.474
	Long 2	14.692	6.329	8.25	26.168	6.113	9.359	31.271	5.916	9.45	35.009	5.904	9.599
	Long 3	11.116	5.817	7.336	18.043	5.311	7.798	21.746	4.965	7.671	23.851	4.874	7.662
F	Com. 1	14.643	8.051	9.532	25.009	7.773	11.18	29.272	7.467	11.314	32.35	7.454	11.541
	Com. 2	16.433	11.941	12.784	27.994	12.093	15.793	32.989	11.486	16.126	36.354	11.421	16.47
	Long 1	8.114	2.341	3.377	14.378	2.162	3.607	16.393	2.023	3.464	17.993	2.012	3.474
	Long 2	14.714	6.353	8.273	26.168	6.113	9.359	31.271	5.916	9.45	35.009	5.904	9.599
	Long 3	11.116	5.817	7.336	18.043	5.311	7.798	21.746	4.965	7.671	23.851	4.874	7.662

Tabla 5.5 Resultados para conjunto no controladas con umbral de 1%.

5.3.2 Experimento 2

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo
- Conjunto: No controlado
- Formato: No

Total de patrones: Longitud1 = 62; Longitud2 = 206; Longitud3 = 136.

Longitud 1	Longitud 2	Longitud 3
OF @LINK AND	@COMA @LINK @LINK @COMA AND	THE @LINK @LINK @LINK IS
@COMA @LINK @COMA AND	AND @LINK @LINK @PUNTO THE	OF @LINK @LINK @LINK AND
OF @LINK @COMA	OF THE @LINK @LINK @COMA	A @LINK @LINK @LINK @PUNTO THE
THE @LINK OF THE	OF @LINK @LINK FOR	A @LINK @LINK @LINK @PBR
OF @LINK @PUNTO	OF @LINK @LINK @PUNTO THE	A @LINK @LINK @LINK WITH
TO @LINK THE	FOR @LINK @LINK @COMA	OF @LINK @LINK @LINK @COMA
AND @LINK @COMA	THE @LINK @LINK FOR	AND @LINK @LINK @LINK AR
THE @LINK IS	OF @LINK @LINK OF	OF THE @LINK @LINK @LINK @COMA
THE @LINK @GM	THE @LINK @LINK @COMA WHICH	OF @LINK @LINK @LINK @PBR
IN @LINK AND	THE @LINK @LINK @PBR	OF @LINK @LINK @LINK FOR

Tabla 5.6 Mejores patrones léxicos para conjunto no controlado con umbral de 0.1%.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	24.532	25.088	23.188	36.661	23.1	26.576	41.744	21.686	26.868	44.003	21.113	26.783
	Com 2	24.855	29.011	25.598	36.306	30.207	31.299	41.651	29.035	32.462	44.224	28.367	32.653
	Long 1	20.444	10.038	12.45	29.793	8.319	12.359	33.772	7.633	11.814	35.495	7.382	11.571
	Long 2	26.775	18.685	20.734	40.634	17.981	23.106	46.726	16.918	23.08	49.227	16.505	22.937
	Long 3	22.162	14.425	16.779	32.078	13.954	18.247	38.077	13.451	18.661	40.711	13.116	18.594
P	Com 1	24.722	25.524	23.477	36.479	22.976	26.431	41.508	21.541	26.696	43.765	20.968	26.615
	Com 2	25.661	30.223	**26.557**	36.686	30.693	**31.718**	42.056	29.404	**32.824**	44.787	28.814	**33.117**
	Long 1	20.444	10.038	12.45	29.793	8.316	12.357	33.732	7.612	11.786	35.442	7.36	11.541
	Long 2	27.437	19.518	21.462	40.656	18.001	23.127	46.726	16.918	23.08	49.227	16.505	22.937
	Long 3	22.387	14.687	17.032	32.135	14.009	18.303	38.077	13.451	18.661	40.711	13.116	18.594
R	Com 1	24.448	24.576	22.947	36.476	22.857	26.374	41.532	21.524	26.686	20.952	43.819	26.607
	Com 2	25.164	28.76	25.686	36.22	29.816	31.076	41.418	28.807	32.228	44.201	28.271	32.572
	Long 1	20.431	10.014	12.433	29.767	8.295	12.334	33.732	7.61	11.785	35.442	7.359	11.539
	Long 2	27.131	19.155	21.132	40.645	17.992	23.117	46.726	16.918	23.08	49.227	16.505	22.937
	Long 3	22.22	14.572	16.891	32.119	13.988	18.284	38.077	13.451	18.661	40.711	13.116	18.594
F	Com 1	24.569	25.108	23.212	36.532	22.992	26.458	41.573	21.551	26.718	43.86	20.976	26.637
	Com 2	25.202	29.463	25.982	36.222	30.173	31.229	41.644	29.064	32.459	44.441	28.525	32.812
	Long 1	20.444	10.038	12.45	29.78	8.305	12.345	33.732	7.61	11.785	35.442	7.359	11.539
	Long 2	19.122	27.106	21.103	40.645	17.992	23.117	46.726	16.918	23.08	49.227	16.505	22.937
	Long 3	22.283	14.621	16.949	32.119	13.995	18.288	38.077	13.451	18.661	40.711	13.116	18.594

Tabla 5.7 Resultados para conjunto no controlado con umbral de 0.1%.

5.3.3 Experimento 3

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo.
- Conjunto: Combinado
- Formato: No

Total de patrones: Longitud1 = 152; Longitud2 = 204; Longitud3 = 140.

Longitud 1	Longitud 2	Longitud 3
@CBR @LINK @CCI	AND THE @LINK @LINK @COMA	@COMA THE @LINK @LINK @LINK AND THE
SUCH AS @LINK AND	DIFFER @LINK @LINK @COMA	USING @LINK @LINK @LINK THIS PAPER
OF @LINK @PUNTO WE	WITH @LINK @LINK IN THIS PAPER	USING @LINK @LINK @LINK @PBR
@COMA AND @LINK @PUNTO	USING THE @LINK @LINK OF	OF A @LINK @LINK @LINK @PBR
OF @LINK @COMA A	USING A @LINK @LINK @PUNTO	THE @LINK @LINK @LINK @PUNTO THIS
ON THE @LINK @COMA	THE @LINK @LINK @PUNTO IN THIS PAPER	A @LINK @LINK @LINK AND A
OF @LINK AS	IN @LINK @LINK IN	@PCI AND @LINK @LINK @LINK @PBR
OF @LINK ON	FOR @LINK @LINK WE	A @LINK @LINK @LINK FOR A
IN @LINK @PUNTO THE	FOR @LINK @LINK IN THIS PAPER	A @LINK @LINK @LINK @PUNTO THE
@COMA @LINK @COMA AND	WITH A @LINK @LINK @PUNTO	A @LINK @LINK @LINK FOR THE

Tabla 5.8 Mejores patrones léxicos para conjunto combinado con umbral de 0.1% y segundo stemming.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	5.598	5.818	5.145	8.866	5.153	5.858	11.048	4.659	6.069	12.103	4.408	6.076
	Com 2	9.814	8.638	**8.433**	15.017	7.343	**9.068**	18.649	6.729	9.262	20.612	6.466	9.281
	Long 1	4.431	2.521	2.899	7.542	2.09	3.031	9.674	2.024	3.179	10.618	1.877	3.047
	Long 2	4.985	3.974	4.097	7.359	3.608	4.488	9.192	3.164	4.419	10.268	2.991	4.392
	Long 3	5.345	3.187	3.816	7.681	2.949	3.999	9.287	2.797	4.05	10.26	2.73	4.061
P	Com 1	5.81	6.006	5.335	8.93	5.194	5.903	11.237	4.742	6.183	12.343	4.497	6.205
	Com 2	9.585	8.404	8.229	14.923	7.327	9.037	18.646	6.774	**9.312**	20.629	6.509	**9.334**
	Long 1	4.431	2.521	2.899	7.542	2.09	3.031	9.674	2.024	3.179	10.618	1.877	3.047
	Long 2	5.052	4.024	4.156	7.349	3.595	4.476	9.192	3.164	4.419	10.268	2.991	4.392
	Long 3	5.169	3.043	3.659	7.681	2.949	3.999	9.287	2.797	4.05	10.26	2.73	4.061
R	Com 1	5.631	5.755	5.129	8.92	5.168	5.877	11.171	4.698	6.125	12.258	4.456	6.145
	Com 2	9.489	8.337	8.14	14.695	7.251	8.921	18.205	6.648	9.113	20.156	6.377	9.125
	Long 1	4.431	2.521	2.899	7.542	2.09	3.031	9.674	2.024	3.179	10.618	1.877	3.047
	Long 2	4.963	3.947	4.072	7.349	3.595	4.476	9.192	4.476	4.419	10.268	2.991	4.392
	Long 3	5.144	3.008	3.628	7.667	2.937	3.986	9.287	2.797	4.05	10.26	2.73	4.061
F	Com 1	5.656	5.771	5.15	8.894	5.135	5.848	11.163	4.684	6.112	12.249	4.448	6.135
	Com 2	9.516	8.311	8.147	14.752	7.231	8.923	18.283	6.646	9.126	20.227	6.388	9.148
	Long 1	4.431	2.521	2.899	7.542	2.09	3.031	9.674	202400	3.179	10.618	1.877	3.047
	Long 2	4.987	3.963	4.092	7.349	3.595	4.476	9.192	3.164	4.419	10.268	2.991	4.392
	Long 3	5.172	3.032	3.655	7.681	2.949	3.999	9.287	2.797	4.05	10.26	2.73	4.061

Tabla 5.9 Resultados para conjunto combinado con umbral de 0.1% y segundo stemming.

5.3.4 Experimento 4

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: Combinado
- Formato: No

Total de patrones: Longitud1 = 150; Longitud2 = 213; Longitud3 = 141.

Longitud 1	Longitud 2	Longitud 3
OF @LINK @COMA A	WITH A @LINK @LINK @PUNTO	OF A @LINK @LINK @LINK @PBR
INFORM SYSTEM @PBR @LINK @PCI	DIFFER @LINK @LINK @COMA	US @LINK @LINK @LINK THI PAPER
@CBR @LINK @CCI	WITH @LINK @LINK IN THI	@PCI AND @LINK @LINK @LINK @PBR
SUCH AS @LINK AND	TO @LINK @LINK @DOSPUNTO	US @LINK @LINK @LINK @PBR
OF @LINK @PUNTO WE	THE @LINK @LINK @PUNTO IN THI PAPER	AN @LINK @LINK @LINK @PBR
@COMA AND @LINK @PUNTO	SYSTEM OF @LINK @LINK @PUNTO	A @LINK @LINK @LINK WE
ON THE @LINK @COMA	OF @LINK @LINK IN THI	A @LINK @LINK @LINK WITH
US @LINK AND	IN @LINK @LINK IN	AN @LINK @LINK @LINK @PUNTO
OF @LINK AS	FOR @LINK @LINK WE	AND THE @LINK @LINK @LINK @PUNTO
IN @LINK @PUNTO THE	EFFICI @LINK @LINK FOR	A @LINK @LINK @LINK FOR THE

Tabla 5.10 Mejores patrones léxicos para conjunto combinado con umbral de 0.1% primer stemming.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	24.607	25.033	23.044	37.687	21.935	25.884	43.498	18.978	25.006	45.873	17.46	24.039
	Com 2	25.272	28.861	25.613	37.612	28.303	30.416	43.386	25.048	30.138	45.793	23.05	29.163
	Long 1	18.416	9.443	11.558	28.097	7.305	11.012	32.843	6.202	9.957	34.718	5.683	9.343
	Long 2	27.284	19.202	21.264	42.261	17.47	23.02	48.934	15.085	21.701	51.385	13.811	20.585
	Long 3	20.772	13.158	15.481	30.48	12.382	16.496	36.598	11.195	16.139	38.869	10.221	15.299
P	Com 1	25.794	25.031	23.592	37.738	21.993	25.947	43.583	19.024	25.067	46.015	17.495	24.1
	Com 2	25.828	29.803	**26.308**	37.798	28.55	**30.634**	43.923	25.427	**30.569**	46.686	23.515	**29.758**
	Long 1	18.402	9.418	11.541	28.084	7.291	10.998	32.817	6.186	9.937	34.692	5.672	9.328
	Long 2	28.063	20.15	22.114	42.283	17.489	23.04	48.934	15.085	21.701	51.385	13.811	20.585
	Long 3	21.337	13.671	16.018	30.495	12.388	16.505	36.598	11.195	16.139	38.869	10.221	15.299
R	Com 1	24.662	25.167	23.137	37.675	21.955	25.899	43.435	18.935	24.958	45.858	17.419	23.998
	Com 2	25.283	28.848	25.616	37.272	28.033	30.125	43.115	24.861	29.919	45.931	23.066	29.215
	Long 1	18.416	9.443	11.558	28.097	7.305	11.012	32.843	6.202	9.957	34.718	5.683	9.343
	Long 2	27.78	19.79	21.796	42.273	17.475	23.029	48.934	15.085	21.701	51.385	13.811	20.585
	Long 3	20.785	13.239	15.537	30.48	12.371	16.489	36.598	11.195	16.139	38.869	10.221	15.299
F	Com 1	24.612	25.137	23.096	37.559	21.877	25.807	43.324	18.879	24.885	45.739	17.363	23.925
	Com 2	25.326	28.94	25.68	37.297	28.092	30.166	43.212	24.949	30.01	46.034	23.14	29.295
	Long 1	18.416	9.443	11.558	28.097	7.305	11.012	32.843	6.202	9.957	34.718	5.683	9.343
	Long 2	28.556	14.466	18.775	43.195	16.249	22.514	49.071	14.829	21.559	51.424	13.795	20.586
	Long 3	20.998	13.391	15.714	30.509	12.4	16.518	36.598	11.195	16.139	38.869	10.221	15.299

Tabla 5.11 Resultados para conjunto combinado con umbral de 0.1% primer stemming.

5.3.5 Experimento 5

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: Si

Total de patrones: Longitud1 = 92; Longitud2 = 191; Longitud3 = 131.

Longitud 1	Longitud 2	Longitud 3
OF @LINK IN COMPUT	DIFFER @LINK @LINK @COMA	CONTROL OF @LINK @LINK @LINK @PUNTO
SUCH AS @LINK AND	WITH @LINK @LINK @PUNTO IN	@COMA THE @LINK @LINK @LINK AND THE
AND @LINK TECHNOLOG	THE @LINK @LINK @PUNTO IT IS	US @LINK @LINK @LINK @PUNTO THI PAPER
@GM BASE @LINK @COMA	ON @LINK @LINK @PUNTO THI	WITH A @LINK @LINK @LINK @PUNTO
OF @LINK ON	A @LINK @LINK @PUNTO WE	A @LINK @LINK @LINK WITH
THE @LINK @PUNTO IN	SUCH AS @LINK @LINK AND	THE @LINK @LINK @LINK @PUNTO THI
@COMA AND @LINK @PUNTO	BASE ON @LINK @LINK @PUNTO	THE @LINK @LINK @LINK @PUNTO A
FOR @LINK SYSTEM	THE @LINK @LINK @COMA WHICH	AN @LINK @LINK @LINK @COMA
AND THE @LINK @PUNTO	SYSTEM OF @LINK @LINK @PUNTO	AN @LINK @LINK @LINK OF
IN @LINK @PUNTO THE	OF AN @LINK @LINK @PUNTO	A @LINK @LINK @LINK @PUNTO THE

Tabla 5.12 Mejores patrones léxicos para conjunto no controlado con umbral de 0.1% primer stemming y con formato.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	23.764	25.808	23.232	36.045	23.996	27.137	40.851	22.544	27.409	43.162	21.975	27.339
	Com 2	24.371	29.678	25.713	35.913	31.029	31.718	40.992	29.632	32.696	43.631	29.024	32.929
	Long 1	19.459	11.366	13.348	28.705	9.393	13.481	32.866	8.711	13.076	34.707	8.47	12.889
	Long 2	26.316	19.089	20.904	40.556	18.449	23.661	46.464	17.358	23.597	48.922	16.92	23.421
	Long 3	21.942	14.808	17.032	32.37	14.448	18.755	37.633	13.63	18.764	40.261	13.286	18.722
P	Com 1	24.179	26.549	23.775	36.305	24.205	27.354	41.117	22.711	27.604	43.463	22.154	27.554
	Com 2	25.138	30.875	**26.636**	36.259	31.548	**32.135**	41.501	30.103	**33.164**	44.276	29.448	**33.406**
	Long 1	19.472	11.391	13.365	28.705	9.393	13.481	32.84	8.697	13.058	34.681	8.459	12.874
	Long 2	27.003	19.867	21.622	40.598	18.495	23.706	46.464	17.358	23.597	48.922	16.92	23.421
	Long 3	22.529	15.424	17.636	32.455	14.527	18.836	37.633	13.63	18.764	40.261	13.286	18.722
R	Com 1	23.977	26.165	23.977	36.048	23.955	27.109	40.882	22.502	27.389	43.267	21.956	27.356
	Com 2	24.53	29.842	25.874	36.039	31.184	31.85	41.173	29.753	32.837	43.971	29.13	33.105
	Long 1	19.459	11.366	13.348	28.679	9.372	13.458	32.853	8.705	13.068	34.694	8.465	12.882
	Long 2	26.912	19.754	21.525	40.577	18.475	23.685	46.464	17.358	23.597	48.922	16.92	23.421
	Long 3	22.246	15.16	17.357	32.383	14.471	18.773	37.633	13.63	18.764	40.261	13.286	18.722
F	Com 1	24.057	26.291	23.603	36.043	23.94	27.097	40.846	22.484	27.364	43.243	21.945	27.34
	Com 2	24.791	30.236	26.184	36.122	31.287	31.938	41.301	29.858	32.943	44.099	29.22	33.198
	Long 1	19.459	11.366	13.348	28.679	9.372	13.458	32.853	8.705	13.068	34.694	8.465	12.882
	Long 2	26.981	19.842	21.6	40.577	18.475	23.685	46.464	17.358	23.597	48.922	16.92	23.421
	Long 3	22.358	15.292	17.485	32.384	14.47	18.773	37.633	13.63	18.764	40.261	13.286	18.722

Tabla 5.13 Resultados para conjunto no controlado con umbral de 0.1% primer stemming y con formato.

5.3.6 Experimento 6

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo.
- Conjunto: No controlado
- Formato: Si

Total de patrones: Longitud1 = 176; Longitud2 = 213; Longitud3 = 154.

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 5.14 Mejores patrones léxicos para conjunto no controlado con umbral de 0.1% segundo stemming y con formato.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	20.406	26.162	21.248	33.357	25.045	26.729	39.099	23.291	27.41	42.003	22.621	27.546
	Com 2	21.717	30.311	23.997	33.893	32.186	31.195	39.74	30.375	32.576	43.003	29.66	33.144
	Long 1	15.581	11.316	11.971	25.357	9.441	12.92	30.023	8.565	12.587	32.323	8.26	12.433
	Long 2	22.99	19.837	20.013	37.491	19.585	23.962	44.338	18.259	24.146	47.635	17.71	24.044
	Long 3	19.311	14.755	16.076	28.61	14.288	17.865	33.531	13.169	17.766	36.945	12.799	17.908
P	Com 1	20.719	26.641	21.637	33.327	24.99	26.689	39.201	23.333	27.48	42.194	22.684	27.655
	Com 2	22.124	30.9	**24.498**	33.901	32.351	**31.289**	39.993	30.644	**32.85**	43.515	29.978	**33.534**
	Long 1	15.581	11.316	11.971	25.357	9.441	12.92	29.997	8.546	12.565	32.297	8.245	12.415
	Long 2	23.572	20.456	20.617	37.501	19.597	23.973	44.338	18.259	24.146	47.635	17.71	24.044
	Long 3	19.305	14.724	16.049	14.384	14.384	17.957	33.531	13.169	17.766	36.945	12.799	17.908
R	Com 1	20.194	25.723	20.969	33.006	24.736	26.407	38.757	23.02	27.12	41.712	22.377	27.294
	Com 2	21.452	29.783	23.65	33.558	31.871	30.874	39.58	30.171	32.406	42.969	29.526	33.063
	Long 1	15.581	11.316	11.971	25.37	9.459	12.935	30.01	8.555	12.576	32.31	8.252	12.424
	Long 2	23.222	19.999	20.217	37.48	19.568	23.948	44.338	18.259	24.146	47.635	17.71	24.044
	Long 3	19.3	14.772	16.079	28.668	14.356	17.929	33.531	13.169	17.766	36.945	12.799	17.908
F	Com 1	20.248	25.775	21.031	32.996	24.702	26.389	38.747	23.003	27.104	41.715	22.371	27.287
	Com 2	21.682	30.107	23.915	33.568	31.859	30.87	39.547	30.166	32.386	42.92	29.525	33.044
	Long 1	15.581	11.316	11.971	25.37	9.459	12.935	30.01	8.555	12.576	32.31	8.252	12.424
	Long 2	23.288	20.071	20.289	37.48	19.568	23.948	44.338	18.259	24.146	47.635	17.71	24.044
	Long 3	19.3	14.74	16.063	28.669	14.35	17.926	33.531	13.169	17.766	36.945	12.799	17.908

Tabla 5.15 Resultados para conjunto no controlado con umbral de 0.1% segundo stemming y con formato.

5.3.7 Experimento 7

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: Si

Para este experimentos se utilizó los patrones de entrenamiento tomados del trabajo de Hernández (Hernández, 2016) y se aplicaron al conjunto test del presente trabajo de tesis. Total de patrones: Longitud1 = 161; Longitud2 = 227; Longitud3 = 201.

Longitud 1	Longitud 2	Longitud 3
THE @LINK PROTOCOL	A @LINK @LINK CAN BE	BY THE @LINK @LINK @LINK @PUNTO
SET OF @LINK @COMA	@COMA THE @LINK @LINK FOR	A @LINK @LINK @LINK @PUNTO THE
AN @LINK IS	@COMA THE @LINK @LINK CAN	THE @LINK @LINK @LINK @PUNTO THIS
EACH @LINK IS	A @LINK @LINK @PUNTO WE	OF @LINK @LINK @LINK @PUNTO IN
OF @LINK ON	OF @LINK @LINK @PUNTO NUM	FOR @LINK @LINK @LINK @PUNTO WE
A @LINK IN	THE AVERAG @LINK @LINK OF	A @LINK @LINK @LINK WITH
THE @LINK HAS	OF A @LINK @LINK IS	A @LINK @LINK @LINK @COMA WE
THE @LINK PAGE	OF A @LINK @LINK @PUNTO	WITH @LINK @LINK @LINK @PBR
THE @LINK MAY	THE @LINK @LINK @COMA WHICH	USING @LINK @LINK @LINK @PUNTO
AN @LINK THAT	USING @LINK @LINK @PUNTO	USING THE @LINK @LINK @LINK @PUNTO

Tabla 5.16 Mejores patrones léxicos conjunto no controlado con umbral de 0.1% primer stemming con formato.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	22.662	24.523	21.99	34.69	22.98	25.999	39.906	21.693	26.594	42.214	21.122	26.568
	Com 2	23.388	27.899	24.32	34.972	29.336	**30.258**	40.366	28.229	**31.522**	42.923	27.545	**31.719**
	Long 1	20.053	11.774	13.533	29.669	9.886	13.985	33.301	9.088	13.502	35.04	8.806	13.289
	Long 2	25.14	17.84	19.687	38.378	17.258	22.273	45.361	16.408	22.679	48.046	15.968	22.543
	Long 3	20.898	13.246	15.536	30.676	12.701	16.844	35.726	12.002	16.789	38.117	11.718	16.755
P	Com 1	22.633	24.445	21.939	34.404	22.766	25.753	39.563	21.495	26.351	41.968	20.987	26.404
	Com 2	23.508	28.102	24.471	34.727	29.143	30.04	40.255	28.045	31.368	42.993	27.482	31.709
	Long 1	20.053	11.774	13.533	29.683	9.898	13.998	33.315	9.097	13.512	35.054	8.812	13.298
	Long 2	25.672	18.4	20.222	38.408	17.296	22.307	45.361	16.408	22.679	48.046	15.968	22.543
	Long 3	20.883	13.233	15.519	30.694	12.697	16.85	35.717	11.993	16.78	38.117	11.718	16.755
R	Com 1	22.651	24.558	21.987	34.404	22.766	25.753	39.563	21.495	26.351	41.968	20.987	26.404
	Com 2	23.652	28.416	**24.676**	34.822	29.181	30.099	40.226	27.993	31.32	42.96	27.428	31.659
	Long 1	20.053	11.774	13.533	29.683	9.898	13.998	33.315	9.097	13.512	35.054	8.812	13.298
	Long 2	25.72	18.516	20.304	38.408	17.296	22.307	45.361	16.408	22.679	48.046	15.968	22.543
	Long 3	20.789	13.195	15.456	30.608	12.629	16.775	35.717	11.993	16.78	38.117	11.718	16.755
F	Com 1	22.686	24.54	22.002	34.41	22.752	25.743	39.581	21.495	26.353	41.976	20.981	26.399
	Com 2	23.629	28.336	24.636	34.838	29.182	30.106	40.276	28.019	31.354	43.004	27.448	31.687
	Long 1	20.053	11.774	13.533	29.669	9.888	13.986	33.289	9.084	13.495	35.028	8.802	13.283
	Long 2	25.724	18.475	20.288	38.408	17.296	22.307	45.361	16.408	22.679	48.046	15.968	22.543
	Long 3	20.877	13.257	15.529	30.623	12.645	16.79	35.717	11.993	16.78	38.117	11.718	16.755

Tabla 5.17 Resultados para conjunto no controlado con umbral de 0.1% primer stemming con formato.

5.3.8 Experimento 8

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo.
- Conjunto: No controlado
- Formato: Si

Para este experimentos se utilizó los patrones de entrenamiento tomados del trabajo de Hernández (Hernández, 2016) y se aplicaron al conjunto test del presente trabajo de tesis. Total de patrones: Longitud1 = 161; Longitud2 = 227; Longitud3 = 201.

Longitud 1	Longitud 2	Longitud 3
THE @LINK PROTOCOL	A @LINK @LINK CAN BE	BY THE @LINK @LINK @LINK @PUNTO
SET OF @LINK @COMA	@COMA THE @LINK @LINK FOR	A @LINK @LINK @LINK @PUNTO THE
AN @LINK IS	@COMA THE @LINK @LINK CAN	THE @LINK @LINK @LINK @PUNTO THIS
EACH @LINK IS	A @LINK @LINK @PUNTO WE	OF @LINK @LINK @LINK @PUNTO IN
OF @LINK ON	OF @LINK @LINK @PUNTO NUM	FOR @LINK @LINK @LINK @PUNTO WE
A @LINK IN	THE AVERAG @LINK @LINK OF	A @LINK @LINK @LINK WITH
THE @LINK HAS	OF A @LINK @LINK IS	A @LINK @LINK @LINK @COMA WE
THE @LINK PAGE	OF A @LINK @LINK @PUNTO	WITH @LINK @LINK @LINK @PBR
THE @LINK MAY	THE @LINK @LINK @COMA WHICH	USING @LINK @LINK @LINK @PUNTO
AN @LINK THAT	USING @LINK @LINK @PUNTO	USING THE @LINK @LINK @LINK @PUNTO

Tabla 5.18 Mejores patrones léxicos para conjunto no controlado con umbral de 0.1% segundo stemming con formato.

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15			Candidatas 20		
		P	R	F-M	P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	19.96	26.721	21.391	31.919	25.05	26.321	37.991	23.49	27.487	40.835	22.716	27.63
	Com 2	8.538	12.531	9.658	13.555	13.604	12.996	16.473	13.029	13.987	18.178	12.649	14.324
	Long 1	17.319	13.935	13.809	26.503	11.101	14.576	31.038	9.977	14.249	33.261	9.527	13.996
	Long 2	22.881	19.481	19.806	36.4	18.971	23.122	43.856	17.955	23.829	47.445	17.455	23.899
	Long 3	19.58	14.933	16.281	30.266	14.845	18.628	35.386	13.676	18.363	38.82	13.317	18.484
P	Com 1	19.943	26.604	21.382	31.967	25.094	26.373	38.086	23.531	27.55	40.987	22.81	27.746
	Com 2	20.858	29.975	**23.543**	32.448	32.328	**30.788**	38.658	30.833	**32.666**	41.792	30.012	**33.219**
	Long 1	17.293	13.864	13.771	26.504	11.097	14.575	31.013	9.951	14.223	33.236	9.504	13.971
	Long 2	23.216	19.813	20.14	36.442	19.032	23.172	43.856	17.955	23.829	47.445	17.455	23.899
	Long 3	19.307	14.509	15.941	30.3	14.857	18.65	35.386	13.676	18.363	38.82	13.317	18.484
R	Com 1	19.735	26.431	21.152	31.577	24.777	26.038	37.481	23.189	27.124	40.342	22.487	27.331
	Com 2	20.54	29.742	23.227	31.793	31.576	30.099	37.867	30.154	31.962	41.031	29.382	32.561
	Long 1	17.319	13.935	13.809	26.517	11.113	14.589	31.039	9.976	14.249	33.248	9.519	13.986
	Long 2	23.176	19.803	20.112	36.42	19.007	23.149	43.856	17.955	23.829	47.445	17.455	23.899
	Long 3	18.881	14.15	15.552	30.23	14.807	18.591	35.386	13.676	18.363	38.82	13.317	18.484
F	Com 1	19.88	26.681	21.33	31.648	24.848	26.104	37.57	23.235	27.182	40.404	22.509	27.364
	Com 2	20.748	30.042	23.473	31.933	31.733	30.239	37.966	30.254	32.052	41.147	29.486	32.663
	Long 1	17.28	13.824	13.751	26.491	11.077	14.559	31.013	9.951	14.223	33.222	9.497	13.962
	Long 2	17.28	13.824	13.751	26.491	11.077	14.559	31.013	9.951	14.223	33.222	9.497	13.962
	Long 3	19.026	14.261	15.678	30.216	14.793	18.577	35.386	13.676	18.363	38.82	13.317	18.484

Tabla 5.19 Resultados para conjunto no controlado con umbral de 0.1% segundo stemming con formato.

5.4 Discusión de resultados

Las frases extraídas por el método de Hernández (Hernández, 2016), aplicado al corpus Inspec dieron como resultado más alto un F-Measure de 33.53 para frases candidatas 20, permitiendo posicionarse en la gráfica (Gráfica 5.20) como el cuarto lugar, muy cerca del método de n-gramas de Hulth (Hulth, 2003).

Enfoque	Método	F-Measure
Supervisado	N-gramas con etiquetas (Hulth, 2003)	33.9
	Chunking con etiquetas (Hulth, 2003)	33
	Patrones con etiquetas (Hulth, 2003)	28.1
	Patrones (Hulth, 2003)	25.6
	Chunking (Hulth, 2003)	22.7
	N-gramas (Hulth, 2003)	17.6
No Supervisado	TextRank Co-occ. = 2 (Mihalcea & Tarau, 2004)	36.2
	TextRank Directo Co-occ. = 2 (Mihalcea & Tarau, 2004)	35.9
	Propuesto 6 (2° alg. Porter y formato)	**33.53**
	Propuesto 5 (1° alg. Porter y formato)	33.406
	Propuesto 8 (2° alg. Porter, formato y patrones de Hernández 2016)	33.219
	Propuesto 2 (1° alg. Porter sin formato)	33.117
	TextRank Indirecto Co-occ. = 3 (Mihalcea & Tarau, 2004)	32.6
	TextRank Indirecto Co-occ. = 5 (Mihalcea & Tarau, 2004)	32.2
	TextRank Indirecto Co-occ. = 10 (Mihalcea & Tarau, 2004)	32.2

Tabla 5.20 Resultados del estado del arte con resultados del presente trabajo de tesis.

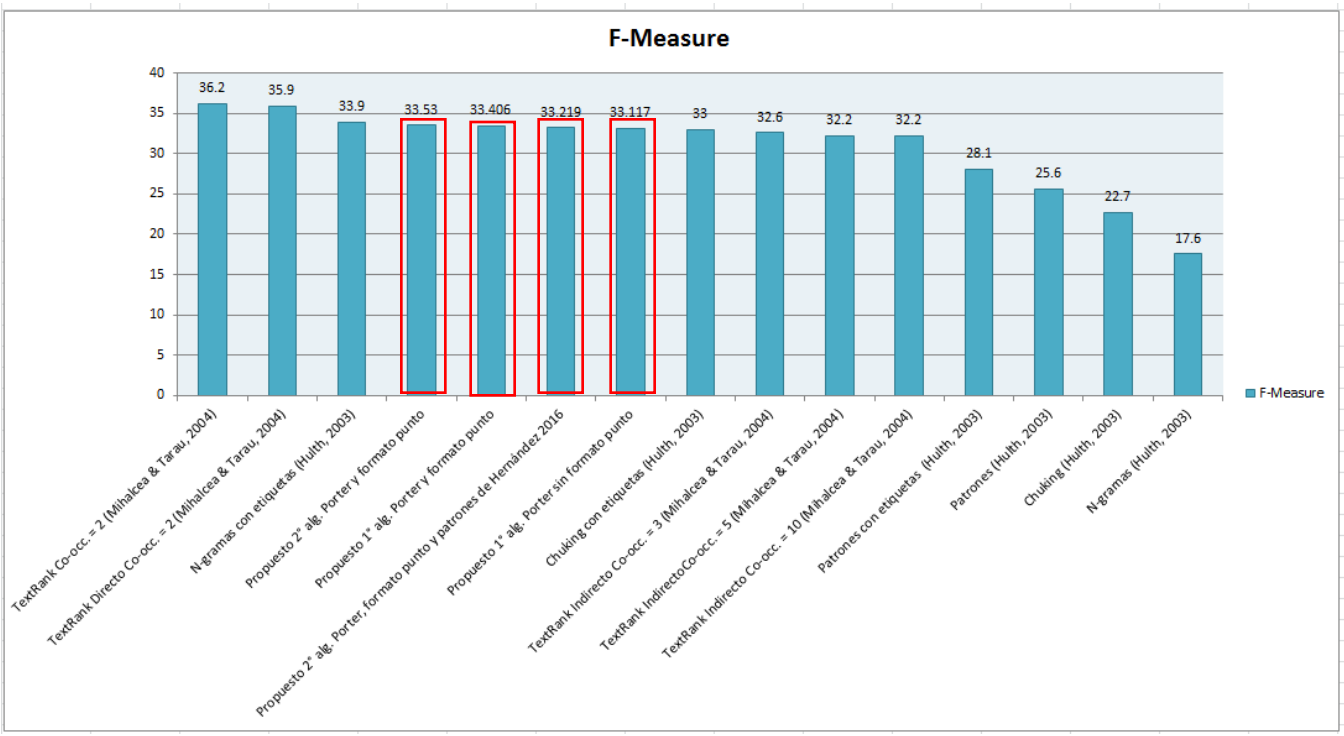

Figura 5.6 Gráfica de los métodos del estado del arte y el método aplicado en el presente trabajo de tesis.

CAPÍTULO 6.

Conclusiones

En este capítulo, se presentan las conclusiones del trabajo de tesis, así como los trabajos futuros.

6.1 Conclusiones

En la tesis se cumplieron los siguientes objetivos:

Se identificaron técnicas y métodos para la tarea de extracción de frases clave; como el sistema KEA que presentó buenos resultados para frases candidatas 20 obteniendo para el conjunto no controlado un F-Measure de 29.596, para el conjunto controlado se obtuvo 26.087 de F-Measure y por último para el conjunto combinado se obtuvo 15.372. Con base en estos resultados se pudo concluir que el conjunto combinado no es la mejor opción como *Gold Standard* debido a que presenta bajos resultados en comparación con el conjunto no controlado, que se ha utilizado en trabajos del estado del arte.

En la aplicación del método se pudo observar que presenta una frecuencia donde los mejores resultados son para el pesado *P* en la combinación 2, mismos esquemas se presentaron en el trabajo de Hernández (Hernández, 2016).

Por otra parte, el corpus seleccionado mostró un buen desempeño para la tarea de extracción de frases clave del experimento 1, donde el umbral fue de 1% la F-Measure fue de 16.6, pero en el experimento 2, cuando el umbral se asigna a 0.1%, con los mismos parámetros el F-Measure incrementó a 33.17. Posteriormente, para los demás experimentos el umbral se conservó a 0.1% y los resultados ya no incrementaron tan drásticamente.

Otra conclusión importante se deriva de la asignación de un punto al final del título de cada abstract y otro al finalizar el texto. En el experimento 5 se obtiene F-Measure de 33.41 en comparación con el experimento 2 que obtuvo 33.21. Los resultados no son notorios pero confirma que la consideración del título es fundamental para una mejor obtención de frases clave.

También se pudo concluir que los patrones extraídos de artículos científicos son suficientes para extraer frases clave. Esto con base en los experimentos 5 y 7 y los experimentos 6 y 8, en los cuales los primeros (experimento 5 y experimento 6) fueron obtenidos mediante el entrenamiento del presente trabajo de tesis consiguiendo

como F-Measure 33.406 y 33.534 respectivamente. Mientras que los experimentos 7 y 8 fueron obtenidos con los patrones de entrenamiento del trabajo de Hernández (Hernández, 2016) presentando como F-Measure de 31.719 y 33.219 respectivamente.

Para finalizar, los mejores resultados obtenidos en el presente trabajo de tesis fue para frases candidatas 20 con un F-Measure de 33.534, donde el umbral fue 0.1%, aplicando el segundo algoritmo de Porter, con formato del "." y para el conjunto no controlado.

Retomando lo mencionado anteriormente, para el pesado *P* en su combinación 2.

6.2 Trabajo futuro

Aplicar el método de TextRank (Mihalcea & Tarau, 2004) para la tarea de extracción de frases clave, ya que presenta buen desempeño en el estado del arte.

Realizar pruebas sobre un conjunto de test diferente, para conocer qué tan importante es la posición de una frase clave en el texto.

Crear un corpus en idioma español para la tarea de extracción de frases clave.

¿Un corpus en idioma español, podrá funcionar igual o mejor a un corpus en idioma inglés para la tarea de extracción de frases clave utilizando patrones léxicos?

Referencias

(Camacho, 2015) Camacho, A. Marcela. (2015). Detección de Fragmentos de Texto como Candidato a Hipervínculo. UAEM. Tesis de Maestría. Tianguistenco, Edo de México. p. 73.

(Eíto & Senso, 2004) Eíto Brun R., Senso J. (2004). Minería textual. *El profesional de la información*. p. 17. Recuperado el 20 de Septiembre de 2016 de http://www.elprofesionaldelainformacion.com/contenidos/2004/enero/2.pdf

(Frank, et al, 1999) Frank, E., Paynter, G.D. and Witten, I. (1999) Domain-Specific Keyphrase Extraction. *In Proceeding of 16th International Joint Conference on Artificial Intelligence*. p. 6. Recuperado el 18 de Marzo de 2016 de http://researchcommons.waikato.ac.nz/handle/10289/1508

(Garcia, 2004) 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 vol. 3287, pp. 478-486. Springer-Verlag.

(Garcia, 2006) 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 vol. 3878, pp. 514–523, Springer-Verlag.

(García, 2007) García Hernández Rene Arnulfo. (2007). Desarrollo de Algoritmos para el Descubrimiento de Patrones Secuenciales Maximales. INAOE. Puebla, México 2007.

(Gelbukh & Sidorov, 2006) Gelbukh, Alexander & Sidorov, Grigori. (2006). Procesamiento automático del español con enfoque en recursos léxicos grandes .Centro de Investigación en Computación Instituto Politécnico Nacional México. Pág. 133. Recuperado el 8 de Junio de 2016 de http://www.gelbukh.com/libro-procesamiento/LibroPLN.pdf

(Gelbukh, 2010) Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones.Korpus Sapiens. Sociedad Mexicana de inteligencia artificial, 1.
Recuperado el 29 de Junio de 2016 de http://nlp.gelbukh.com/Publications/2010/Procesamiento%20de%20lenguaje%20natural%20y%20sus%20aplicaciones.pdf

(Hasan & Ng, 2010) 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. Recuperado el 18 de Marzo de 2016 de http://dl.acm.org/citation.cfm?id=1944608

(Hasan & Ng, 2014) 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.
Recuperado el 09 de Febrero de 2016 de http://acl2014.org/acl2014/P14-1/pdf/P14-1119.pdf

(Hernández, 2016) Hernández Casimiro, Y. (2016). Extracción de Frases Clave Usando Patrones Léxicos en Artículos Científicos. Tesis Maestría en Ciencias de la Computación. UAEM. p.124.

(Hulth, 2003) Hulth, A. (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *Department of Computer and Systems Sciences Stockholm University Swedez*. p. 8. Recuperado el 18 de Marzo de 2016 de

http://www.aclweb.org/anthology/W03-1028

(Hulth, 2004) Hulth, A. (2004). Enhancing Linguistically Oriented Automatic Keyword Extraction. *Department of Computer and Systems Sciences Stockholm University*. p. 4. Recuperado el 18 de Marzo de 2016 de http://dl.acm.org/citation.cfm?id=1613989

(Jones & Paynter , 2001) Jones, S. and Paynter, G. W. (2001). Human evaluation of KEA, an Automatic Keyphrasing System. *Department of Computer Science University of Waikato*. June p. 9. Recuperado el 20 de Marzo de 2016 de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1847&rep=rep1&type=pdf

(Kan, Luong & Nguyen, 2010) Kan, M-Y., Luong, M-T., and Nguyen, T. D. (2010). Logical Structure Recovery in Scholarly Articles with Rich Document Features. *National University of Singapore, Singapore*. p. 25. Recuperado el 20 de Marzo de 2016 de http://dl.acm.org/citation.cfm?id=2436647

(Kim & Kan, 2009) 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. Recuperado el 20 de Marzo de 2016 de http://www.aclweb.org/anthology/W09-2902.pdf

(Kim et al., 2010) 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. Recuperado el 20 de Marzo de 2016 de http://www.aclweb.org/anthology/S10-1004

(Kim, et al., 2012) Kim, S. N., Medelyan, O., Kan, M-Y. And Baldwin, T. (2012). Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation.* p. 22. Recuperado el 20 de Marzo de 2016 de https://www.comp.nus.edu.sg/~kanmy/papers/10.1007_s10579-012-9210-3.pdf

(Ledeneva, 2008) Ledeneva Yulia Nikolaevna. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization (Doctoral). Instituto Politécnico Nacional. November.

(Ledeneva, 2008a) Ledeneva Yulia, Alexander Gelbukh, René García Hernández. (2008). Terms Derived from Frequent Sequences for Extractive Text Summarization. LNCS 4919, pp. 593-604, Springer-Verlag, ISSN 0302-9743.

(Ledeneva, 2010) Ledeneva Yulia, René García Hernández, Alexander Gelbukh. (2010). Multi-document Summarization using Maximal Frequent Sequences. Research in Computer Science, pp.15-24, vol. 47, ISSN 1870-4069. (INDIZADO POR LATINEX)

(Ledeneva, 2014) Ledeneva Yulia, René García-Hernández, Alexander Gelbukh. (2014). Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization. LNCS Springer-Verlag, ISSN 0302-9743, vol. 8404, pp. 466-480. DOI: 10.1007/978-3-642-54903-8_39.

(Lin, 2004) Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8). July 2004. p. 8. Recuperado el 01 de Julio de 2016 de http://anthology.aclweb.org/W/W04/W04-1013.pdf

(Liu et al, 2009) 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. Recuperado el 18 de Marzo de 2016 de http://www.aclweb.org/anthology/D09-1027

(Liu et al., 2010) Liu Z, Huang W, Zheng Y, and Sun M. (2010). Automatic keyphrase extraction via topic decomposition. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. p. p. 366–376. Recuperado el 2 de Junio de 2016 de http://www.personal.psu.edu/wzh112/publications/emnlp2010.pdf

(Medelyan & Witten, 2006) Medelyan, O., Witten, I. H., (2006). Thesaurus Based Automatic Keyphrase Indexing. *JCDL*. p. 2. Recuperado el 20 de Marzo de 2016 de http://www.cs.waikato.ac.nz/~ihw/papers/06-OM-IHW-Thesaurus-auto-keyphrase.pdf

(Medelyan et al , 2009) Medelyan, O., Frank, E., Witten, I. H., (2009). Human-competitive tagging using automatic keypharase extraction. *Computer Science Department University of Waikato. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. p. p. 1318–1327.* Recuperado el 16 de Marzo de 2016 de http://www.aclweb.org/anthology/D09-1137

(Mihalcea & Tarau, 2004) 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. Recuperado el 04 de Mayo de 2016 de https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

(Mihalcea, 2004) Mihalcea, Rada (2004). Graph-based Ranking Algorithms for Sentence Extraction,
Applied to Text Summarization. *Department of Computer Science University of North Texas.* p. 4.
Recuperado el 16 de Marzo de 2016 de http://www.aclweb.org/anthology/P04-3020

(Mondal & Maji, 2013) Mondal, Kumar Amit & Maji Kumar Dipak. (2013).
Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text. Department of Computer Science and Engineering Indian Institute of Technology, Kanpur Kanpur, India.
Recuperado el 24 de Junio de 2016 de
http://www.cs.northwestern.edu/~akm175/docs/btp.pdf

(Montes, 2003) Montes y Gómez, Manuel. (2003). Minería de texto: Un nuevo reto computacional. *Laboratorio de Lenguaje Natural, Centro de Investigación en Computación, IPN.*
Recuperado el 02 de Noviembre de 2015 de https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

(Montiel, 2009) Montiel Soto Romyna, René García Hernández, Yulia Ledeneva, Rafael Cruz Reyes. (2009). *Comparación de Tres Modelos de Representación de Texto en la Generación Automática de Resúmenes.* Sociedad Española para el Procesamiento del Lenguaje Natural, vol. 43, pp. 303-311, ISSN 1135-5948.

(Nguyen & Kan, 2007) Nguyen, T. D. and Kan, M-Y. (2007). Keyphrase Extraction in Scientific Publications. *Department of Computer Science, School of Computing*. p. 10.
Recuperado el 29 de Febrero de 2016 de https://www.comp.nus.edu.sg/~kanmy/papers/icadl2007.pdf

(Nguyen, 2007) Nguyen, T. D. (2007), Automatic Keyphrase Generation, *Department of Computer Science School of Computing National University of Singapore*, Tesis. Recuperado el 18 de Marzo de 2016 de http://wing.comp.nus.edu.sg/downloads/keyphraseCorpus/

(Ortiz, 2010) 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. Recuperado el 15 de Junio de 2015.

(Padilla, 2016) Padilla Camacho, Jesús Ernesto (2016). Evaluación de sistemas de extracción de frases clave. Tesis de Licenciatura. UAEM.

(Popova et al., 2013) Popova, S., Kovriguina, L., Mouromtsev, D., & Khodyrev, I. (2013). Stop-words in keyphrase extraction problem. In *Open Innovations Association (FRUCT)*, November *2013 14th Conference of* IEEE. p. p. 113-121. Recuperado el 21 de Junio de 2016 de https://fruct.org/publications/fruct14/files/Pop_23.pdf

(Porter, 1980) Porter, M.F. (1980). An algorithm for suffix stripping. *Computer Laboratory*. Vol. 14 No.3, pp. 130-137. Recuperado el 15 de Febrero de 2016 de http://www.cs.toronto.edu/~frank/csc2501/Readings/R2_Porter/Porter-1980.pdf

(Sidorov, 2013) Sidorov. G. (2013). Construcción No lineal de N-gramas en la lingüística computacional. Sociedad Mexicana de la inteligencia artificial. México 2013. p. 166. Recuperado el 8 de Junio de 2016 de http://www.cic.ipn.mx/~sidorov/sn_gramas.pdf

(Tolosa & Bordignon, 2008) Tolosa G. H. y Bordignon F.R.A. (2008).Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos. *Universidad Nacional de Luján, Argentina.* p. 149. Recuperado el 2 de Junio del 2016. http://eprints.rclis.org/12243/1/Introduccion-RI-v9f.pdf

(Tsatsaronis et al., 2010) 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. Recuperado el 24 de Junio de 2016 de https://www.dit.hua.gr/~varlamis/Varlamis-papers/C47.pdf

(Turney, 1999) Turney, P. D. (1999). Learning to Extract Keyphrases from Text. p. 45. Recuperado el 20 de Marzo de 2016 de http://cogprints.org/1802/5/ERB-1057.pdf

(Turney, 2000) Turney, P. D. (2000) Learning Algorithms for Keyphrase Extraction. p. 46. Recuperado el 20 de Marzo de 2016 de http://www.extractor.com/IR2000.pdf

(Witten et al, 1999) Witten, I. H., Paynter, G. W., Frank, E. Gutwin, C. and Nevill-Manning C. G. (1999) KEA: Practical Automatic Keyphrase Extraction. p. 9. Recuperado el 20 de Marzo de 2016 de http://dl.acm.org/citation.cfm?id=313437

(Zesch & Gurevych 2009) Zesch, T., & Gurevych, I. (2009). Approximate Matching for Evaluating Keyphrase Extraction. In *RANLP*, September 2009. p. p. 484-489.
Recuperado el 21 de Junio de 2016 de
http://www.aclweb.org/anthology/R09-1086

Anexos

Anexo 1. Lista de Stopwords

Se adjunta las listas de stopwords utilizadas en el trabajo.

a. Lista 1

a,about,after,again,all,almost,also,although,always,am,among,an,and,another,any,approximately,are,as,at,be,because,been,before,being,between,both,but,by,can,can't,could,couldn't,did,didn't,do,don't,does,doesn't,done,due,during,each,either,enough,especially,etc,even,ever,first,followed,following,for,found,from,further,give,given,giving,had,hardly,has,have,having,here,he,he's,her,his,how,however,if,i'm,in,into,is,isn't,it,its,it's,itself,just,kg,km,largely,like,made,mainly,make,may,max,me,might,more,most,mostly,must,my,myself,nearly,neither,no,nor,not,now,obtain,obtained,of,often,on,only,or,other,our,out,over,overall,per,perhaps,possible,previously,quite,rather,really,regarding,resulted,resulting,same,seem,seen,several,she,should,show,showed,shown,shows,significant,significantly,since,so,some,somehow,such,suggest,than,that,the,their,theirs,them,then,there,there's,these,they,this,those,through,thus,to,under,until,up,upon,use,used,using,various,very,was,we,were,what,when,whereas,which,who,while,with,within,without,would,you

b. Lista 2

HI, BEEN, DON'T, ABOUT ,COULDN'T, SIGNIFICANTLI, WHEN ,WITHOUT, THESE, VERI, MOSTLI, HER, WOULD, APPROXIM, DOESN'T, PER, IF, BETWEEN, YOU, THERE'S, GIVE, IN, MADE, MYSELF, IS, THEM, ESPECI, IT, THEN, AM ,AN, EACH, THEI, AR, EVER, AS, BEFOR, ITSELF, AT, EVEN, AMONG, MUST, DOE, OTHER, BE, ISN'T, OUR, I'M, SEEN, OUT, SEEM, RESULT, HOW, INTO, FOUND, SAME, MAI, BY, HAVE, LARG, AFTER, ANOTH, KG, SO, ALWAI, CAN'T, VARIOU, A, KM, MAX, WITHIN, COULD, MORE, SUGGEST, THE, SUCH, THI, REGARD, QUIT,

TO, UNDER, DID, FIRST, MAINLI, BUT, THROUGH, THU, ALMOST, HAD, DO, UPON, WHILE, THAT, EITHER, SHOWN, ETC, SEVER, THAN, ME, SHOULD, FROM, UP, THOSE, US, ALL, OBTAIN, WHICH, GIVEN, SIGNIFIC, WHEREA, LIKE, IT'S, MIGHT, DIDN'T, ONLI, OFTEN, DURE, HARDLI, MY, DONE, BOTH, MOST, SHE, WERE, PERHAP, SINC, WHO, HERE, NO, SOME, RATHER, BECAUS, FOR, SHOW, THEIR, WA, WE, NOR, CAN, NO, AND, NEARLI, NOW, OF, HE'S, POSSIBL, SOMEHOW, ANI, JUST, MAKE, REALLI, ON, OVER, OR, PREVIOUS, AGAIN, OVERAL, ALSO ,ENOUGH, FOLLOW, HOWEV, WITH, WHAT, ALTHOUGH, DUE, NEITHER, THERE, HA, UNTIL, FURTHER, HE

Anexo 2. Lista de formateo y codificación

Se adjuntas las listas de codificación y etiquetado utilizado en la etapa de pre-procesamiento.

a. Caracteres inválidos

(?m)(\n)

Ñ
N
(À|Á|Â|Ã|Ä|Å)
A
(È|É|Ê|Ë)
E
(Ì|Í|Î|Ï)
I
(Ò|Ó|Ô|Õ|Ö)
O
(Ù|Ú|Û|Ü)
U
ñ
n
(à|á|â|â|ã|ä|å)
a
(è|é|ê|ë)
e
(ì|í|î|ï)
i
(ò|ó|ô|õ|ö)

o

(ù | ú | û | ü)

u

Ç

Z

ç

z

b. Caracteres y signos de puntuación

((\{)*[a-zA-Z0-9]+([_\.\,\,][a-zA-Z0-9]+)*(\})*\@([a-zA-Z0-9]+\.)+([a-zA-Z]{2,9})([a-zA-Z]{2,3})?)

@EMAIL

((((jan) | (feb) | (mar) | (apr) | (may) | (jun)(e)* | (jul)(y)* | (aug) | (sep) | (oct) | (nov) | (dec)))

+((r)*(uary) | (ch) | (il) | (ust) | (tember) | (ober) | (ember))*)+(\s | \. | \, | \-

| \\)+(((\d+){1,2}) | ((\d+){1,4}))(\s | \. | \, | \- | \\)*((\d+){4})*

@FECHA

\¿

@INTINI

\?

@INTFIN

\¡

@ADMINI

\!

@ADMFIN

\~

\$?(\d+)((\. | \, | \:)\d+)?(% | , | | \.?$)

@NUM

```
\,
 @COMA
\.
 @PUNTO
\;
 @PUNTOCOMA
\:
 @DOSPUNTOS
\(
 @PBR
\)
 @PCI
\[
 @CBR
\]
 @CCI
\{
 @LLAVEA
\}
 @LLAVEC
\-
 @GM
\_
 @GB
\%
 @PC
\$
 @SP
\/
 @DIDE
\\
```

@DIZQ
\"
@COMILL
\<
@MENQ
\>
@MAYQ
\#
@SG
\'
@APOS
ñ
@n
Ñ
@N

c. Números

\,\$?(\d+)((\.|\,|\:)\d+)?(% |,|, | |\.?$)
,@NUM,
1
@UNO
2
@DOS
3
@TRES
4
@CUATRO
5
@CINCO
6
@SEIS

7
@SIETE
8
@OCHO
9
@NUEVE
0
@CERO

Anexo 3. Lista de reformateo y recodificación

([^a-zA-Z0-9\ _\-\,\:\...\;\(\)\[\]\{\}\.\<\>\/\\\'\"\@\%\$\#]*)*
$1
@COMA
,
@PUNTO
.
@PUNTOCOMA
;
@DOSPUNTOS
:
@PBR
(
@PCI
)
@CBR
[
@CCI
]
@LLAA

{
@LLAC
}
@GM
-
@GB
_
@PC
%
@DIDE
/
@DIZQ
\
@COMILL
"
@MENQ
<
@MAYQ
>
@INTFIN
ʔ
@INTINI
ʕ
@ADMFIN
!
@ADMINI
ˡ
@APOS / @ APO
'
@GRAD
°

```
 @SG
 #
 @SP
 \$
@UNO
1
@DOS
2
@TRES
3
@CUATRO
4
@CINCO
5
@SEIS
6
@SIETE
7
@OCHO
8
@NUEVE
9
@CERO
0
@
```

Anexo 4. Ejemplo de un documento Traing del corpus Inspec

a. Abstract

Perceptual audio coding using adaptive pre- and post-filters and lossless compression

This paper proposes a versatile perceptual audio coding method that achieves high compression ratios and is capable of low encoding/decoding delay. It accommodates a variety of source signals (including both music and speech) with different sampling rates. It is based on separating

irrelevance and redundancy reductions into independent functional units. This contrasts traditional audio coding where both are integrated within the same subband decomposition. The separation allows for the independent optimization of the irrelevance and redundancy reduction units. For both reductions, we rely on adaptive filtering and predictive coding as much as possible to minimize the delay. A psycho-acoustically controlled adaptive linear filter is used for the irrelevance reduction, and the redundancy reduction is carried out by a predictive lossless coding scheme, which is termed weighted cascaded least mean squared (WCLMS) method. Experiments are carried out on a database of moderate size which contains mono-signals of different sampling rates and varying nature (music, speech, or mixed). They show that the proposed WCLMS lossless coder outperforms other competing lossless coders in terms of compression ratios and delay, as applied to the pre-filtered signal. Moreover, a subjective listening test of the combined pre-filter/lossless coder and a state-of-the-art perceptual audio coder (PAC) shows that the new method achieves a comparable compression ratio and audio quality with a lower delay

b. *Frases clave no controladas*

Perceptual audio coding; adaptive pre-filters; adaptive post-filters; lossless compression; high compression ratio; low encoding/decoding delay; source signals; music; sampling rates; redundancy reduction; adaptive filtering; predictive coding; psycho-acoustically controlled adaptive linear filter; irrelevance reduction; predictive lossless coding; weighted cascaded least mean squared; WCLMS lossless coder; subjective listening test; pre-filter/lossless coder; audio quality

c. *Frases clave controladas*

Adaptive filters; adaptive signal processing; audio coding; data compression; delays; filtering theory; hearing; least squares approximations; prediction theory

Anexo 5. Ejemplo de un documento Test del corpus Inspec

a. Abstract

A new graphical user interface for fast construction of computation phantoms and MCNP calculations: application to calibration of in vivo measurement systems

Reports on a new utility for development of computational phantoms for Monte Carlo calculations and data analysis for in vivo measurements of radionuclides deposited in tissues. The individual properties of each worker can be acquired for a rather precise geometric representation of his (her) anatomy, which is particularly important for low energy gamma emitting sources such as thorium, uranium, plutonium and other actinides. The software enables automatic creation of an MCNP input data file based on scanning data. The utility includes segmentation of images obtained with either computed tomography or magnetic resonance imaging by distinguishing tissues according to their signal (brightness) and specification of the source and detector. In addition, a coupling of individual voxels within the tissue is used to reduce the memory demand and to increase the calculational speed. The utility was tested for low energy emitters in plastic and biological tissues as well as for computed tomography and magnetic resonance imaging scanning information.

b. Frases clave no controladas

computational phantoms; Monte Carlo calculations; in vivo measurements; radionuclides; tissues; worker; precise geometric representation; MCNP input data file; scanning data; computed tomography; brightness; graphical user interface; computation phantoms; calibration; in vivo measurement systems; Th; U; Pu; signal; detector; individual voxels; memory demand; calculational speed; plastic; biological

tissues; magnetic resonance imaging scanning information; anatomy; low energy gamma ray emitting sources; actinides; software; automatic creation

c. Frases clave controladas

biological tissues; biomedical MRI; calibration; computerised tomography; graphical user interfaces; lung; Monte Carlo methods; physics computing; radioisotopes

Anexo 6. Resultados para evaluador Performance

6.1 Sistema KEA

Top	Coincidencias	Precisión	Recuerdo	F-Measure
05	378	15.12%	6.71%	9.30%
10	479	9.58%	8.50%	9.01%
15	505	6.73%	8.96%	7.69%

Tabla A6.1 Resultados conjunto no controlado, utilizando en sistema KEA de Witten et al., 1999.

Top	Coincidencias	Precisión	Recuerdo	F-Measure
5	121	4.84%	4.99%	4.91%
10	164	3.28%	6.76%	4.42%
15	177	2.36%	7.29%	3.57%

Tabla A6.2 Resultados conjunto controlado, utilizando en sistema KEA de Witten et al., 1999.

Top	Coincidencias	Precisión	Recuerdo	F-Measure
5	328	13.12%	4.58%	6.79%
10	391	7.82%	5.46%	6.43%
15	406	5.41%	5.67%	5.54%

Tabla A6.3 Resultados conjunto combinado utilizando en sistema KEA de Witten et al., 1999.

En la figura 5.5, se muestran los mejores resultados de cada conjunto evaluado, y se observa una gran diferencia para el conjunto no controladas y controlado; Por otro lado, se esperaba que el conjunto combinado superara o se igualara al conjunto no controlado, sin embargo, no se presentó de esta forma.

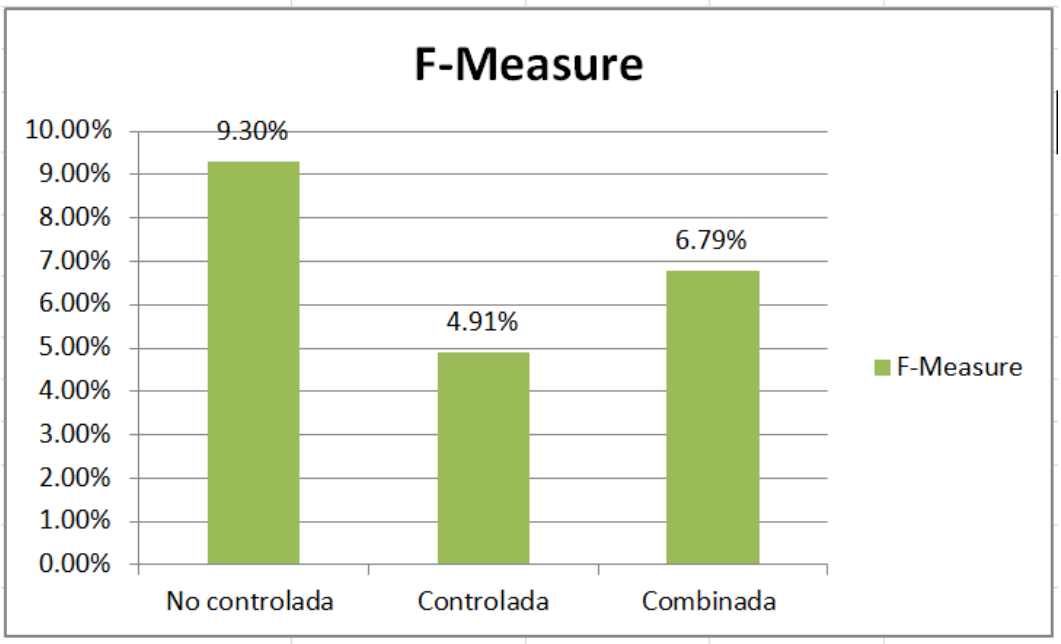

Figura A6.1 Gráfica de los resultados obtenidos con KEA para los tres conjuntos.

6.2 Experimento 1

Parámetros

- Umbral: 1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: No

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	9.28	4.12	5.71	5.86	5.2	5.51	3.91	5.2	4.46
	Com 2	11.32	5.02	6.96	7.16	6.35	6.73	4.77	6.35	5.45
	L1	3.04	1.35	1.87	1.68	1.49	1.58	1.12	1.49	1.28
	L2	8.92	3.96	5.48	4.5	3.99	4.23	3	3.99	3.42
	L3	3.32	1.47	2.04	1.66	1.47	1.56	1.11	1.47	1.26
P	Com 1	9.48	4.21	5.83	5.9	5.24	5.55	3.93	5.24	4.49
	Com 2	11.52	5.11	**7.08**	7.32	6.5	**6.89**	4.88	6.5	**5.57**
	L1	3.12	1.38	1.91	1.68	1.49	1.58	1.12	1.49	1.28
	L2	8.92	3.96	5.48	4.5	3.99	4.23	3	3.99	3.42
	L3	3.32	1.47	2.04	1.66	1.47	1.56	1.11	1.47	1.26
R	Com 1	9.4	4.17	5.78	5.86	5.2	5.51	3.91	5.2	4.46
	Com 2	11.28	5.01	6.94	7.2	6.39	6.77	4.8	6.39	5.48
	L1	3.04	1.35	1.87	1.66	1.47	1.56	1.12	1.49	1.28
	L2	8.92	3.96	5.48	4.5	3.99	4.23	3	3.99	3.42
	L3	3.32	1.47	2.04	1.66	1.47	1.56	1.11	1.47	1.26
F	Com 1	9.44	4.19	5.8	5.86	5.2	5.51	3.91	5.2	4.46
	Com 2	11.32	5.02	6.96	7.22	6.41	6.79	4.81	6.41	5.5
	L1	3.04	1.35	1.87	1.66	1.47	1.56	1.12	1.49	1.28
	L2	8.92	3.96	5.48	4.5	3.99	4.23	3	3.99	3.42
	L3	3.32	1.47	2.04	1.66	1.47	1.56	1.11	1.47	1.26

6.5 Experimento 2

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: No

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	10.28	4.56	6.32	4.52	4.01	4.25	0.24	0.32	0.27
	Com 2	14	6.21	8.6	6.42	5.7	6.04	0.51	0.67	0.58
	Long 1	7.56	3.35	4.64	4.7	4.17	4.42	3.16	4.21	3.61
	Long 2	19.04	8.45	11.71	10.38	9.21	9.76	6.92	9.21	7.9
	Long 3	7.08	3.14	4.35	3.58	3.18	3.37	2.39	3.18	2.73
P	Com 1	11.84	5.25	7.27	5.02	4.46	4.72	0.24	0.32	0.27
	Com 2	16.84	7.47	10.35	7.14	6.34	6.72	0.57	0.76	0.65
	Long 1	8.6	3.82	5.29	4.72	4.19	4.44	3.15	4.19	3.6
	Long 2	19.72	8.75	**12.12**	10.38	**9.21**	9.76	6.92	9.21	**7.9**
	Long 3	7.16	3.18	4.4	3.58	3.18	3.37	2.39	3.18	2.73
R	Com 1	11.08	4.92	6.81	4.7	4.17	4.42	0.24	0.32	0.27
	Com 2	15.44	6.85	9.49	6.78	6.02	6.38	0.56	0.75	0.64
	Long 1	8	3.55	4.92	4.64	4.12	4.36	3.13	4.17	3.58
	Long 2	19.4	8.61	11.93	10.38	9.21	9.76	6.92	9.21	7.9
	Long 3	7.04	3.12	4.32	3.58	3.18	3.37	2.39	3.18	2.73
F	Com 1	11.16	4.95	6.86	4.68	4.15	4.4	0.24	0.32	0.27
	Com 2	15.72	6.98	9.67	6.82	6.05	6.41	0.56	0.75	0.64
	Long 1	7.96	3.53	4.89	4.64	4.12	4.36	3.13	4.17	3.58
	Long 2	19.36	8.59	11.9	10.38	9.21	9.76	6.92	9.21	7.9
	Long 3	7.04	3.12	4.32	3.58	3.18	3.37	2.39	3.18	2.73

6.3 Experimento 3

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo.
- Conjunto: Combinado
- Formato: No

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	10.44	3.64	5.4	7.6	5.29	6.24	5.23	5.46	5.34
	Com 2	11.36	3.96	5.87	7.8	5.43	6.4	5.28	5.52	5.4
	Long 1	9.16	3.19	4.73	4.84	3.37	3.97	3.24	3.39	3.31
	Long 2	13.32	4.64	6.88	6.98	4.86	5.73	4.69	4.9	4.79
	Long 3	9.68	3.37	5	4.84	3.37	3.97	3.23	3.37	3.3
P	Com 1	11.08	3.86	5.73	7.82	5.45	6.42	5.35	5.59	5.47
	Com 2	12.44	4.33	6.42	8.06	5.61	**6.62**	5.47	5.71	**5.59**
	Long 1	9.48	3.3	4.9	4.86	3.39	3.99	3.24	3.39	3.31
	Long 2	13.52	4.71	**6.99**	7	4.88	5.75	4.69	4.9	4.79
	Long 3	9.68	3.37	5	4.84	3.37	3.97	3.23	3.37	3.3
R	Com 1	10.72	3.73	5.53	7.68	5.35	6.31	5.27	5.5	5.38
	Com 2	11.8	4.11	6.1	7.88	5.49	6.47	5.35	5.59	5.47
	Long 1	9.2	3.2	4.75	4.84	3.37	3.97	3.24	3.39	3.31
	Long 2	13.52	4.71	**6.99**	7.02	4.89	5.76	4.69	4.9	4.79
	Long 3	9.68	3.37	5	4.84	3.37	3.97	3.23	3.37	3.3
F	Com 1	10.8	3.76	5.58	7.74	5.39	6.35	5.29	5.53	5.41
	Com 2	11.88	4.14	6.14	7.94	5.53	6.52	5.36	5.6	5.48
	Long 1	9.28	3.23	4.79	4.84	3.37	3.97	3.24	3.39	3.31
	Long 2	13.52	4.71	**6.99**	7.02	4.89	5.76	4.69	4.9	4.79
	Long 3	9.68	3.37	5	4.84	3.37	3.97	3.23	3.37	3.3

6.4 Experimento 4

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: Combinado
- Formato: No

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	14.6	5.09	7.55	15.74	10.98	12.94	11.24	11.76	11.49
	Com 2	19.16	6.68	9.91	18.46	12.88	15.17	13.01	13.62	13.31
	Long 1	7.56	2.64	3.91	4.42	3.08	3.63	2.96	3.1	3.03
	Long 2	23.84	8.32	12.34	13.38	9.34	11	8.99	9.41	9.2
	Long 3	8.68	3.03	4.49	4.38	3.06	3.6	2.92	3.06	2.99
P	Com 1	15.96	5.57	8.26	16.66	11.62	13.69	11.51	12.04	11.77
	Com 2	21.8	7.61	11.28	19.44	13.56	**15.98**	13.72	14.36	**14.03**
	Long 1	8.16	2.85	4.22	4.42	3.08	3.63	2.96	3.1	3.03
	Long 2	25.16	8.78	**13.02**	13.48	9.41	11.08	8.99	9.41	9.2
	Long 3	8.76	3.06	4.54	4.38	3.06	3.6	2.92	3.06	2.99
R	Com 1	14.8	5.16	7.65	15.94	11.12	13.1	11.21	11.74	11.47
	Com 2	19.36	6.75	10.01	18.4	12.84	15.13	13.09	13.7	13.39
	Long 1	7.32	2.55	3.78	4.38	3.06	3.6	2.96	3.1	3.03
	Long 2	24.68	8.61	12.77	13.46	9.39	11.06	8.99	9.41	9.2
	Long 3	8.72	3.04	4.51	4.38	3.06	3.6	2.92	3.06	2.99
F	Com 1	14.72	5.14	7.62	15.9	11.09	13.07	11.2	11.72	11.45
	Com 2	19.68	6.87	10.18	18.68	13.03	15.35	13.28	13.9	13.58
	Long 1	7.24	2.53	3.75	4.38	3.06	3.6	2.96	3.1	3.03
	Long 2	24.64	8.6	12.75	13.48	9.41	11.08	8.99	9.41	9.2
	Long 3	8.76	3.06	4.54	4.38	3.06	3.6	2.92	3.06	2.99

6.5 Experimento 5

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo.
- Conjunto: No controlado
- Formato: Si

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	12.52	5.56	7.7	13.78	12.23	12.96	9.75	12.97	11.13
	Com 2	15.8	7.01	9.71	15.64	13.88	14.71	11.08	14.75	12.65
	Long 1	7.96	3.53	4.89	5.16	4.58	4.85	3.55	4.72	4.05
	Long 2	19.6	8.7	12.05	10.84	9.62	10.19	7.24	9.64	8.27
	Long 3	7.4	3.28	4.55	3.8	3.37	3.57	2.53	3.37	2.89
P	Com 1	14.04	6.23	8.63	14.66	13.01	13.79	10.13	13.49	11.57
	Com 2	18.92	8.4	11.63	16.96	15.05	**15.95**	11.95	15.9	**13.64**
	Long 1	9.08	4.03	5.58	5.3	4.7	4.98	3.53	4.7	4.03
	Long 2	20.56	9.12	**12.64**	10.86	9.64	10.21	7.24	9.64	8.27
	Long 3	7.6	3.37	4.67	3.8	3.37	3.57	2.53	3.37	2.89
R	Com 1	13.12	5.82	8.06	14.34	12.73	13.49	9.87	13.13	11.27
	Com 2	17.08	7.58	10.5	16.16	14.34	15.2	11.39	15.16	13.01
	Long 1	8.28	3.67	5.09	5.18	4.6	4.87	3.53	4.7	4.03
	Long 2	20.56	9.12	**12.64**	10.86	9.64	10.21	7.24	9.64	8.27
	Long 3	7.48	3.32	4.6	3.8	3.37	3.57	2.53	3.37	2.89
F	Com 1	13.12	5.82	8.06	14.34	12.73	13.49	9.88	13.15	11.28
	Com 2	17.24	7.65	10.6	16.24	14.41	15.27	11.4	15.18	13.02
	Long 1	8.32	3.69	5.11	5.18	4.6	4.87	3.53	4.7	4.03
	Long 2	20.56	9.12	**12.64**	10.86	9.64	10.21	7.24	9.64	8.27
	Long 3	7.52	3.34	4.63	3.8	3.37	3.57	2.53	3.37	2.89

6.6 Experimento 6

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo
- Conjunto: No controlado
- Formato: Si

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	10.92	5.56	7.37	13.74	13.98	13.86	9.88	15.08	11.94
	Com 2	15.32	7.8	10.34	16.38	16.67	16.52	11.76	17.95	14.21
	Long 1	5.88	2.99	3.96	3.7	3.77	3.73	2.52	3.85	3.05
	Long 2	21.84	11.11	14.73	12.12	12.33	12.22	8.08	12.33	9.76
	Long 3	9.36	4.76	6.31	4.72	4.8	4.76	3.15	4.8	3.8
P	Com 1	12.36	6.29	8.34	14.66	14.92	14.79	10.15	15.49	12.26
	Com 2	18.6	9.46	12.54	17.48	17.79	**17.63**	12.49	19.07	**15.09**
	Long 1	6.8	3.46	4.59	3.74	3.81	3.77	2.52	3.85	3.05
	Long 2	22.84	11.62	**15.4**	12.1	12.31	12.2	8.08	12.33	9.76
	Long 3	9.4	4.78	6.34	4.72	4.8	4.76	3.15	4.8	3.8
R	Com 1	11.32	5.76	7.64	13.82	14.06	13.94	9.75	14.88	11.78
	Com 2	16.76	8.53	11.31	16.52	16.81	16.66	11.83	18.05	14.29
	Long 1	5.56	2.83	3.75	3.68	3.75	3.71	2.52	3.85	3.05
	Long 2	22.32	11.36	15.06	12.08	12.29	12.18	8.08	12.33	9.76
	Long 3	9.4	4.78	6.34	4.72	4.8	4.76	3.15	4.8	3.8
F	Com 1	11.36	5.78	7.66	13.86	14.11	13.98	9.77	14.92	11.81
	Com 2	16.84	8.57	11.36	16.58	16.87	**16.72**	11.87	18.12	14.34
	Long 1	5.64	2.87	3.8	3.7	3.77	3.73	2.52	3.85	3.05
	Long 2	22.36	11.38	15.08	12.08	12.29	12.18	8.08	12.33	9.76
	Long 3	9.4	4.78	6.34	4.72	4.8	4.76	3.15	4.8	3.8

6.7 Experimento 7

Parámetros

- Umbral 0.1%
- Stopwords: Lista 2 (Anexo 1.b.)
- Stemming de Porter: Primer algoritmo
- Conjunto: No controlado
- Formato: Si

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	11.2	4.97	6.88	13.14	11.66	12.36	9.39	12.5	10.72
	Com 2	15.28	6.78	9.39	14.86	13.19	13.98	10.65	14.18	12.16
	Long 1	7.84	3.48	4.82	5.28	4.69	4.97	3.61	4.81	4.12
	Long 2	18.44	8.18	11.33	10.38	9.21	9.76	6.93	9.23	7.92
	Long 3	6.24	2.77	3.84	3.24	2.88	3.05	2.16	2.88	2.47
P	Com 1	12.28	5.45	7.55	13.78	12.23	12.96	9.61	12.8	10.98
	Com 2	16.84	7.47	10.35	15.36	13.63	**14.44**	11.01	14.66	**12.58**
	Long 1	8.48	3.76	5.21	5.28	4.69	4.97	3.61	4.81	4.12
	Long 2	19.28	8.56	**11.86**	10.4	9.23	9.78	6.93	9.23	7.92
	Long 3	6.44	2.86	3.96	3.24	2.88	3.05	2.16	2.88	2.47
R	Com 1	12.4	5.5	7.62	13.76	12.21	12.94	9.61	12.8	10.98
	Com 2	16.72	7.42	10.28	15.34	13.61	14.42	10.99	14.63	12.55
	Long 1	8.48	3.76	5.21	5.28	4.69	4.97	3.61	4.81	4.12
	Long 2	19.28	8.56	**11.86**	10.4	9.23	9.78	6.93	9.23	7.92
	Long 3	6.44	2.86	3.96	3.24	2.88	3.05	2.16	2.88	2.47
F	Com 1	12.52	5.56	7.7	13.78	12.23	12.96	9.65	12.85	11.02
	Com 2	16.76	7.44	10.31	15.34	13.61	14.42	10.99	14.63	12.55
	Long 1	8.6	3.82	5.29	5.28	4.69	4.97	3.61	4.81	4.12
	Long 2	19.16	8.5	11.78	10.4	9.23	9.78	6.93	9.23	7.92
	Long 3	6.44	2.86	3.96	3.24	2.88	3.05	2.16	2.88	2.47

6.8 Experimento 8

Parámetros

- Umbral 0.1%
- Stopwords: Lista 1 (Anexo 1.a.)
- Stemming de Porter: Segundo algoritmo
- Conjunto: No controlado
- Formato: Si

Pesado	Conjunto	Candidatas 5			Candidatas 10			Candidatas 15		
		P	R	F-M	P	R	F-M	P	R	F-M
B	Com 1	8.56	4.36	5.78	13.16	13.39	13.27	9.63	14.7	11.64
	Com 2	14.16	7.21	9.55	15.5	15.77	15.63	11.48	17.52	13.87
	Long 1	5.32	2.71	3.59	3.56	3.62	3.59	2.48	3.79	3
	Long 2	21.16	10.77	14.27	12.08	12.29	12.18	8.09	12.35	9.78
	Long 3	8.88	4.52	5.99	4.62	4.7	4.66	3.08	4.7	3.72
P	Com 1	10.2	5.19	6.88	14.24	14.49	14.36	10	15.27	12.09
	Com 2	16.84	8.57	11.36	16.84	17.14	**16.99**	12.33	18.83	**14.9**
	Long 1	6.4	3.26	4.32	3.66	3.72	3.69	2.48	3.79	3
	Long 2	22.48	11.44	**15.16**	12.12	12.33	12.22	8.09	12.35	9.78
	Long 3	9.08	4.62	6.12	4.62	4.7	4.66	3.08	4.7	3.72
R	Com 1	9.76	4.97	6.59	13.94	14.19	14.06	9.88	15.08	11.94
	Com 2	15.44	7.86	10.42	16.36	16.65	16.5	12.04	18.38	14.55
	Long 1	6	3.05	4.04	3.6	3.66	3.63	2.48	3.79	3
	Long 2	22.12	11.26	14.92	12.12	12.33	12.22	8.09	12.35	9.78
	Long 3	9.12	4.64	6.15	4.62	4.7	4.66	3.08	4.7	3.72
F	Com 1	9.8	4.99	6.61	14.02	14.27	14.14	9.89	15.1	11.95
	Com 2	15.64	7.96	10.55	16.52	16.81	16.66	12.07	18.42	14.58
	Long 1	6.04	3.07	4.07	3.6	3.66	3.63	2.48	3.79	3
	Long 2	22.24	11.32	15	12.12	12.33	12.22	8.09	12.35	9.78
	Long 3	9.12	4.64	6.15	4.62	4.7	4.66	3.08	4.7	3.72