UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Generación Automática de Resúmenes Independientes del Lenguaje"

Tesis
Para Obtener el Grado de
Maestra en Ciencias de la Computación

Que Presenta
Ing. Griselda Areli Matias Mendoza

Tutor Académico:
Dra. Yulia Nikolaevna Ledeneva

Tutores Adjuntos:
Dr. René Arnulfo García Hernández
Dr. Grigori Sidorov

TIANGUISTENCO, MÉXICO ENERO 2016

Dedico este trabajo

A los dos amores de mi vida

Artemio Becerril García

Y

Moisés Becerril Matias

Son lo mejor que me ha pasado

Agradecimientos

Dios, gracias por darme la oportunidad de ser una mejor persona, por darme la vida y la sabiduría para poder realizar este sueño.

Artemio, gracias por tu apoyo incondicional y por siempre impulsarme a ser mejor. Te amo

Mis padres, gracias por apoyarme no solo económicamente sino también con amor, comprensión, y sobre todo por creer en mi ante las adversidades.

Doctores Yulia y René, gracias por creer en mí a pesar de mis errores y fallas, son un verdadero ejemplo a seguir.

Maestro Rafa, gracias por sus regaños, consejos y críticas constructivas, que me permitieron mejorar seminario a seminario.

Resumen

En la actualidad la información en formato digital crece de manera exponencial y ante ello surgen diversas problemáticas, como la sobrecarga de información, redundancia de información, pérdida de información, entre otras. Este tipo de problemas puede ocasionar en los usuarios deficiencia en su trabajo, al no tener el tiempo disponible necesario, para procesar toda la información, ante esto surge la importante necesidad de contar con métodos que permitan la generación automática de resúmenes. Pero además de contar con un método que nos permite generar resúmenes, sería ideal que los métodos generaran resúmenes en cualquier lenguaje, principalmente en el lenguaje que domina el usuario (en nuestro caso el español).

Un método de generación automática de resúmenes independientes del lenguaje, trata de contrarrestar los efectos negativos de la sobrecarga de información, además de que permite generar un resumen, independientemente del lenguaje en el que se encuentre el texto original. Según (Ledeneva, 2008) un resumen es un texto corto que transmite la información más importante de un documento de origen.

Actualmente existen métodos del estado del arte que dicen ser independientes del lenguaje, pero solo prueban en el lenguaje inglés. Existen otros que son independientes del lenguaje y prueban más de una colección de documentos, pero no en español. Entre los métodos del estado del arte que dicen ser independientes del lenguaje está el propuesto por (Matias, 2013), el cual obtienen buenos resultados para el lenguaje inglés y puede trabajar con otros lenguajes. Entonces con referencia a los resultados que se obtienen con el método de (Matias, 2013), en este trabajo se propone el método en los lenguajes: inglés, portugués y español. Además se ajustaron los parámetros de las etapas: pre-procesamiento, modelo de texto, importancia de las oraciones, función de aptitud y el operador de selección, para tratar de mejorar la calidad de los resúmenes.

Las colecciones de documentos utilizadas en este trabajo son, para inglés la colección DUC2002, para portugués la colección TeMário y para el lenguaje español TER. La colección TER es una aportación de este trabajo, la cual es una colección de noticias de un periódico mexicano (La crónica) especialmente para el uso de resúmenes. Los resúmenes resultantes son evaluados con la herramienta ROUGE la cual permite comparar los resúmenes generados a partir del método con los resúmenes generados por un humano.

Los resultados obtenidos de los experimentos con cada una de las colecciones se comparan con los resultados obtenidos con los resúmenes generados con las herramientas comerciales

y otros métodos del estado del arte. Los resultados obtenidos con el método propuesto en todos los lenguajes superan tanto a las herramientas comerciales como a los métodos del estado del arte.

Contenido

Página

Lista de Figuras

Página

Lista de tablas

CAPÍTULO 1

Introducción

En la actualidad, la información disponible de manera digital crece de manera exponencial. Según un estudio realizado por el Instituto Nacional de Estadística y Geografía (INEGI) el uso de internet ha ido en incremento. Actualmente el 47% de la población mexicana son usuarios activos de internet (INEGI, 2015).

En la Figura 1, podemos observar cuáles son los principales usos que se le da al internet.

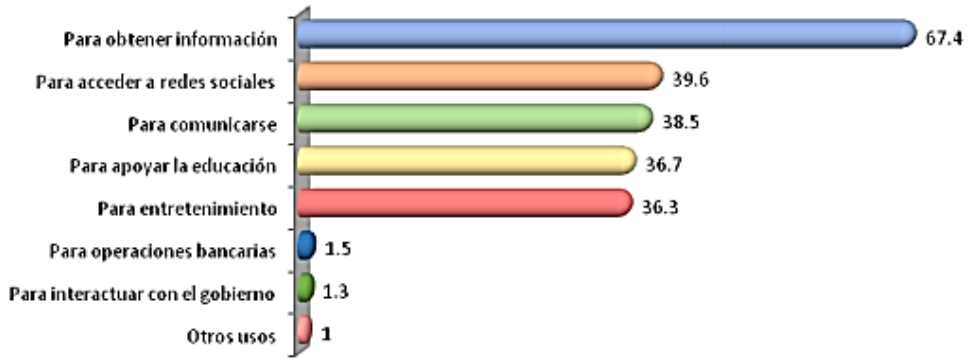

Figura 1. Usuarios de internet por tipo de uso en el año 2014. Las categorías no son excluyentes, por lo que la suma de las proporciones no es el 100 por ciento. Fuente: (INEGI, 2015).

Como se puede observar en la gráfica anterior, el principal uso de internet es la obtención de información, esto quiere decir que la información se ha convertido en un bien necesario y que cada día es más difícil para los usuarios poder acceder a ella y poder procesar la gran cantidad que existe.

La información en formato digital crece de manera rápida, pero el tiempo disponible para procesarla sigue siendo un recurso valioso y limitado. Estudios como el de (Klingberg, 2009) han subrayado cómo el exceso de información puede hacer nuestro trabajo menos productivo, lo que ocasiona ansiedad y estrés. Por ello ha surgido la importante necesidad de contar con métodos que permitan la generación automática de resúmenes.

La generación automática de resúmenes trata en cierta manera de contrarrestar los efectos negativos de la sobre carga de información sobre la capacidad de los usuarios para obtener aquella que realmente les interesa y transformarla en conocimiento (Plaza, 2010).

Según Ladda Saunmali (Suanmali, 2011) el objetivo del resumen de texto es presentar la información más importante en una versión más corta del texto original, manteniendo su contenido principal y ayudando al usuario a comprender rápidamente el gran volumen de información.

Actualmente se puede ver el uso de los resúmenes en diferentes áreas. Se emplean por ejemplo para videos (Yahiaoui, 2003), (Mei, 2015), noticias de periódicos (García, 2013) (Mihalcea, 2005), artículos científicos (Qazvinian, 2013) y actualmente para resumir información producida en las redes sociales como twitter (Nichols, 2012), (Yang, 2011), donde la información cambia rápidamente y se requiere de tecnologías que permitan tener acceso a la información en tiempo real.

Según Alfonseca, Berker, Da Cunha Faneg entre otros, (Alfonseca, 2003), (Berker, 2011), (Da Cunha, 2008) los resúmenes se clasifican según su estrategia de condensación en resúmenes abstractivos y extractivos. Los resúmenes abstractivos son aquellos resúmenes generados a partir de la comprensión del documento y describen el contenido con palabras u oraciones que en algunas ocasiones no encontramos en el texto original (Saunmali, 2011), (Qazvinian, 2008), (Da Cunha, 2008), (Ledeneva, 2008), (Montiel, 2009), (Plaza, 2010), (García, 2009). En cambio, los resúmenes extractivos son generados a partir de la selección de frases clave, oraciones o párrafos considerados importantes del texto original; por lo que no requieren de la comprensión del documento (Saunmali, 2011), (Qazvinian, 2008), (Da Cunha, 2008), (Ledeneva, 2008), (Montiel, 2009), (Plaza, 2010), (García, 2009).

Regularmente cuando realizamos una búsqueda de información en internet, lo hacemos en el lenguaje que dominamos (español) y esperamos que la información obtenida se encuentre en esté. Sin embargo, no siempre es así, por lo regular debemos tener más de una opción en los lenguajes que podemos dominar. Actualmente uno de los lenguajes más utilizados es el inglés.

Sería ideal contar con herramientas generadoras de resúmenes que permitan obtener un resumen de un texto en varios lenguajes, que sean independientes del lenguaje y que además los resúmenes generados sean de calidad.

Un método de generación automática de resúmenes independientes del lenguaje según (Plaza, 2010) consiste en tener un método que teniendo como entrada un texto base en cierto lenguaje genera el resumen en este lenguaje y posteriormente se traduce a diferentes lenguajes. Sin embargo, otros autores como (Patel, 2007), (Mihalcea, 2005), (Wang, 2013) y (Last, 2010) dicen que un método generador automático de resúmenes independiente del lenguaje consiste en que teniendo una colección de documentos multilingües (colección de documentos escritos en varios idiomas) se genere el resumen mediante una única herramienta. Un requisito importante para cualquier método que trabaje independiente del lenguaje, es que demuestre un funcionamiento igual en diversos idiomas, sin adaptaciones especiales como modificaciones al algoritmo o datos adicionales de cada lengua.

Cuando se habla de generar resúmenes para varios lenguajes se complica, ya que las características de cada uno de ellos es diferente, sin embargo si se utilizan métodos de tipo estadístico (extractivos) se pueden simplificar los problemas, aunque en algunas ocasiones se sacrifica calidad.

Existen métodos del estado del arte que dicen ser independientes del lenguaje, pero sólo prueban con una sola colección de documentos (Ledeneva, 2008), (Ledeneva, 2008a), (García, 2008), (Montiel, 2009), (Matias, 2013), (García, 2013).

Existen otros métodos del estado del arte que dicen ser independientes del lenguaje y en realidad prueban con más de una colección de documentos entre ellas están (Patel, 2007), (Mihalcea, 2005), (Last, 2010), (Saggion, 2011), entre otros. Sin embargo, las colecciones que utilizan para realizar sus pruebas en su mayoría son colecciones propias que no se encuentran disponibles. De los métodos que son independientes del lenguaje ninguno prueba con alguna colección en español.

En la Figura 2, se muestran los resultados de los métodos del estado del arte que dicen ser independiente del lenguaje, además de las heurísticas baseline. Como se puede observar los métodos propuestos por (Matias, 2013) y (García, 2013) son lo que obtienen los mejores resultados.

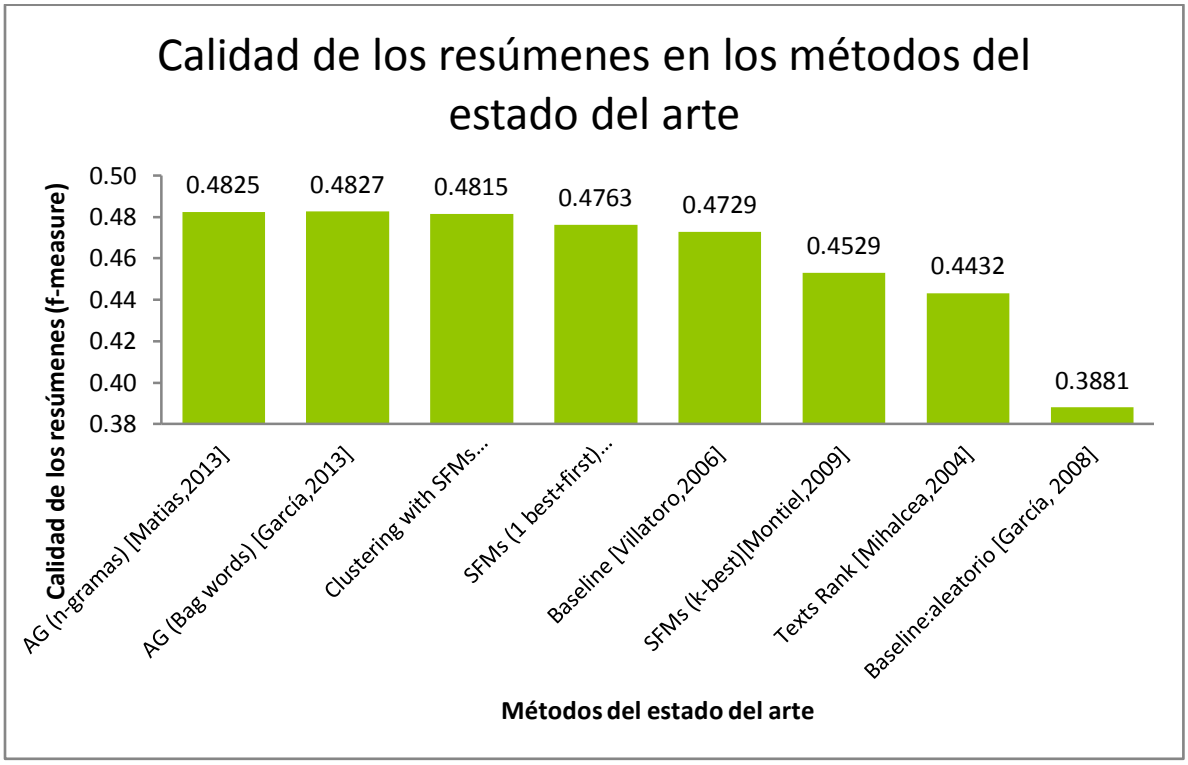

Figura 2. Calidad de los resúmenes en los métodos del estado del arte que dicen ser independientes del lenguaje.

Se tiene disponible el método de (Matias, 2013), el cual como se puede observar en la Figura 2 obtiene buenos resultados para la generación de resúmenes pero solo fue probado en inglés, aunque puede trabajar con otros idiomas. El método de (Matias, 2013)] es un método para la generación de resúmenes de textos extractivos para un solo documento.

Entonces con referencia a los resultados anteriores y considerando que se tiene acceso al método de (Matias, 2013) surge el siguiente problema de investigación.

1.1 Planteamiento del Problema

Conocer que tan independiente del lenguaje es el método de (Matias, 2013) especialmente en el lenguaje español y cómo mejora la calidad de sus resúmenes para los diferentes lenguajes?

1.2 Objetivos

1.2.1 Objetivo general

Probar el método de (Matias, 2013) con colecciones en diferentes lenguajes y ajustar sus parámetros para conocer el desempeño con diversos lenguajes.

1.2.2 Objetivos específicos

- Probar el método de (Matias, 2013) con colecciones en diferentes lenguajes.
- Ajustar sus parámetros:
    - Aplicar pre-procesamiento
    - Extraer el modelo de texto
    - Analizar la importancia de las oraciones
    - Ajustar la función de aptitud
    - Modificar el operador de selección
- Realizar las pruebas con los diferentes parámetros
- Evaluar cada una de las pruebas
- Analizar los resultados obtenidos

1.3 Hipótesis

Si se realizan pruebas con colecciones en diferentes lenguajes y se ajustan los parámetros del método se pueden obtener mejores resultados y comprobar su independencia del lenguaje.

1.4 Delimitación del problema

- El método no estará enfocado a responder en un tiempo determinado, sino a la calidad del resumen.
- Se generarán resúmenes con los documentos en tres lenguajes (español, inglés y portugués).
- El dominio para el que se generaran los resúmenes es noticias.

- No se toma en cuenta el tiempo de optimización de los procesos, se busca que el proceso sea rápido, pero por el momento no es una condicionante.

1.5 Motivación y posibles aplicaciones

Uno de los principales razones para realizar este trabajo sobre la generación de resúmenes se remonta a los trabajos previos del estado del arte (Matias, 2013), (García, 2009), (Ledeneva, 2011), (García, 2013), (Ledeneva, 2008). Principalmente en el trabajo de (Matias, 2013), en el cual se obtienen buenos resultados para la generación de resúmenes. Saber que se tenía un método con buenos resultados con el motivo de demostrar que se podían obtener resultados de calidad en el lenguaje español. Y para demostrar su independencia del lenguaje se debía probar con algún otro lenguaje, por esto se decidió probar con la colección de documentos en portugués.

Por otro lado, se sabe que el crecimiento de la información ha provocado que surjan diversos problemas para los usuarios de internet principalmente, ya que no es fácil acceder a la gran cantidad de información o que ésta sea redundante ya que no se cuente con una herramienta que le permita saber de qué trata un documento sin tener que leerlo completamente (tener un resumen del texto). Es por ello que el contar con una herramienta que le permita al usuario no solamente poder tener un resumen del texto, sino que pueda tenerlo en el lenguaje que domina es la herramienta ideal.

1.6 Organización de la tesis

En el presente capítulo, se da una introducción a la importancia de generar resúmenes de forma automática, se habla acerca del crecimiento del uso del internet y de los diferentes dominios en donde se emplean los resúmenes. También se presentan algunos resultados que se tiene en los métodos del estado del arte que dicen ser independientes del lenguaje. Finalmente se presenta el problema, los objetivos generales y específicos, la hipótesis, la delimitación del problema, la motivación y posibles aplicaciones de esta tesis.

El resto del documento está organizado de la siguiente manera:

En el capítulo 2, se definen los conceptos utilizados en este trabajo. Se presenta el concepto de independencia del lenguaje, modelado de texto, el cual a su vez está constituido por la definición de pre-procesamiento, análisis léxico, eliminación de *stopwords, stemming*.y los tipos de modelo de texto (bolsa de palabras y n-gramas). También se menciona lo que son los algoritmos genéticos, su esquema y sus parámetros. Finalmente se mencionan la medida de evaluación más utilizada para la generación de resúmenes.

En el capítulo 3, se presenta el estado del arte para la generación de resúmenes automáticos. Este capítulo está dividido por lenguajes, se muestra el estado del arte para el lenguaje inglés, portugués y español. También se describen las herramientas comerciales instalables y en línea para la generación automática de resúmenes. Finalmente se muestran algunas tablas donde se realiza un análisis de los métodos del estado del arte independientes del lenguaje, así como de las características consideradas para la generación de resúmenes.

En el capítulo 4, se describe el método propuesto en este trabajo. Se detalla cada una de las etapas que se siguen.

En el capítulo 5, se describen los corpus que se utilizaron para generar los resúmenes, también se menciona la herramienta utilizada para la evaluación. Principalmente en este capítulo se muestran las gráficas con los resultados de los experimentos para los lenguajes, inglés, portugués y español.

Finalmente en el capítulo 6, se describen las conclusiones obtenidas del análisis de los resultados y las derivadas en el transcurso de la elaboración del trabajo. Además se presentan las aportaciones derivadas de esta investigación, así como el trabajo futuro

CAPÍTULO 2

Marco Teórico

En este capítulo, se presentan los conceptos principales utilizados en este trabajo. Se da la definición que se utilizará para independencia del lenguaje. Otra de las definiciones importantes para este trabajo es la importancia de las oraciones con respecto a un texto dado. También se describe en que consiste el pre-procesamiento y el modelado de texto, parámetros que se utilizarán en el método. Se presentan los conceptos básicos sobre algoritmos genéticos, los cuales son la base del método con el que se va a trabajar. Finalmente, se presenta la herramienta de evaluación y los corpus utilizados en las pruebas.

2.1 Independencia del lenguaje

Cuando se habla de generación automática de resúmenes para varios idiomas se complica, por las características de cada lenguaje pueden variar. Sin embargo, antes de considerar las características o los métodos que se utilizan en los métodos para la generación de resúmenes independientes del lenguaje se debe tomar en cuenta la definición que se le da en el estado del arte a independencia del lenguaje, multilingüe o plurilingüe.

Algunos sistemas como SUMMARIST (Hovy, 1999) generan resúmenes multilingüe extrayendo las oraciones de un documento en distintos idiomas, y traduciendo el resumen a distintos idiomas. Otros sistemas como NewsBlaster (Blair-Goldensohm, 2004) o los presentados por (Bouayad-Agha, 2009) y (Saggion, 2008), realizan la traducción antes de realizar la extracción de las oraciones; es decir, realizan un paso previo para traducir a un idioma común todos los

documentos a resumir. Sin embargo, la mayoría de estos sistemas presentan un resultado pobre en cuanto a legibilidad y a calidad gramatical que refiere, debido principalmente al software utilizado para realizar la traducción automática.

Según (Patel, 2007), (Mihalcea, 2005), (Wang, 2013) y (Last, 2010), multilenguaje, independiente del lenguaje y plurilingüe tienen el mismo significado. Para estos autores independencia del lenguaje consiste, en que teniendo como entrada un conjunto de documentos en diferentes lenguajes se puede generar el resumen en el lenguaje de entrada del texto para cada uno de ellos, con un método que no se basan en ningún análisis morfológico del texto.

Cuando se habla de generar resúmenes para varios idiomas se complica, porque las características de cada uno de ellos son diferentes. Sin embargo, si se utilizan métodos de tipo estadístico se puede simplificar el problema, aunque en algunas ocasiones se sacrifica calidad.

Utilizar métodos estadísticos, permite tener método más robustos y se adaptan fácilmente a diferentes idiomas. Lo que requiere es solo tener la lista de palabras vacías y un algoritmo de *stemming* para el idioma que se desea realizar el resumen.

Entonces se pudiera decir que un método es independiente del lenguaje cuando no utiliza ninguna ayuda como diccionario, entrenamiento, etc. Y que es poco independiente si solo utiliza por ejemplo la eliminación de *stopwords* (ocupa la lista de estas palabras en el idioma que se trabaja) o *stemming*, sin embargo se vuelve dependiente si utiliza más recursos del idioma en el que se trabaja como son: colocaciones, desambiguación, diccionarios.

2.2 Modelado del texto

Según Romyna Montiel (Montiel, 2009) los modelos de representación de textos son una técnica que se basa en la extracción de los términos de un texto o documento. El modelado de texto consiste en seleccionar los términos que serán extraídos y convertirlos en un patrón que pueda ser analizado posteriormente. La diferencia entre modelos es el tipo de término que se extrae del documento. Para poder utilizar un modelo de texto primero se debe pasar por la etapa de pre-procesamiento.

La tabla 1 muestra un documento separado en cinco oraciones, el cual se utiliza en este capítulo para analizar los modelos de texto.

Tabla 1. Lista de oraciones del documento de ejemplo.

ORACIONES
Oración1: el gobierno de Egipto protege las pirámides
Oración2: las pirámides de Egipto son un patrimonio cultural
Oración3: las pirámides fueron construidas por los faraones
Oración4: un buen gobierno protege su patrimonio cultural
Oración5: las pirámides de Egipto fueron tumbas para los faraones

2.2.1 Pre-procesamiento

El pre-procesamiento es la etapa donde se procesa el texto de entrada para producir el texto de salida que se utilizarán en el programa generador de resúmenes, el texto se transforma a una forma estructurada o semiestructurada de su contenido.

Al pre-procesar los textos se pueden obtener representaciones sencillas que faciliten el análisis del texto. Para poder pre-procesar el texto se pueden utilizar las siguientes herramientas.

2.2.1.1 Análisis léxico

Análisis léxico, es el proceso de convertir un conjunto de caracteres en un conjunto de palabras (Lo Cen, 2012). Uno de los principales objetivos del análisis léxico es identificar las palabras y el reconocimiento de espacios, dígitos, manejo de cadenas HTML, manejo de mayúsculas y minúsculas, signos de puntuación.

- **Espacios.** Generalmente reconocidos como separadores de palabras.
- **Dígitos.** Usualmente no son vistos como buenos términos índices, por lo que no son incluidos.
- **Manejo de etiquetas HTML.** Generalmente son eliminadas, ya que no proporcionan información al texto.
- **Guiones.** Separar todas las palabras divididas por guiones ayuda a un manejo uniforme de los términos. Sin embargo, el separar en algunos casos pueden conducir a la pérdida del significado de la palabra. Si son utilizados para dividir en sílabas una palabra al final de una línea, este se debe de volver a unir. Su uso depende mucho del contexto en el que se esté trabajando.
- **Manejo de mayúsculas/minúsculas.** La combinación de estas no suelen ser un problema para la identificación de términos que formarán un índice, por lo que el analizador suele pasar todo a un mismo tipo de letra.

- **Signos de puntuación.** Normalmente se manejan como separadores de palabras por lo que pueden eliminarse.

2.2.1.2 Eliminación de stopwords

Eliminación de palabras muy repetitivas que no proporcionan información relevante, palabras que aparecen en más del 80% del documento no son consideradas y se llaman *stopwords* (Montiel, 2009). Generalmente las candidatas son los artículos, preposiciones, conjunciones y pronombres. Para cada idioma se puede tener una lista de estas palabras consideradas vacías.

2.2.1.3 Stemming

Stemming, es un método para reducir una palabra a su raíz (en inglés) a stem o lema, con el objetivo de eliminar prefijos, sufijos y de permitir la recuperación de los documentos que tienen variaciones sintácticas de los términos que se están recuperando (Ramírez, 2007).

Los primeros algoritmos de *stemming* se desarrollaron para el idioma inglés, uno de los principales algoritmos para *stemming* es el algoritmo de Porter (Porter, 1980). Sin embargo esta técnica puede ser adaptada a diferentes idiomas. Estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común.

A continuación, en la Tabla 2, se muestra el texto resultante al aplicarle el pre-procesamiento de análisis léxico (separar por palabras, manejo de mayúsculas), eliminar las palabras repetitivas y separar las palabras por comas.

Tabla 2. Representación de la etapa de pre-procesamiento

ORACIONES
Oración1: GOBIERNO,EGIPTO,PROTEGE,PIRAMIDE
Oración2: PIRAMIDE,EGIPTO,PATRIMONIO,CULTURAL
Oración3: PIRAMIDE,FUERON,CONSTRUIDAS,FARAONES
Oración4: BUEN,GOBIERNO,PROTEGE,PATRIMONIO,CULTURAL
Oración5: PIRAMIDE,EGIPTO,TUMBAS,FARAONES

Después de la etapa de pre-procesamiento puede ser utilizado un modelo de texto. A continuación se describen algunos de los modelos más comunes.

2.2.2 Tipos de modelo de texto

2.2.2.1 Modelo bolsa de palabras

Dado un documento, el primer paso consiste en extraer todas las palabras diferentes de dicho documento. A este conjunto de palabras se le conoce como "bolsa de palabras" debido a que éstas no se encuentran ordenadas. Posteriormente, cada oración del documento es indexada, es decir representada por un vector de términos, donde cada término corresponde a una palabra de la bolsa de palabras (García, 2008a). A continuación en la Tabla 3 se muestra un ejemplo tomando como base el texto sin pre-procesamiento.

Tabla 3. Representación del modelo bolsa de palabras

BOLSA DE PALABRAS				
Oración1	Oración2	Oración3	Oración4	Oración5
el	las	las	un	las
Gobierno	pirámides	pirámides	buen	pirámides
de	de	fueron	gobierno	de
Egipto	Egipto	construidas	protege	Egipto
protege	son	por	su	fueron
las	un	los	patrimonio	tumbas
pirámides	patrimonio	faraones	cultural	para
	cultural			los
				faraones

2.2.2.2 modelo n-gramas

Se llama n-grama a una subsecuencia de n elementos consecutivos en una secuencia dada. Se pueden construir n-gramas con base en distintos tipos de elementos como, por ejemplo, fonemas, silabas, letras o palabras. Cabe mencionar, que los 1-gramas también se llaman unigramas; los 2-gramas también se llaman bigramas; los 3-gramas también se llaman trigramas. Sin embargo, se pueden definir valores mayores de n dependiendo el problema que se quiera resolver.

Su definición es la siguiente:
Sea una secuencia S de elementos ordenados $s_1 s_2 s_3 \dots s_k$ se denomina n-grama a cualquier subsecuencia $A = s_{i+1} s_{i+2} \dots s_{i+n}$ donde i es un valor entre 0 y $|s| - n$ para garantizar que la longitud de A sea siempre n o lo que es lo mismo $|A| = n; n > 1$ (EcuRed, 2013).

Por ejemplo, dado el texto del ejemplo anterior, si se establecen como elementos a las palabras del texto, sus bigramas se muestran en la Tabla 4, tomando como base el texto sin pre-procesamiento:

Tabla 4. Representación del modelo de n-gramas por oración

		N-GRAMA		
Oración1	Oración2	Oración3	Oración4	Oración5
el gobierno	las pirámides	las pirámides	un buen	las pirámides
gobierno de	pirámides de	Pirámides fueron	buen gobierno	pirámides de
de Egipto	de Egipto	fueron construidas	gobierno protege	de Egipto
Egipto protege	Egipto son	construidas por	protege su	Egipto fueron
protege las	son un	por los	su patrimonio	fueron tumbas
las pirámides	un patrimonio	los faraones	patrimonio cultural	tumbas para
	patrimonio cultural			para los
				los faraones

2.3 Algoritmos genéticos

Los algoritmos genéticos son una técnica de resolución de problemas de búsqueda y optimización inspirada en la teoría de la evolución de las especies y la selección natural. Estos algoritmos reúnen características de búsqueda aleatoria con características de búsqueda dirigida que provienen del mecanismo de selección de los individuos más adaptados (Araujo, 2009).

A cada individuo se le asigna una puntuación en relación a sus características. En la naturaleza esto sería equivalente a la capacidad que tiene un individuo para competir con otros y sobrevivir; ya que en el siguiente paso se selecciona de acuerdo al valor de su aptitud, para pasar a una siguiente generación o desaparecer. Los individuos más aptos pueden reproducirse cruzando su material genético con otro individuo que ha sido seleccionado de la misma forma. Este proceso permite la generación de nuevos individuos (nuevas posibles soluciones), los cuales seguirán siendo evaluados según sus características. La cruza permitirá pasar a la descendencia las características que permitan a un individuo hijo mejorar. Sin embargo, a lo largo de algunas generaciones puede llegarse a una estabilización y no conseguir tener mejores individuos, por lo cual se usa otro operador genético llamado mutación, el cual permite cambiar de manera aleatoria alguna de las características del individuo, lo que permitirá tener una mayor diversidad genética y por ende más posibilidades de obtener mejores soluciones.

Los algoritmos genéticos cuentan con características que los hacen mejores que los métodos puramente aleatorios. Es importante mencionar que los algoritmos genéticos no garantizan una solución exacta, sino una aproximación que dependerá de los recursos dedicados a la búsqueda, es decir, tiempo y memoria, a parte claro está el diseño adecuado de los componentes del algoritmo computacional, proporcionando una solución aproximada que en muchos casos es suficiente para las necesidades del usuario.

2.3.1 Esquema general de un algoritmo genético

Los algoritmos genéticos procesan simultáneamente, no una solución al problema, sino todo un conjunto de ellas (Araujo, 2009). Los algoritmos genéticos trabajan con alguna forma de representación de soluciones llamadas individuos y el conjunto de ellos forman una población, que es con la que trabaja el algoritmo genético. La población se va modificando a lo largo de las iteraciones del algoritmo que se denominan generaciones. A lo largo de cada una de las generaciones se crean nuevos individuos mediante operaciones de transformación, los cuales se conocen como operadores genéticos de cruza.

Cada generación incluye un proceso de selección, que da mayor probabilidad de permanecer en la población y participar en las operaciones de reproducción a los mejores individuos. Esta selección debe ser aleatoria para poder dar oportunidad a los individuos con menor adaptación de poder ser elegidos, aunque con menor probabilidad.

Los mejores individuos son aquellos que tienen el mejor valor de la función de aptitud, por esto son los que tienen la mayor probabilidad de sobrevivir y reproducirse.

A continuación, en la Figura 3, se muestra el diagrama de flujo del algoritmo genético (Kuri, 2007), (Araujo, 2009), (Suanmali, 2011).

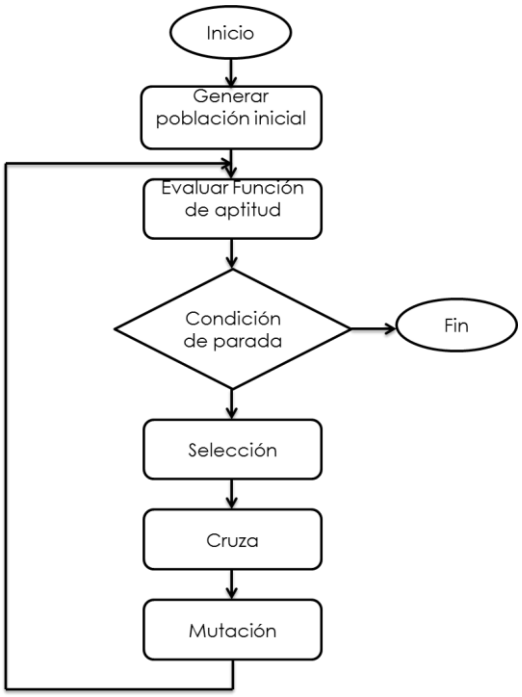

Figura 3. Diagrama de flujo del algoritmo genético.

De acuerdo a lo anterior a continuación se describe en qué consiste cada una de las principales etapas del algoritmo genético. Cabe mencionar, que para poder implementar cada una de las etapas mencionadas en el diagrama, primero se debe tener una representación de individuo.

2.3.1.1 Función de Aptitud

La función de aptitud dependerá completamente del problema que se quiera resolver, busca determinar cuál de los individuos es el más apto para sobrevivir y reproducirse.

Para determinar qué individuo es una buena solución es necesario calificarlo de alguna manera. Cada individuo de cada generación de un algoritmo genético recibe una calificación o, para usar el término biológico, una medida de su grado de adaptación (fitness). Éste es un número real no negativo tanto más grande cuanto mejor sea la solución (Kuri, 2007).

Al hablar de que un individuo de la población se le asigna una y solo una calificación, se está hablando de una función denominada, función de adaptación (Kuri, 2007).

Según Ladda Saunmali (Suanmali, 2011), el valor de la función de aptitud refleja lo bien que un cromosoma se compara con el otro cromosoma en la población. El cromosoma tiene una mayor probabilidad de supervivencia y reproducción entonces puede representar a la próxima generación.

2.3.1.2 Selección

Una parte fundamental del funcionamiento de un algoritmo genético es el proceso de selección de candidatos a reproducirse. En el algoritmo genético este proceso de selección suele realizarse de forma probabilística (es decir, aún los individuos menos aptos tienen cierta oportunidad de sobrevivir).

Existen diferentes técnicas de selección, entre ellas se encuentran: selección por ruleta (Araujo, 2009), (Kuri, 2007), (Coello, 2010), (Guervós, 2013), muestreo estocástico universal, muestreo por restos, selección por torneo (Araujo, 2009), selección por jerarquías, escalamiento sigma, selección de Botzmann (Carlos, 2010), entre otros.

En este trabajo se hace uso de la selección por ruleta y selección por torneo, por lo que serán las que se expliquen a detalle.

2.3.1.2.1 Selección por ruleta

Se crea la selección una ruleta con los cromosomas presentes en una generación. Cada cromosoma tendrá una parte de esa ruleta mayor o menor en función a la puntuación que tenga cada uno. Se hace girar la ruleta y se selecciona en el cromosoma en el que se para la ruleta. Obviamente el cromosoma con mayor puntuación tendrá mayor probabilidad. En caso de que las probabilidades difieran mucho, este método de selección dará problemas puesto que si un cromosoma tiene un 90% de posibilidades de ser seleccionado, el resto apenas saldrá lo que reduciría la diversidad genética.

La idea principal para la selección de tipo ruleta, es que los mejores individuos tengan más oportunidad de ser elegidos, ya que ésta probabilidad se da proporcional a la puntuación obtenida en la función de aptitud, por lo cual se asigna a cada individuo una parte de la ruleta proporcional a su función de aptitud (Araujo, 2009).

Según Carlos A. Coello (Coello, 2010) el algoritmo de ruleta es el siguiente:

- Calcular la suma de valores esperados T
- Repetir N veces (N es el tamaño de la población)
    - Generar un número aleatorio r entre 0.0 y T
    - Ciclar a través de los individuos de la población sumando los valores esperados hasta que la suma sea mayor o igual a r
    - El individuo que haga que esta suma exceda el límite es el seleccionado

Si se tuviera una población de 3 individuos y el valor de $\mathrm{T} = 6$, se tiene la siguiente representación gráfica mostrada en la Figura 4.

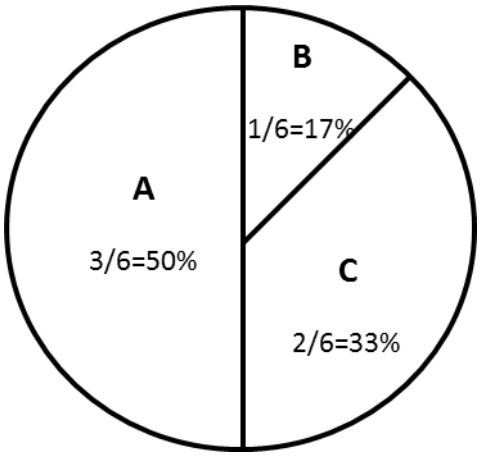

Figura 4. Ejemplo de selección por ruleta.

En la Figura 4, se puede ver que el individuo A tiene un 50 % de ser seleccionado, por lo que es el que tiene más posibilidad de ser elegido para pasar a la cruza. Sin embargo, el individuo B a pesar de ser el que menos posibilidad tiene de ser elegido con un 17 %, también puede ser seleccionado.

2.3.1.2.2 Selección por torneo

La selección por torneo, constituye un procedimiento de selección de padres muy extendido y en el cual la idea consiste en escoger al azar un número de individuos de la población, tamaño del torneo, (con o sin reemplazamiento), seleccionar el mejor individuo de este grupo, y repetir el proceso hasta que el número de individuos seleccionados coincida con el tamaño de la población. Habitualmente el tamaño del torneo es 2, y en tal caso se ha utilizado una versión probabilística en la cual se permite la selección de individuos sin que necesariamente sean los mejores.

2.4 Métodos de evaluación

Una vez que un método genera un resumen automático se puede evaluar automáticamente con los realizados por un humano, para esto se ocupan sistemas de evaluación.

En el principio los métodos de evaluación eran manuales, es decir que los que juzgaban la calidad de los resúmenes eran directamente los humanos, posteriormente se desarrollaron métodos de evaluación automáticos con el objetivo de disminuir el costo y el tiempo que implican los métodos de evaluación manuales y por otro lado aumentar la integridad de la tarea.

El método de evaluación intrínseco más utilizado hoy en día es el que emplea el sistema ROUGE (Lin, 2004), el cual compara el resumen que se desea evaluar (resumen candidato) con resúmenes creados por humanos (resúmenes modelo o de referencia).

Otro método de evaluación es SUMMAC (Summarization Evaluation Confcerence), que fue un sistema de resúmenes automáticos de texto que tuvo lugar en 1998 como parte del programa TIPSTER de la Administración de Proyectos Avanzados de Investigación de Defensa viene del inglés Defense Advanced Research Projects Administration. Participaron 16 sistemas teniendo en cuenta la evaluación extrínseca de dos tareas del mundo real. La primera consistía en procesar una lista de documentos para encontrar los relevantes. La segunda era una tarea de categorización en la que, por ejemplo, se presentaba un conjunto de 1000 documentos que debían ser agrupados en 10 clases (Márquez, 2010).

En el área de generación de resúmenes existen dos medidas de evaluación frecuentemente utilizadas, precisión y recuerdo (Ledeneva, 2008). El recuerdo está definido como la probabilidad de detectar un objeto dado que es relevante, mientras que la precisión se define como la probabilidad de que un objeto es relevante dado que fue detectado (Zhu, 2004).

Para este trabajo las medidas de precisión y recuerdo permiten evaluar las oraciones del resumen hecho por un humano contra uno generado automáticamente. Esta evaluación de realiza considerando las oraciones que conforman los resúmenes.

A continuación se muestran las formulas correspondientes a precisión y recuerdo.

$$Presición = \frac{correctas}{(correctas + incorrectas)}$$

$$Recuerdo = \frac{correctas}{(correctas + olvidadas)}$$

Se define como $correctas$ al número de oraciones extraídas por el sistema y por el humano; $incorrectas$ como el número de oraciones extraídas por el sistema pero no por el humano y $olvidadas$ como el número de oraciones extraídas por el humano pero no por el sistema.

F-Measure es una métrica que combina las ideas de recuerdo y precisión en la recuperación de información (Arco, 2006) De acuerdo con Porta (Porta, 2005), la medida F-Measure está definida por:

$$F - Measure = \frac{2 * recuerdo * precisión}{recuerdo + precisión}$$

2.5 Heurística en la generación automática de resúmenes

2.5.1 Baseline

Dentro del área de generación automática de resúmenes, existen varias heurísticas recientemente utilizadas. Una de las heurísticas aplicadas a la generación automática de resúmenes es conocida como *Baseline*, la cual consiste en tomar las n primeras líneas del texto para conformar el resumen. Este procedimiento se lleva a cabo debido a la hipótesis que asegura que la información más importante de un documento se encuentra en las primeras secciones de este (Ledeneva, 2008)

2.5.2 Baseline aleatorio

Esta heurística no pretende obtener los mejores resultados, pero trata de ayudar a determinar la calidad de los resúmenes, ya que su funcionamiento sólo consiste en tomar de un conjunto de oraciones algunas al azar. La idea es determinar cuan significativos son los resultados con respecto a esta heurística (Ledeneva, 2008)

CAPÍTULO 3

Estado del Arte

En este capítulo, se presentan las herramientas comerciales y los métodos del estado del arte para la generación automática de resúmenes. Primero, se explican las herramientas y los métodos del estado del arte por idioma y posteriormente, se presenta un análisis tanto de las herramientas como de los métodos del estado del arte con respecto a la independencia del lenguaje.

3.1 Métodos del estado del arte

A continuación se presentan los diferentes métodos del estado del arte propuestos para la generación automática de resúmenes.

3.1.1 Métodos del estado del arte para el lenguaje inglés

A continuación se mencionan los métodos del estado del arte para la generación automática de resúmenes en el lenguaje inglés. Los métodos presentados trabajan con la colección de documentos DUC2002 y se evalúan con ROUGE.

3.1.1.1 UnifiedRank

El método UnifiedRank (Wan, 2010) es un método basado en grafos enfocado a generar resúmenes para un solo documento y para multi-documentos. Este trabajo examina la influencia que existe entre la generación de resúmenes para un sólo documento y multiples-documentos. El corpus con el que trabaja para la generación de resúmenes para un solo documentos es DUC 2002 y la herramienta con la que evalúa es ROUGE.

3.1.1.2 MA-SingleDocSum

Es un método Ma-SingleDocSum (Mendoza, 2015) está basado en un algoritmo genético, enfocado en la generación de resúmenes para un solo documento. Además de utilizar operadores genéticos para la generación de los resúmenes utiliza la búsqueda local. Los paramentos que considera para la función de aptitud son: posición de las oraciones, relación de la oración con el título, longitud de la oración, cohesión y la convergencia (conocida como temática del texto). Los experimentos están realizados utilizando la colección de documentos DUC2002 y la herramienta de evaluación utilizada es ROUGE.

3.1.1.3 AG (Bag words)

El método propuesto por (García, 2013), es uno de los que han obtenido los mejores resultados. Esta realizado mediante un algoritmo genético y utiliza el modelo de texto bolsa de palabras. La función de aptitud utilizada en el trabajo de (García, 2013) toma dos características principales, las cuales se mencionan a continuación:

- Las primeras oraciones son más importantes, se considera a las primeras oraciones de un texto como candidatas a formar parte del resumen.
- Evaluar que el resumen tenga diferentes ideas, es decir que no sea repetitivo, pero que a la vez tenga palabras importantes (Precisión-Recuerdo).

3.1.1.4 TextRank

Este método consiste en un algoritmo de ponderación basado en grafos. De acuerdo con Rada Mihalcea (Mihalcea, 2004) construye un grafo para representar el texto, de manera que los nodos son palabras (u otras entidades de texto) interconectadas mediante arcos con relaciones significativas. Para la tarea de extracción de oraciones, el objetivo es calificar oraciones enteras y ordenarlas de mayor a menor importancia. Por lo tanto, se agrega un arco al grafo por cada oración en el texto. Para establecer las conexiones entre oraciones, se define una relación de similitud, donde la relación entre dos oraciones puede ser vista como un proceso de "recomendación": una oración que señala a cierto concepto en el texto da al lector una "recomendación" para referirse a otras oraciones en el texto que

señalan a los mismos conceptos y por tanto, un vínculo puede establecerse entre dos oraciones cualesquiera que compartan un contenido común.

3.1.1.5 Secuencias Frecuentes Maximales (SFMs)

Este trabajo presenta un método basado en estadística, que es independiente del dominio y del lenguaje, para generar el resumen extractivo de un solo documento. En su trabajo, Yulia Ledeneva (Ledeneva, 2008) (Ledeneva, 2008a) muestra experimentalmente que las palabras que son partes de bigramas (secuencias de 2 palabras) que se repiten más de una vez en el texto, son buenos términos para describir el contenido de ese texto, al igual que las llamadas Secuencias Frecuentes Maximales (secuencias de palabras que se repiten cierto número de veces y que además no están contenidas en otras secuencias frecuentes). También se muestra que la frecuencia del término como pesado de términos brinda buenos resultados (mientras solo se cuenten las ocurrencias de un término en bigramas repetitivos).

Ledeneva aplica una técnica de cuatro pasos para generar el resumen. Dichos pasos son la selección de términos, pesado de términos, pesado de oraciones y selección de oraciones. En la selección de términos se extraen: las SFMs, los bigramas repetitivos (deben parecer por lo menos dos veces en el texto), y las palabras simples o unigramas. En el pesado de términos se usa la frecuencia del término, que consiste en el número de veces que el término ocurre en el texto dentro de una SFMs. También se utiliza como pesado la máxima longitud de una SFM que contenga al término, así como el asignar un mismo peso para todos los términos. En el pesado de oraciones, solo se suma el peso de todos los términos contenidos en esa oración. Finalmente, la selección de las oraciones que conformaran el resumen se lleva a cabo mediante dos criterios: primero, se seleccionan las mejores oraciones, es decir, las que obtuvieron mayor peso; esto se lleva a cabo hasta alcanzar la longitud deseada (100 palabras) del resumen. En el segundo criterio, se seleccionan las k oraciones mejores, además de las primeras que aparecen en el documento (kbest+first). Esto se realiza hasta alcanzar la longitud deseada del resumen. El mejor resultado es obtenido con el método de palabras (unigramas) y 1best+first, con un 47% de similitud con los resúmenes hechos por un humano.

3.1.1.6 Clustering with SFMs

En el método anterior de SFMs, las sentencias que tienen mayor peso son seleccionadas para componer el resumen. Sin embargo, si existen sentencias muy similares y se eligen, no proporcionan nueva información al resumen. En este trabajo de agrupamiento de oraciones con SFMs con K-means (García, 2008) se realizan grupos de oraciones, de las cuales se

selecciona la frase más repetitiva de cada grupo y esta compone el resumen. Este trabajo también se realizó con agrupamiento con EM en el trabajo (Ledeneva, 2011).

3.1.1.7 SFMs (1 best + first)

En este trabajo se presenta un método basado en estadística, que es independiente del dominio y del lenguaje, para generar el resumen extractivo de un solo documento. En su trabajo, Ledeneva (Ledeneva, 2008) muestra experimentalmente que, las palabras que son partes de bi-gramas que se repiten más de una vez en el texto, son buenos términos para describir el contenido de ese texto, al igual que las llamadas secuencias frecuentes maximales. También se muestra que la frecuencia del término como pesado de términos brinda buenos resultados (mientras solo se cuenten las ocurrencias de un término en bi-gramas repetitivos).

3.1.1.8 AG (bi-gramas)

En la construcción del método AG (bi-gramas) propuesto por (Matías, 2013) para la generación de resúmenes automáticos se utilizó la técnica de algoritmos genéticos, con el fin de optimizar el proceso y obtener mejores resúmenes que los creados por otros métodos y herramientas. En el trabajo de Rene García (García, 2013) se utilizó un modelo de palabras y se propuso un algoritmos genético. Cabe mencionar, que el algoritmo genético utilizado para la generación automática de resúmenes mediante el modelo de n-gramas está basado en el trabajo antes mencionado. Para el algoritmo genético se utiliza un tipo de codificación binaria para cada individuo, donde cada oración del documento constituye un gen; 1 significa que la oración aparecerá y 0 que no. La población inicial se genera de manera aleatoria. Para seleccionar los mejores individuos de acuerdo a su aptitud se utilizó la selección por ruleta.

Los operadores genéticos se construyen de acuerdo al problema que se quiere resolver, por lo que el operador de cruza se ha adaptado a la generación de resúmenes, ya que en este tipo de problema una oración no puede repetirse, y para el operador de mutación se utiliza la mutación por intercambio. Sin embargo, también se ha adaptado a la generación de resúmenes debido a que se pide un número mínimo de palabras que debe contener el resumen y por esta razón si un bit es mutado y no cumple la condición de tener el mínimo de palabras se debe ajustar (García, 2013). La condición de parada que se aplicó para el término del algoritmo genético es el número máximo de generaciones.

3.1.2 Métodos del estado del arte para el lenguaje portugués

A continuación se mencionan los métodos del estado del arte para la generación automática de resúmenes para el lenguaje portugués. Los métodos descritos prueban con la

colección TeMário. Sin embargo, utilizan diferentes herramientas para la evaluación de sus resúmenes.

3.1.2.1 SuPor

SuPor (SUmmarizer for PORtuguese) es un sistema basado en una máquina de aprendizaje (Módolo, 2003). Por lo tiene dos procesos distintos: el entrenamiento y la extracción basada en el método de Naive-Bayes. Esto le permite la combinación de rasgos lingüísticos y no lingüísticos. Las características que considera SuPor para la generación de los resúmenes son: la longitud de la oración (mínimo 5 palabras), frecuencia de las palabras, señalización de la frase, ubicación de la oración y ocurrencia de nombres propios. A continuación se menciona el funcionamiento de SuPor. En primer lugar se extrae el conjunto de características de cada oración. En segundo lugar, para cada conjunto se aplica el clasificador bayesiano, el cual proporciona la probabilidad de que la oración se incluya en el resumen. Las de mayor probabilidad forman parte del resumen.

3.1.2.2 SABio

SaBio (Automatic Summarizer for the Portuguese language with more Biologically plausible connectionist architecture and learning) está basado en una red neuronal entrenada con noticias del corpus TeMário (Orrú,2006). Este método considera las siguientes características: tamaño de la oración, posición de la oración en el texto, posición de la oración dentro del párrafo en el que pertenece, presencia de las palabras claves, valor de la oración con respecto a la distribución de las palabras en el texto, frecuencia de los términos,

3.1.2.3 GistSumm

GistSumm es un resumidor automático basado en un método de integración llamado gist-based (Pardo, 2003). Se compone de tres procesos: segmentación del texto, rankeo de oraciones y la generación del resumen. El rankeo de oraciones está basado en el método de (Luhn, 1958) el cual está basado en palabras claves, se pondera cada oración del texto original por medio de la frecuencia de las palabras, las palabras claves tienen mayor peso. El resumen se produce considerando la correlación entre la palabra clave y la relevancia que esta tiene con relación al contenido del texto.

3.1.2.4 Evaluación de las herramientas comerciales de generación automática de resúmenes de textos para el idioma portugués

En el trabajo realizado por (Ibañez, 2013), se realiza una evaluación de las herramientas comerciales y los métodos del estado del arte para la generación de resúmenes en el lenguaje portugués. Entre las herramientas comerciales esta la herramienta instalable Word

en sus versiones, 2003 y 2007. Las herramientas en línea que se evalúan son, OTS, Shvoong y Tools4noobs. Los métodos del estado del arte evaluados son, GistSumm y los propuestos por (Mihalcea, 2005).

3.1.3 Métodos del estado del arte para el lenguaje español

A continuación se describen algunos métodos del estado del arte que generan resúmenes automáticos para el lenguaje español. Es importante mencionar que para el lenguaje español no se cuenta con una colección de documentos estándar para la generación de resúmenes, por lo cual cada autor trabaja con una colección propia. Cabe mencionar que algunas de las colecciones que utilizan los diversos autores no se encuentras disponibles para su utilización. Las colecciones que están disponibles para el lenguaje español fueron adaptadas por los autores para la generación de resúmenes, ya que muchas de ellas tenían como función principal el agrupamiento, la relevancia de palabras claves, entre otras funciones.

3.1.3.1 Uso de grafos semánticos en la generación automática de resúmenes y estudio de su aplicación en distintos dominios: Biomedicina, Periodismo y Turismo

El trabajo de (Plaza, 2010) se completa con tres casos de estudio en los que el método diseñado se configura y utiliza para generar distintos tipos de resúmenes de textos de diferentes dominios y con unas características de estructura y estilo muy dispares: artículos científicos de biomedicina, noticias periodísticas y páginas web de información turística en el lenguaje español.

El método que utiliza está basado en el uso de grafos semánticos, el cual está constituido por las siguientes etapas: pre-procesamiento, traducción de las oraciones a conceptos, representación de las oraciones como grados de conceptos, construcción del grafo del documento, *clustering* de conceptos, asignación de oraciones a *clusters*, selección de oraciones para el resumen y finalmente la construcción del resumen.

3.1.3.2 Compresión automática de fases: un estudio hacia la generación de resúmenes en español

El trabajo de (Molina, 2013) propone la generación de resúmenes automáticos para el lenguaje español considerando las siguientes características del texto. La segmentación discursiva, la cual consiste en representar el documento a través de un árbol jerárquico que contiene información tipo retórico/discursivo. La comprensión de frases por eliminación de segmentos discursivos, la cual se basa en la gramaticalidad de la frase resultante; en su normatividad (entendida como la calidad de información importante retenida) y en la tasa

de comprensión. La gramaticalidad la cual consiste en determinar si una frase es correcta o no y finalmente la normatividad la cual está basada en la frecuencia de las palabras.

El trabajo de (Molina, 2013) propone dos algoritmos basados en los puntos anteriores para la generación de resúmenes automáticos. El primero es la generación de resúmenes por eliminación de segmentos y el segundo es generación de resúmenes por eliminación de segmentos con tasa de comprensión como argumento. Para la experimentación (Molina, 2013) utiliza un corpus propio que no está disponible.

3.1.3.3 Generación de resúmenes de múltiples documentos

El trabajo de (Villatoro, 2007) está basado en un clasificador, y el uso de herramientas de aprendizaje supervisado. La idea básica con la que funciona el método es que un proceso inductivo automáticamente construya un clasificador por medio de observar las características de un conjunto de documentos previamente resumidos, lo que se le da al algoritmo de aprendizaje son pares (documento, resumen). De tal forma que el problema de generación resúmenes se convierte en una actividad de aprendizaje supervisado.

Para la experimentación con el lenguaje en español se utiliza el corpus Desastres (Telléz, 2009). Cabe mencionar, que el corpus está diseñado para clasificación y fue adaptado para la generación de resúmenes.

3.2 Herramientas comerciales

Las herramientas comerciales son aquellas que pueden como su nombre lo indica comercializarse, y el método con el que trabajan no es publicado ya que como la herramienta tiene un costo su funcionamiento interno no es de dominio público. Entre ellas están Copernic Summarizer y Microsoft Office Word.

3.2.1 Herramientas comerciales instalables

A continuación se describen las herramientas comerciables que se instalan en una computadora para poder generar un resumen.

3.2.1.1 Copernic Summarizer

Copernic Summarizer (Copernic Inc) es un software que fue desarrollado exclusivamente para la generación de resúmenes automáticos, lo cual hace que sea una herramienta flexible y adecuada porque ofrece las opciones de que el resumen resultante sea del 5%, 10%, 25% o 50% de palabras del texto original; o resúmenes de 100, 250 y 1000 palabras sin importar el tamaño del texto original.

3.2.1.2 Microsoft Office Word

Esta herramienta es una suite ofimática la cual podemos encontrar en las versiones de Microsoft Office Word 2003 (Microsoft ® Office Word 2003 (11.8307.8221) SP3) y Microsoft Office Word 2007 (Microsoft ® Office Word 2007 (12.0.4518.1014) MSO). Esta herramienta permite generar resúmenes de 10 o 20 oraciones; 100 o 500 palabras (o menos); o bien en porcentajes de 10%, 25%, 50% y 75% de palabras del documento original. Si algunos de los porcentajes no es adecuado, el usuario lo puede cambiar según sus necesidades.

3.2.2 Herramientas comerciales en línea

A continuación se describen las herramientas comerciables en línea, a las cuales se acceden desde internet y así poder generar un resumen.

3.2.2.1 Tools4noobs

Tools4noobs (Tools4noobs, 2007) es una herramienta en línea que permite generar resúmenes desde 1 al 100 % del texto original. Para la generación de un resumen Tools4Noobs tiene 3 faces: extracción de las oraciones, identificación de las palabras claves del texto contando la relevancia de cada palabra e identificación de las oraciones de acuerdo a las palabras claves identificadas.

3.2.2.2 Pertinence Summarizer

Pertinence Summarizer (Pertinence, 2009) pertenece a la gama de productos desarrollados con tecnología denominada KENiA© (basada en la extracción de conocimiento y arquitectura de notificación) desarrollada por la empresa francesa *Pertinence Mining*. *Pertinence* es una herramienta en línea que permite generar resúmenes en 12 idiomas (Alemán, Inglés, Árabe, Chino, Coreano, Español, Francés, Italiano, Japonés, Portugués, Ruso y Neerlandés) de los documentos de texto en formatos diversos (html, pdf, doc, rtf y txt).

3.2.2.3 OTS (Open Text Summarizer)

Open Text Summarizer (OTS, 2007) es una aplicación de código abierto para resumir textos, que puede ser descargada de Internet de forma gratuita. Sin embargo, también puede encontrarse la interfaz de ésta en línea (Gohr, 2001-2013 ©). OTS genera resúmenes automáticos en diferentes porcentajes y puede generar resúmenes en 37 idiomas.

3.2.2.4 Shvoong

Shvoong (Svhoong, 2005) fue fundado en 2005 por Avi Shaked y Avner Avrahami. Shvoong es una herramienta que permite generar resúmenes automáticos en 21 idiomas diferentes (Checo, Neerlandés, Danés, Inglés, Finlandés, Francés, Alemán, Griego, Hebreo, Húngaro, Indonesio, Italiano, Malayo, Noruego, Polaco, Portugués, Rumano, Ruso, Español, Sueco y

Turco). A diferencia de otras herramientas *Shvoong* no devuelve el resumen como tal, sino que subraya el texto que considera más importante del documento original.

3.2.2.5 Article Summarizer Online

Article Summarizer Online (Summarizer, 2016) es una herramienta que proporciona en línea una aplicación gratuita para la generación automática de resúmenes. Para poder generar un resumen se siguen los siguientes pasos. Se pega el texto en el recuadro donde dice "*Your Text*" de la herramienta *Summarizing* y una vez que este el texto en el recuadro de *Summary length* daremos el número de palabras que requerimos que sea el resumen (se tiene la opción de 100, 150, 200 y 300 palabras). Cabe mencionar que se pueden generar resúmenes de mayor longitud, pero se debe enviar el documento y la elaboración tiene un costo.

3.2.2.6 Text Compactor

Text Compactor (Edyburn, 2010), es una herramienta gratuita para la generación automática de resúmenes, fue creada por Keith Edyburn y está basada en la herramienta Open Text *Summarizer*. Para generar un resumen se siguen los siguientes pasos. Se escribe o se pega el texto en el recuadro y posteriormente se elige el porcentaje de texto que quieres que mantenga el resumen el cual puede ser desde un 0% a un 100% del texto.

3.3 Análisis de los métodos del estado del arte y de las herramientas comerciales

A continuación se muestra el análisis de los métodos del estado del arte y de las herramientas comerciales, con respecto a los métodos que son independientes del lenguaje, los que dicen que son independientes del lenguaje pero sólo prueban con una colección de documentos y finalmente, un análisis con respecto a la posición de las oraciones, una de las características más utilizadas para la generación automática de resúmenes.

3.3.1 Métodos independientes de lenguajes

En la tabla 5, se muestran algunos de los métodos del estado del arte que dicen ser independientes del lenguaje y prueban con más de una colección de documentos. Sin embargo, se puede observar que ninguno prueba con una colección de documentos en español.

Tabla 5. Métodos del estado del arte independientes del lenguaje

Nombre del artículo	Colección de documentos	Idioma	Herramienta de evaluación
Independiente del lenguaje enfocado a resúmenes de texto multilingües(Patel, 2007)	DUC-2002, Documentos en idiomas de la India	Inglés, Hindi, Gujarati y Urdu	Evaluación intrínseca
Algoritmo independiente del lenguaje para resúmenes de uno y múltiples documentos (Mihalcea, 2005)	DUC-2002 y TEMARIO	Inglés y Portugués	ROUGE
Técnicas independientes del lenguaje para resumen de texto automático (Last, 2010)	DUC-2002 y noticias en Hebreo	Inglés y Hebreo	ROUGE
Usando SUMMA para resúmenes independientes del lenguaje en TAC2011(Saggion, 2011)	TAC multilingüe 2011	Arabic, English, French and Hindi	ROUGE

3.3.2 Métodos que son independientes del lenguaje pero prueban con sólo una colección

En la tabla 6, se muestran los métodos del estado del arte que dicen ser independientes del lenguaje. Sin embargo, solo prueban con una colección de documentos, por lo que no está comprobado que en realidad lo sean.

Tabla 6. Métodos del estado del arte que dicen ser independientes del lenguaje y sólo prueban con una colección de documentos.

Nombre del artículo	Colección de documentos	Idioma	Herramienta de evaluación
Términos Derivados de secuencias frecuentes para resúmenes de textos extractivos (Ledeneva, 2008)	DUC-2002	Inglés	ROUGE
Lenguaje Automático- Detección Independiente de descripciones de varias palabras para resúmenes de texto (Ledeneva, 2008a)	DUC-2002	Inglés	ROUGE
Resúmenes de texto por extracción de oraciones usando aprendizaje no supervisado (Hernández,2008)	DUC-2002	Inglés	ROUGE
Generación automática de resúmenes mediante aprendizaje no supervisado (Montiel, 2009)	DUC-2002	Inglés	ROUGE
Generación automática de resúmenes usando algoritmos genéticos (Matias, 2013)	DUC-2002	Inglés	ROUGE
Resumen extractivo de texto basado en un algoritmo genético (García, 2013)	DUC-2002	Inglés	ROUGE

Entre los métodos que se mencionan está el de (Matias, 2013), el cual obtiene uno de los mejores resultados. Sin embargo, aunque menciona que es independiente del lenguaje sólo prueba con el lenguaje inglés. Considerando que se tiene un método que obtiene buenos resultados se puede probar con otros lenguajes para comprobar su independencia y posteriormente, se pueden proponer mejoras.

3.3.3 Análisis de las características del texto utilizadas para la generación automática de resúmenes.

Tabla 7. Características del texto utilizadas para la generación automática de resúmenes.

Características del texto	(Matias,2013), Inglés	(Mendoza,2014), Inglés	(Bossard,2008), Inglés	(Ouyang,2010), Inglés	(Nandhini,2014),Inglés	(Lin,1999), Inglés, Francés,	(Hirao,2002),Japonês	(Katragadda,2009), Inglés	(Nizam,2007),Bangla	(Orasan,2003), Inglés	(Berker,2011), Inglés	(Alfonseca,2003),Inglés	(Ladda,2011), Inglés	(Qazvinian,2008), Inglés	(Mateo,2003), Español
Posición de la oración	x	x	x	x	x	x	x	x		x	x	x	x		x
Longitud de las oraciones		x	x		x	x	x		x		x	x	x		
Relación de la oración con el título		x	x		x	x			x	x			x	x	x
Temática (frecuencia)	x	x				x			x		x		x		x
Datos numéricos						x			x		x		x		
Nombres propios					x	x							x		x
Similitud con una consulta			x			x						x			x
Centralidad			x		x						x		x		
similitud con los fragmentos			x			x								x	
frases de referencia					x					x	x				
Palabras de activación -Trigger words					x		x								x
Cohesión /Similitud		x												x	
Nombre de entidades							x				x				
Sentimiento			x												
Similitud con la primera oración			x												
Longitud de la palabra					x										
palabras polisílabas					x										
Ocurrencia de sustantivos					x										
pronombre y adjetivo						x									
Día de la semana y el mes						x									
Cita						x									
Tipografía del texto															x

En la tabla 7, se muestra un análisis de las características del texto que se utilizan en la generación automática de resúmenes, además se menciona para que lenguaje se aplican.

3.3.4 Análisis de la posición de las oraciones

Como se puede observar en la tabla 8 una de las características más implementadas es la posición de las oraciones. Por lo que en este trabajo se realizó un análisis de esta característica. Primero se buscó en el estado del arte las formas propuestas para determinar la importancia de las oraciones y posteriormente se implementaron en el método. A continuación se muestran las formulas propuestas en (Ouyang, 2010), (Mendoza, 2014), (Vázquez, 2015) y (García, 2013), las cuales fueron probadas para determinar cual da mejores resultados.

Tabla 8. Fórmulas propuestas en el estado del arte para la posición de las oraciones.

(Ouyang, 2010)	$I = \frac{(N - i + 1)}{N}$
(Vázquez, 2015)	$f = \frac{(-49.8563 + 2 * X + 0.5\,(-90.4102 + X) * X^{-X})}{\left(\left(\frac{\left(\frac{X}{14.7}\right)^{14.7}}{X}\right) - 92.5\right)}$
	$\frac{(-28.7 - N)}{-57.4}$
(Mendoza,2014)	$I = \sqrt{\frac{1}{i}}$
(García, 2013)/(Matías, 2013)	$\delta = \frac{\sum_{\lvert c_i \rvert=1}^{n} m(i-x)+x}{\sum_{j=1}^{k} m(j-x)+1},\ x = 1 + \frac{(n-1)}{2}.$

CAPÍTULO 4

Método Propuesto

En este capítulo, se presenta el método propuesto y se describen cada una de sus etapas.

4.1 Método propuesto

En la figura 5 se muestra el método propuesto en este trabajo. Como se mencionó en el estado del arte, el método de (Matias, 2013) es un método que dice ser independiente del lenguaje. Sin embargo, sólo prueba con una colección de documentos en el lenguaje inglés obteniendo buenos resultados. Por lo que para este trabajo se retomó el trabajo de (Matias, 2013) para probar con otras colecciones de documentos en el lenguaje portugués y español. Además se modificaron algunos de sus parámetros para mejorar los resultados.

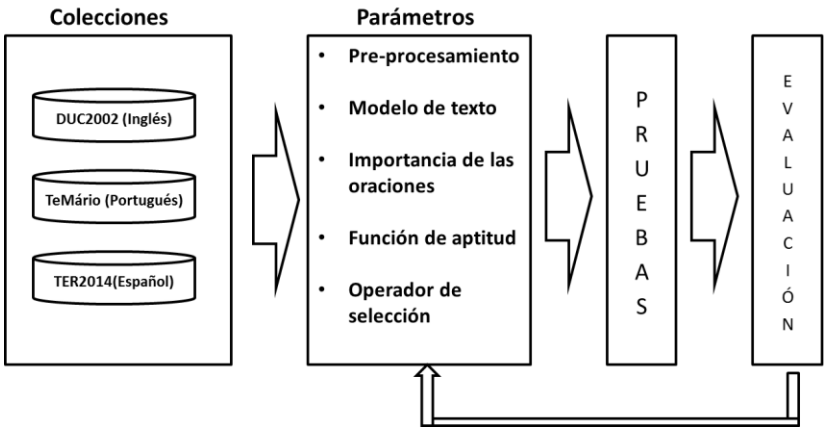

Figura 5. Metodología de trabajo.

El método propuesto está dividido en cuatro etapas, las cuales se describen a continuación.

4.1.1 Colecciones

Como primera etapa se obtienen las colecciones de documentos en diferentes lenguajes. Para este trabajo se utilizaron tres colecciones de documentos en los lenguajes: inglés, portugués y español.

A continuación se describen las colecciones utilizadas para generar los resúmenes en los lenguajes inglés, portugués y español. Cabe mencionar que para los lenguajes inglés y portugués se utilizaron corpus estándar disponibles para la generación de resúmenes, mientras que para el lenguaje español se creó el corpus debido a que no se contaba con un corpus de documentos que fuera especial para la generación de resúmenes.

4.1.1.1 Corpus en Inglés (DUC2002)

Document Understanding Conference (DUC, 2002), es una colección de documentos creada por *National Institute of Standards and Technology* (NIST) para el uso de los investigadores en generación de resúmenes. Esta colección está compuesta por 567 noticias en inglés de diversas longitudes, sobre temas de tecnología, alimentación, política, finanzas, entre otros. Para cada documento de la colección se le crearon dos resúmenes por dos humanos expertos con una longitud mínima de 100 palabras.

4.1.1.2 Corpus en Portugués (TeMário)

TExtos com suMÁRIOS (TeMário), es una colección de documentos compuesta por 100 textos en el idioma portugués. El corpus TeMário está compuesto por textos periodísticos adquiridos de dos periódicos de Brasil, el Folha de Sao Paulo y del Journal de Brasil, sobre temas de mundo, opinión, especial, política e internacional. Para cada documento de la colección un experto humano creo su resumen. Para esta colección cada resumen tienen una longitud de 25 - 30% del tamaño de su texto fuente.

4.1.1.3 Corpus en Español mexicano

Textos en Español para Resúmenes (TER), es una colección de documentos compuesta por 240 noticias en el idioma español. El corpus TER está compuesta por noticias periodísticas adquiridos del periódico mexicano Crónica, sobre 12 diferentes categorías, academia, bienestar, ciudad, cultura, deportes, espectáculos, estados, mundo, nacional, negocios, opinión y sociedad. Para cada documento de la colección se crearon dos resúmenes por dos humanos expertos. En el anexo 5 se muestra la documentación correspondiente a la creación de TER.

4.1.2 Parámetros

En la segunda etapa se modificaron los siguientes parámetros:

4.1.2.1 Pre-procesamiento

Las opciones para esta etapa son:

- Con pre-procesamiento – Se da formato a la colección de documentos aplicando *stemming* y eliminando palabras vacías (*stopwords*).
- Sin pre-procesamiento – Se prueba la colección de documentos sin aplicar *stemming* ni eliminar *stopwords*.

Para cada lenguaje se realizaron dos versiones de la colección, una con pre-procesamiento y una sin pre-procesamiento.

4.1.2.2 Modelo de texto

La modificación del parámetro modelo de texto consiste en determinar cuál será la unidad de texto con la que se va a trabajar el método. Los modelos de texto considerados para este trabajo son:

- bolsa de palabras.
- bi-gramas.
- tri-gramas.
- cuatri-gramas.
- cinco-gramas.

Para cada colección de documentos se realizó una versión para cada modelo de texto.

4.1.2.3 Importancia de las oraciones

La modificación del parámetro importancia de las oraciones consiste en probar las diferentes fórmulas propuestas en el estado del arte para la característica, importancia de las oraciones (Tabla 8).

En (Matias, 2013) la importancia de las oraciones es determinada por la siguiente fórmula.

$$\delta = \frac{\sum_{|C_i|=1}^{n} m(i-x)+x}{\sum_{j=1}^{k} m(j-x)+1},\ x = 1 + \frac{(n-1)}{2}.$$

La idea de la fórmula parte de que si todas las oraciones tuvieran la misma importancia se forma una recta para indicar la importancia de las oraciones. Se puede utilizar el punto medio de esa recta para determinar la pendiente de la recta; suavizando así la importancia de las oraciones. Esto nos permitiría saber qué tan importante es una oración con respecto de las siguientes. Entonces podemos decir que la pendiente indica la importancia que se le da a las primeras oraciones o últimas oraciones, si es negativa las primeras oraciones tienen más importancia, (-1) significa que baja hacia la derecha en ángulo de 45 grados, cero quiere decir que todas las oraciones tienen la misma importancia y positiva que las últimas oraciones tienen más importancia (1) significa que sube hacia la derecha en ángulo de 45 grados.

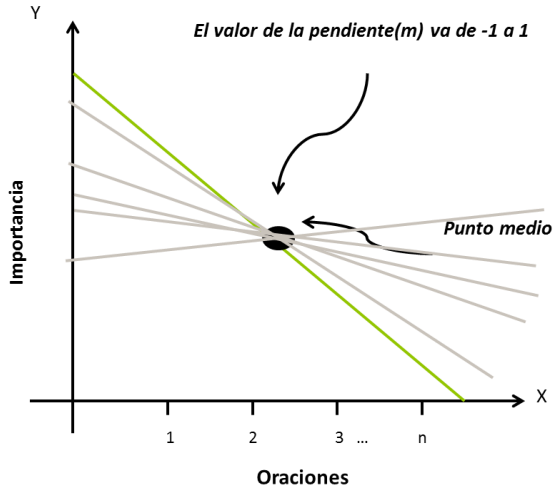

Figura 6. Representación de la importancia de las oraciones usando una recta.

En la Figura 6, se muestra la representación gráfica de la importancia de las oraciones usando una recta. Entonces considerando que si se ajusta el valor de la pendiente (la cual

va de -1 a 1) se puede determinar el valor que se debe usar para la importancia de las oraciones. Para este trabajo se consideraron aleatoriamente los valores mostrados en la Tabla 9.

Tabla 9. Valores de pendiente considerados para determinar el mejor valor para la importancia de las oraciones.

	Valor de pendiente															
Inglés Portugués Español	-0.25	-0.3	-0.375	-0.45	-0.5	-0.55	-0.6	-0.625	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	-1

4.1.2.4 Función de aptitud

La modificación del parámetro función de aptitud consiste, en que considerando lo propuesto en (Matias, 2013) para la función de aptitud, se ajuste como se muestra a continuación.

Posición de las oraciones

$$\boldsymbol{\delta} = \frac{\sum_{|C_i|=1}^{n} m(i-x)+x}{\sum_{j=1}^{k} m(j-x)+1},\ x = 1 + \frac{(n-1)}{2}.$$

Frecuencia de los términos

$$\boldsymbol{\beta} = \frac{\sum_{p=\{word\ \epsilon\ S\}}^{m} frecuency(p,T)}{\sum_{q=\{word\ \epsilon\ T\}}^{m} frecuency(q,T)}.$$

Si se considera que la función de aptitud equivale a 1 se puede decir que $\boldsymbol{\delta}$ equivale a 0.5 y $\boldsymbol{\beta}$ al otro 0.5. Entonces los ajustes realizados se muestran en la tabla 10.

Tabla 10. Ajustes en la función de aptitud.

Posición de las oraciones	Frecuencia de términos
$\mathbf{0.3}\,\boldsymbol{\delta}$	$\mathbf{0.7}\boldsymbol{\beta}$
$\mathbf{0.4}\boldsymbol{\delta}$	$\mathbf{0.6}\boldsymbol{\beta}$
$\mathbf{0.5}\boldsymbol{\delta}$	$\mathbf{0.5}\boldsymbol{\beta}$
$\mathbf{0.6}\boldsymbol{\delta}$	$\mathbf{0.4}\boldsymbol{\beta}$
$\mathbf{0.7}\boldsymbol{\delta}$	$\mathbf{0.3}\boldsymbol{\beta}$

4.1.2.5 Operador de selección

La modificación del parámetro operador de selección consiste en probar dos operadores de selección para el algoritmo genético. Los operadores considerados para este trabajo son: Ruleta y Torneo.

4.1.3 Pruebas

La tercera etapa de la metodología propuesta son las pruebas, donde se realizan los diferentes experimentos modificando cada uno de los parámetros, para cada una de las colecciones en los lenguajes inglés, portugués y español. Cabe mencionar que como se está usando un algoritmo genético y los resultados obtenidos de cada prueba pueden variar, para este trabajo se realizaron dos experimentos por cada una de ellas.

Los resultados mostrados en el capítulo 5 de experimentos son el promedio de los dos experimentos realizados. En el anexo 6 se muestran las tablas con resultados de cada uno de los experimentos, así como el promedio obtenido.

4.1.4 Evaluación

La herramienta de evaluación utilizada en este trabajo fue ROUGE 1.5.5. ROUGE es un sistema automático para la evaluación de resúmenes, propuesto por Lin (Lin, 2004), el cual tiene la capacidad de medir la similitud y determinar la calidad de un resumen automático comparado por el creado por un humano.

CAPÍTULO 5

Experimentación

En este capítulo, se muestran los resultados obtenidos de los experimentos realizados con cada una de las colecciones de documentos, para cada parámetro modificado. Los resultados mostrados en cada gráfica es el promedio de dos experimentos realizados.

Primero, se muestran los resultados obtenidos con el método de (Matias, 2013) para determinar el valor de la pendiente que se consideró para la importancia de la posición de las oraciones y para determinar el mejor modelo de texto para cada colección de documentos. Posteriormente, se muestran los resultados obtenidos de la modificación de cada parámetro.

5.3 Resultados en lenguaje inglés (DUC2002)

5.3.1 Resultados con el valor de la pendiente adecuado para cada modelo de texto

En la figura 7, se muestran los mejores resultados de cada modelo de texto y el valor de la pendiente que mejor se adapta a cada uno de ellos, realizados con la configuración de (Matias, 2013).

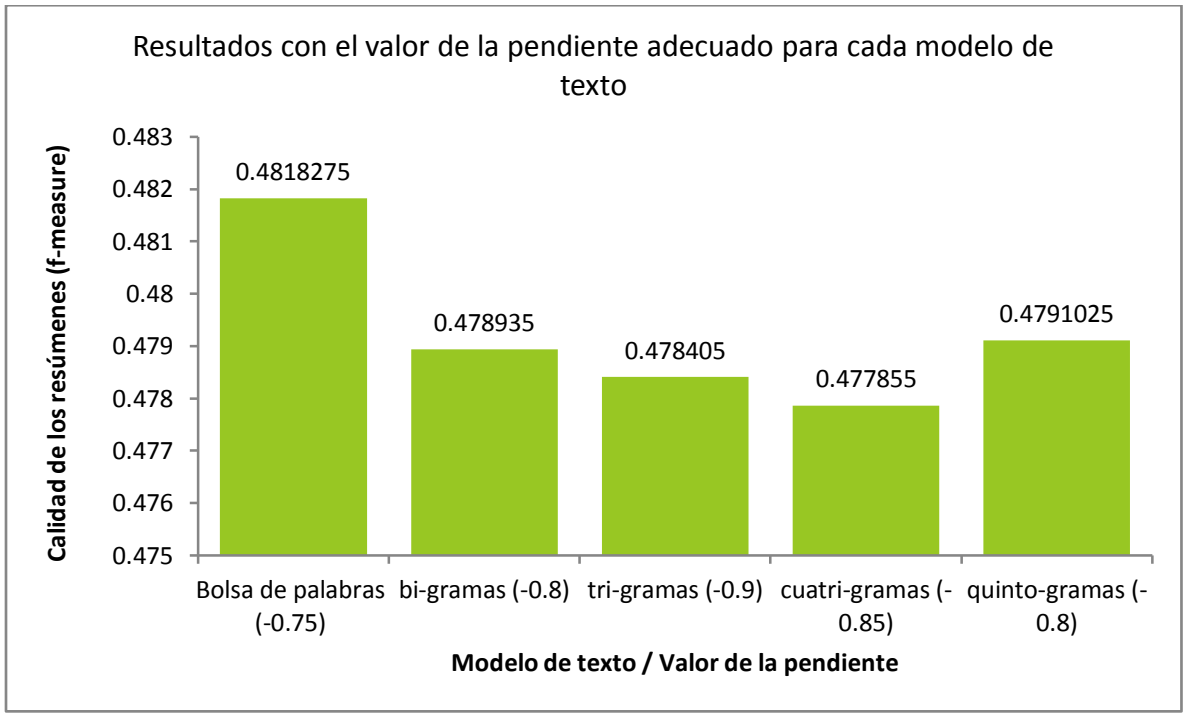

Figura 7. Resultados con el valor de la pendiente adecuado para cada modelo de texto.

En los trabajos de (Matias, 2013) y (García, 2013) se prueba la colección DUC2002 para el lenguaje inglés. En el trabajo de (García, 2013) se reporta como mejor resultado 0.4827 utilizando el modelo de texto bolsa de palabras y un valor de la pendiente de 0.625. Para el trabajo de (Matias, 2013) se reporta como mejor resultado 0.4825 utilizando el modelo de texto bi-gramas con un valor de pendiente de 0.75.

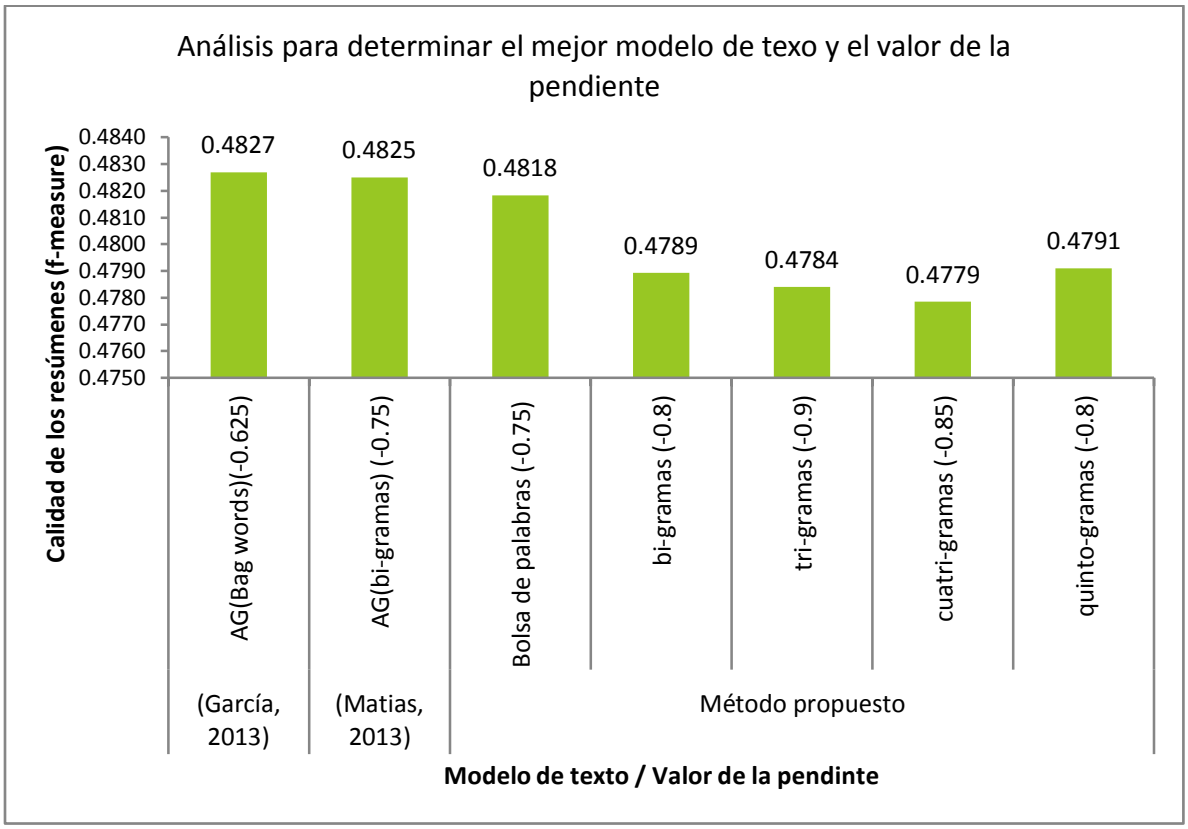

Figura 8. Análisis para determinar el mejor modelo de texto y el valor de la pendiente.

En la Figura 8, se muestran los resultados del método propuesto y de los métodos en el estado del arte para determinar el valor de la pendiente para la importancia de la posición de las oraciones. Como se puede observar los resultados entre (García, 2013) y el método propuesto no se diferencian mucho, mientras que los resultados de (Matias, 2013) y el método propuesto usando bi-gramas si difieren, por lo que se determina que el mejor modelo de texto para los experimentos en el lenguaje inglés es bolsa de palabras. Sin embargo, como la diferencia entre los resultados obtenidos es poca, se realizaron nuevos experimentos para buscar el valor de la pendiente que mejorara los resultados. Los valores de la pendiente probados se muestran en la tabla 11.

Tabla 11. Valores de la pendiente para mejorar los resultados de la colección DUC2002.

Valor de pendiente															
-0.61	-0.62	-0.63	-0.64	-0.66	-0.67	-0.68	-0.69	-0.7	-0.71	-0.72	-0.73	-0.74	-0.75	-0.76	-0.77

De los nuevos experimentos realizados se obtuvieron los siguientes resultados.

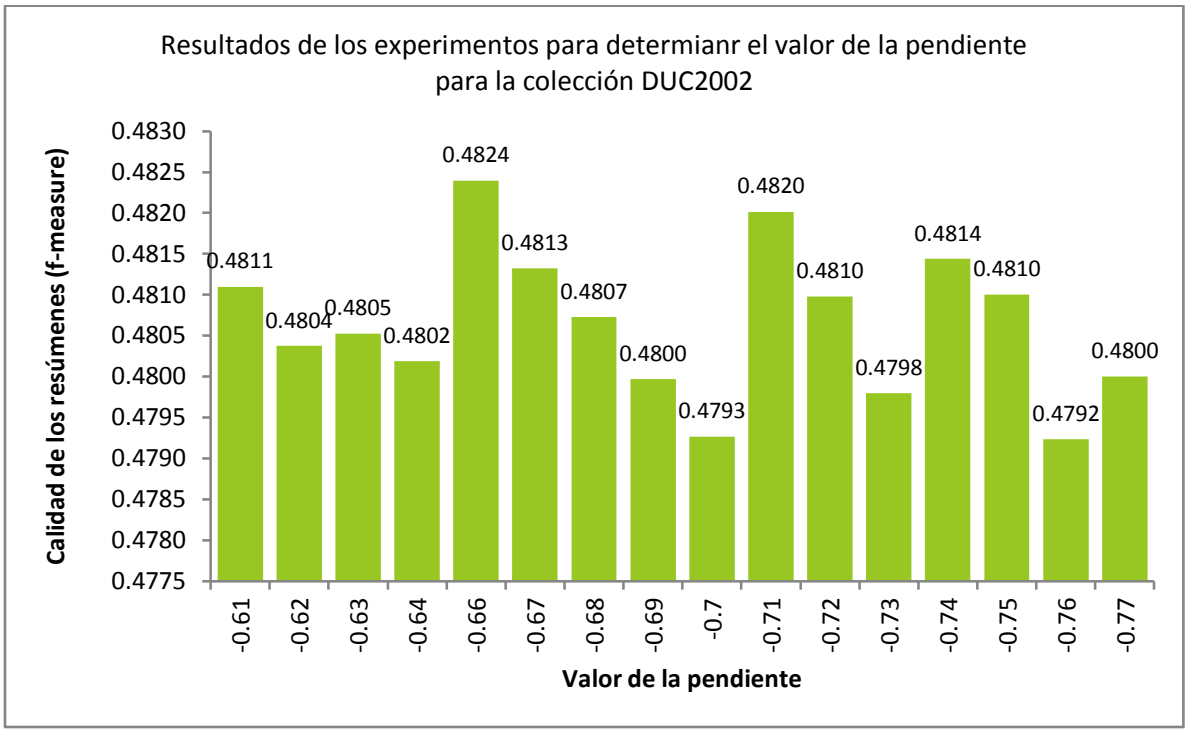

Figura 9. Resultados de los experimentos para determinar el valor de la pendiente para la colección DUC2002.

Como se puede observar en la Figura 9, el mejor resultado 0.4824 se obtuvo con la pendiente 0.66. Este es la configuración que se considera para el siguiente experimento.

5.3.2 Pre-procesamiento y modelo de texto

En la figura 10, se muestran los resultados obtenidos con la colección de documentos DUC2002 del lenguaje inglés para los parámetros pre-procesamiento y modelo de texto.

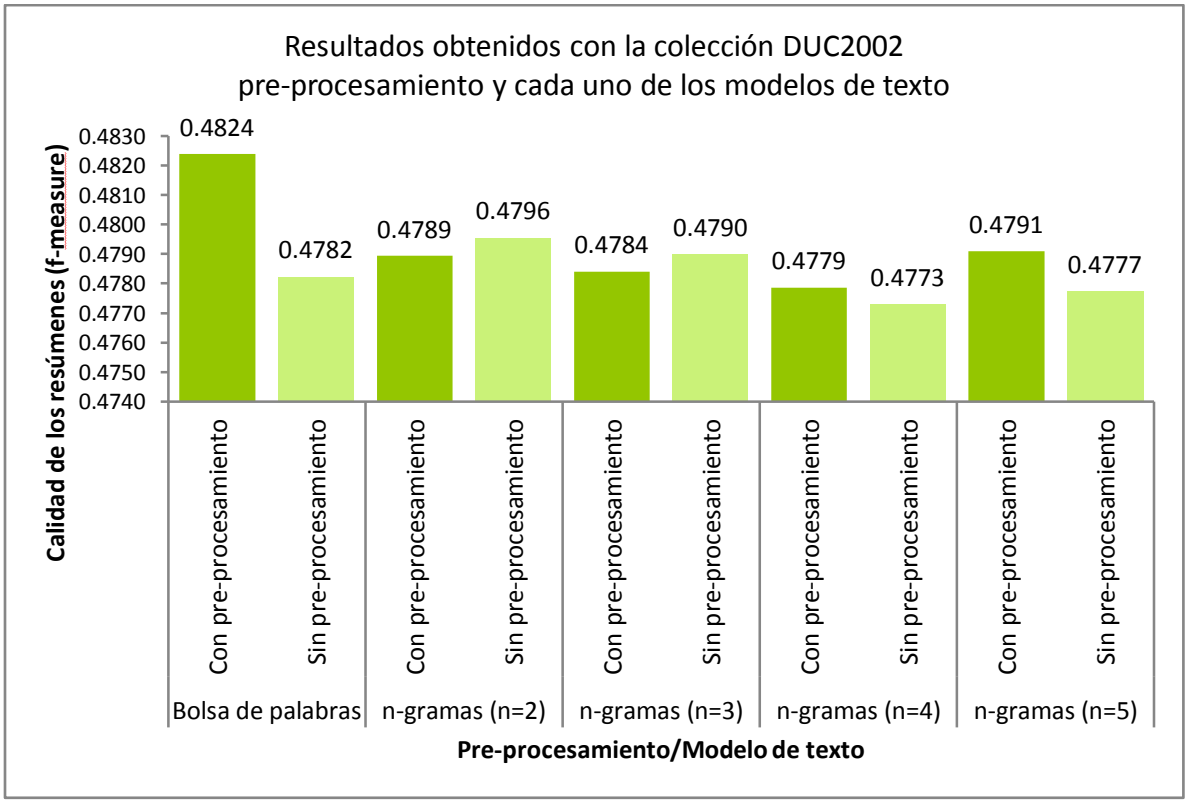

Figura 10. Resultados obtenidos con la colección en el lenguaje inglés con los parámetros pre-procesamiento y con el modelo de texto.

Como se puede observar los mejores resultados se obtuvieron pre-procesando la colección de documentos y utilizando el modelo de texto bolsa de palabras.

5.3.3 Importancia de las oraciones

En la figura 11, se muestran los resultados obtenidos con la colección de documentos DUC2002 para el lenguaje inglés en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

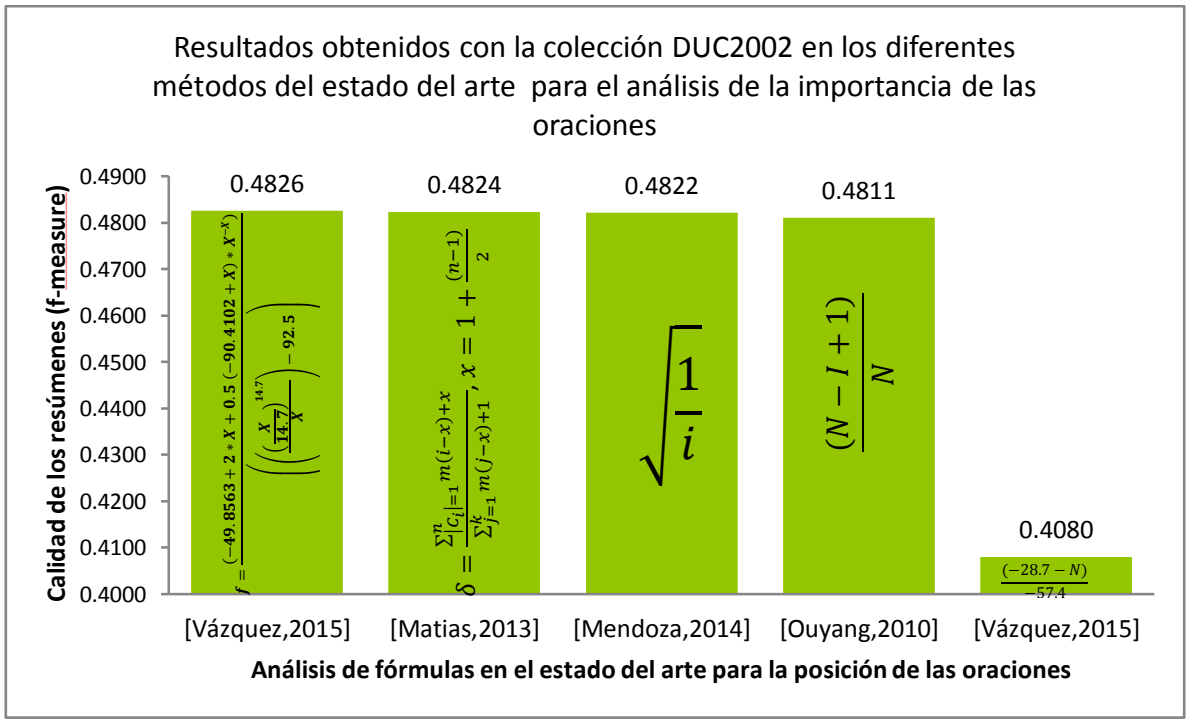

Figura 11. Resultados obtenidos con la colección en el lenguaje inglés en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

Como se puede observar en la gráfica anterior fórmula propuesta en el trabajo de (Vázquez, 2015) es la que obtiene mejores resultados. Cabe mencionar que para este experimento se ocuparon los mejores parámetros anteriores; bolsa de palabras y con pre-procesamiento.

5.3.4 Función de aptitud

En la figura 12, se muestran los resultados obtenidos con el lenguaje inglés con el ajuste de la función de aptitud.

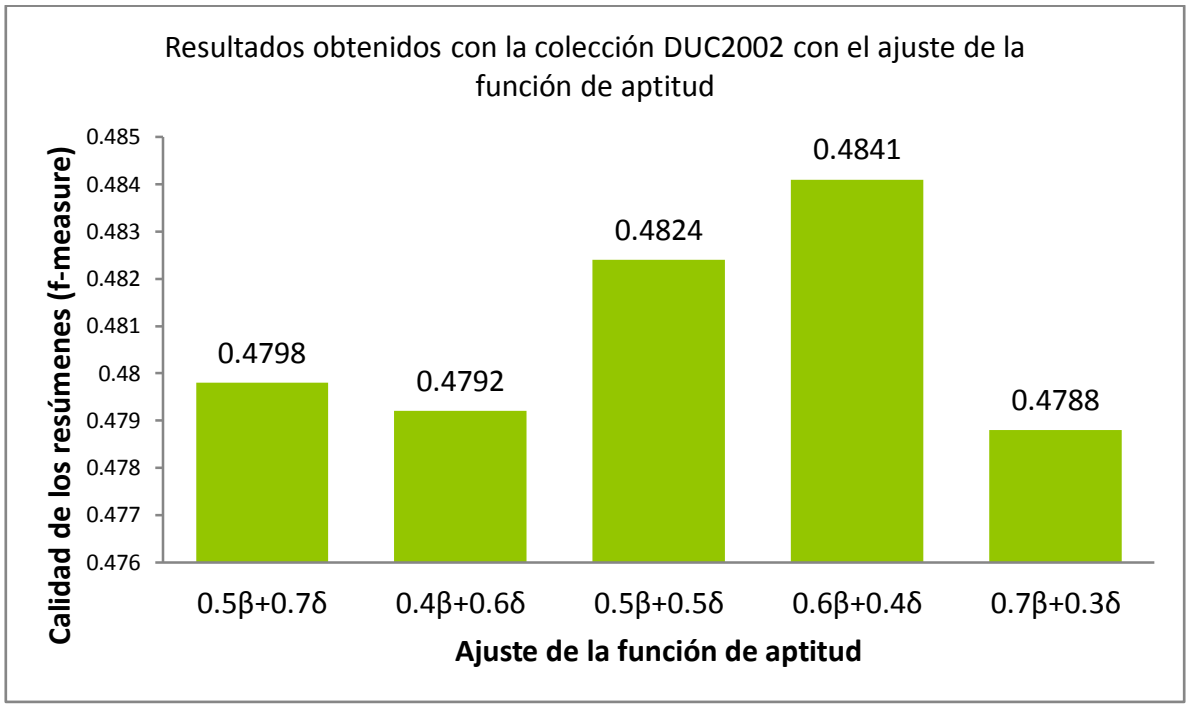

Figura 12. Resultados obtenidos con la colección en el lenguaje inglés con el ajuste de la función de aptitud.

Como se puede observar en la gráfica anterior los mejores resultados se obtuvieron dando mayor valor a la frecuencia de las oraciones $\boldsymbol{\beta = 0.6}$ y un menos a la característica de la posición de las oraciones $\boldsymbol{\delta = 0.4}$. Sin embargo, a pesar de que ligeramente la frecuencia de las oraciones debe tener mayor valor, no es mucha la diferencia entre los resultados obtenidos con la fusión de aptitud equilibrada. Para este experimento los parámetros utilizados son: pre-procesamiento, bolsa de palabras y se utiliza la fórmula propuesta por (Vázquez, 2015).

5.3.5 Operador de selección

En el trabajo de (Matias, 2013) se utilizó el operador de selección ruleta. Todos los resultados mostrados en las gráficas anteriores se realizaron utilizando el operador ruleta. Sin embargo, se tiene antecedente en el trabajo de (Vázquez, 2015) de que el operador de selección por torneo permitía mejorar los resultados del algoritmo genético, por ello se probó en este trabajo obteniendo los resultados mostrado en la Figura 13.

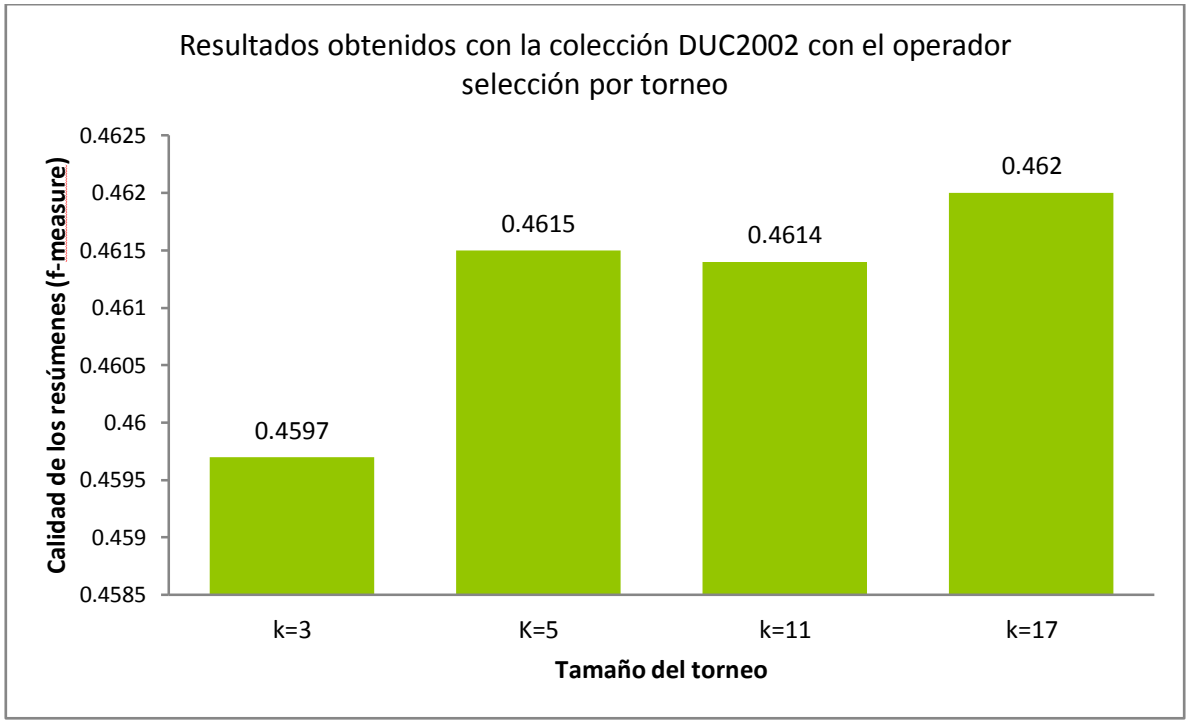

Figura 13. Resultados obtenidos con la colección en el lenguaje inglés con el operador de selección torneo.

En la gráfica anterior se puede observar que el resultado no supera a los anteriores donde se utilizó el operador de selección ruleta. Sin embargo, se muestran los resultados como evidencia que para el método de (Matias, 2013) el operador de selección por torneo no mejora la calidad de los resúmenes para el lenguaje inglés.

5.3.6 Comparación con los métodos del estado del arte y las herramientas comerciales

En la Figura 14, se muestran los resultados de los experimentos con el lenguaje inglés con las herramientas comerciales y los métodos del estado del arte. Los experimentos se realizaron utilizando la colección de documentos DUC2002 y se evaluaron con la herramienta ROUGE.

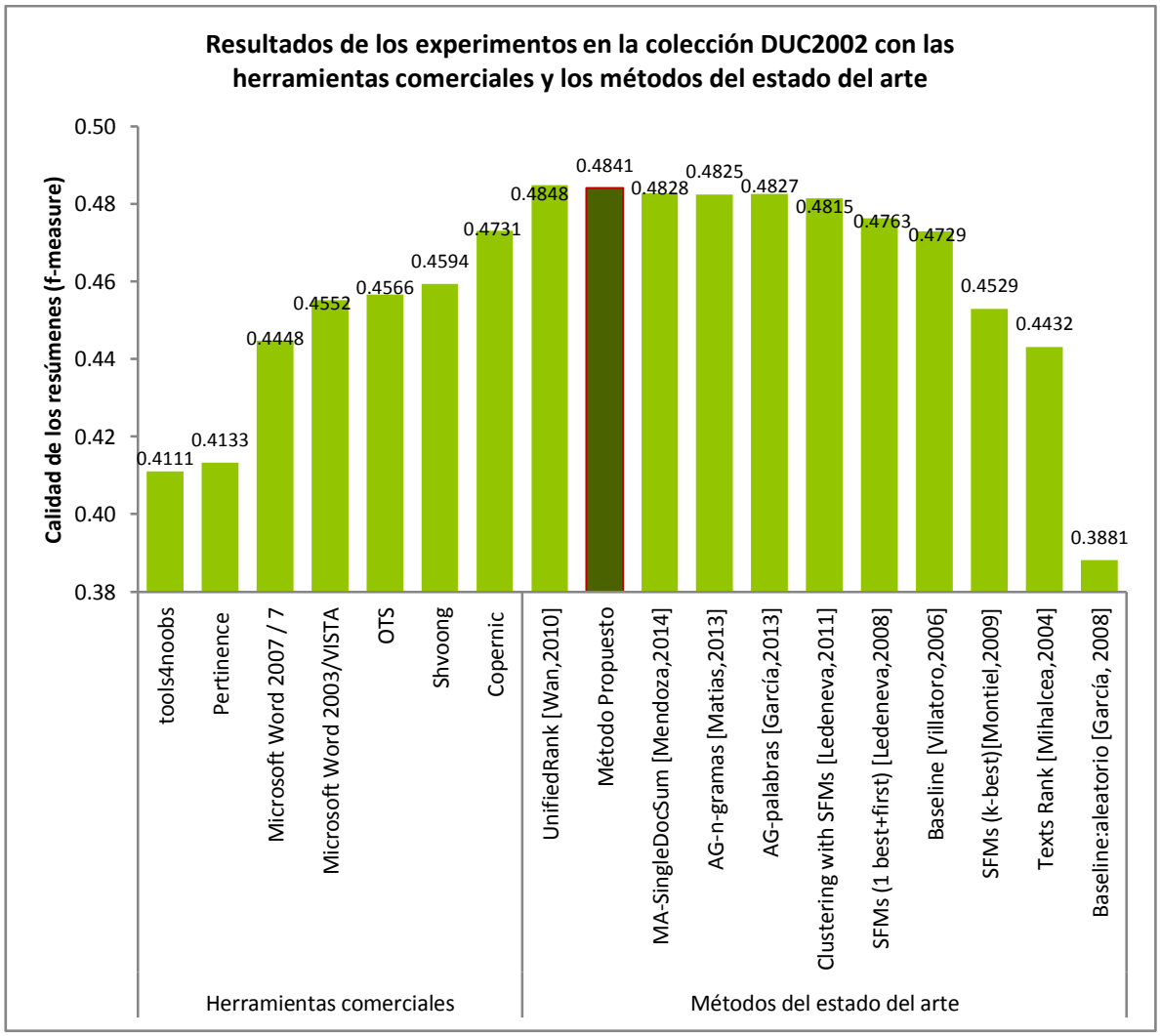

Figura 14. Resultados obtenidos con la colección en el lenguaje inglés con los métodos del estado del arte y las herramientas comerciales.

Como se puede observar el método propuesto supera a todas las herramientas comerciales y es uno de los mejores del estado del arte. La heurística baseline es la base para medir la calidad de los resúmenes, por lo que si un método la supera se puede decir que es un método de calidad. El método propuesto supera a la heurística baseline y la diferencia entre el mejor método UnifiedRank es muy poca, por lo que se puede decir que los dos métodos presentan resúmenes de calidad.

5.4 Resultados en lenguaje portugués (TeMário)

5.4.1 Resultados con el valor de la pendiente adecuado para cada modelo de texto

En la figura 15, se muestran los mejores resultados de cada modelo de texto y el valor de la pendiente que mejor se adapta a cada uno de ellos.

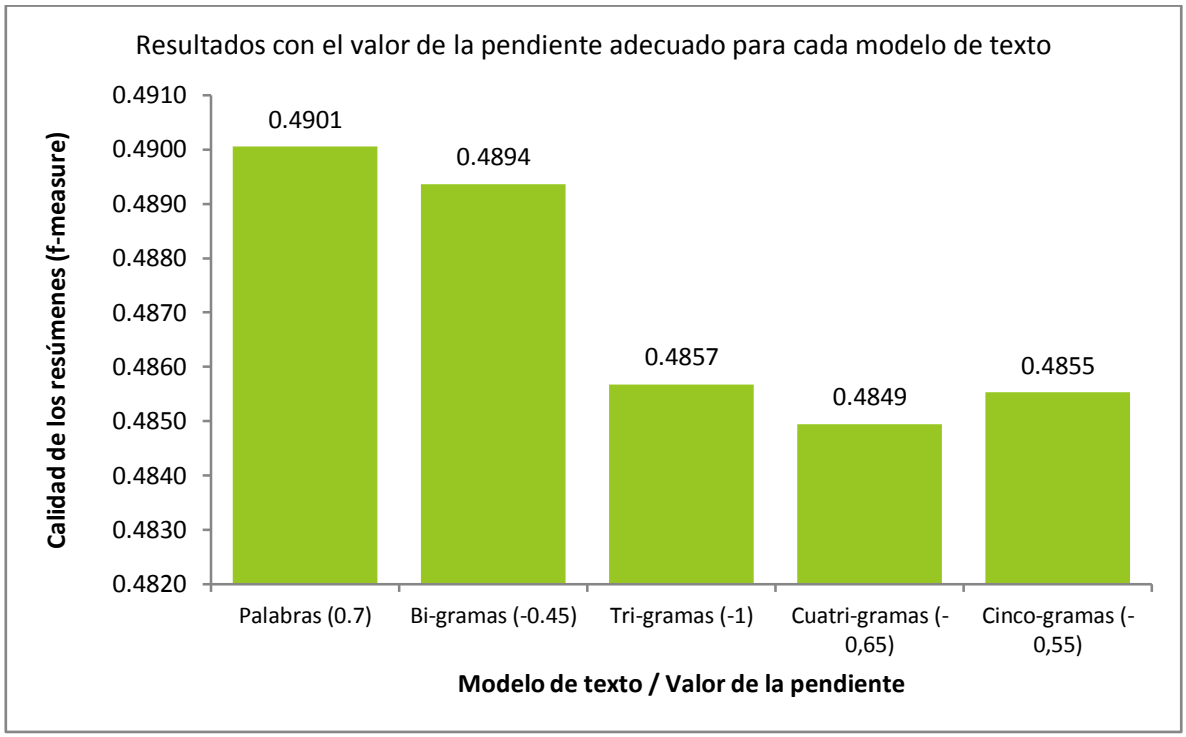

Figura 15. Resultados con el valor de la pendiente adecuado para cada modelo de texto.

Como se puede observar en la Figura 15, el mejor resultado 0.4901 se obtuvo con la pendiente 0.7. Este es la configuración que se considera para el siguiente experimento.

5.4.2 Pre-procesamiento y modelo de texto

En la figura 16, se muestran los resultados obtenidos con la colección de documentos TeMário del lenguaje portugués para los parámetros pre-procesamiento y modelo de texto.

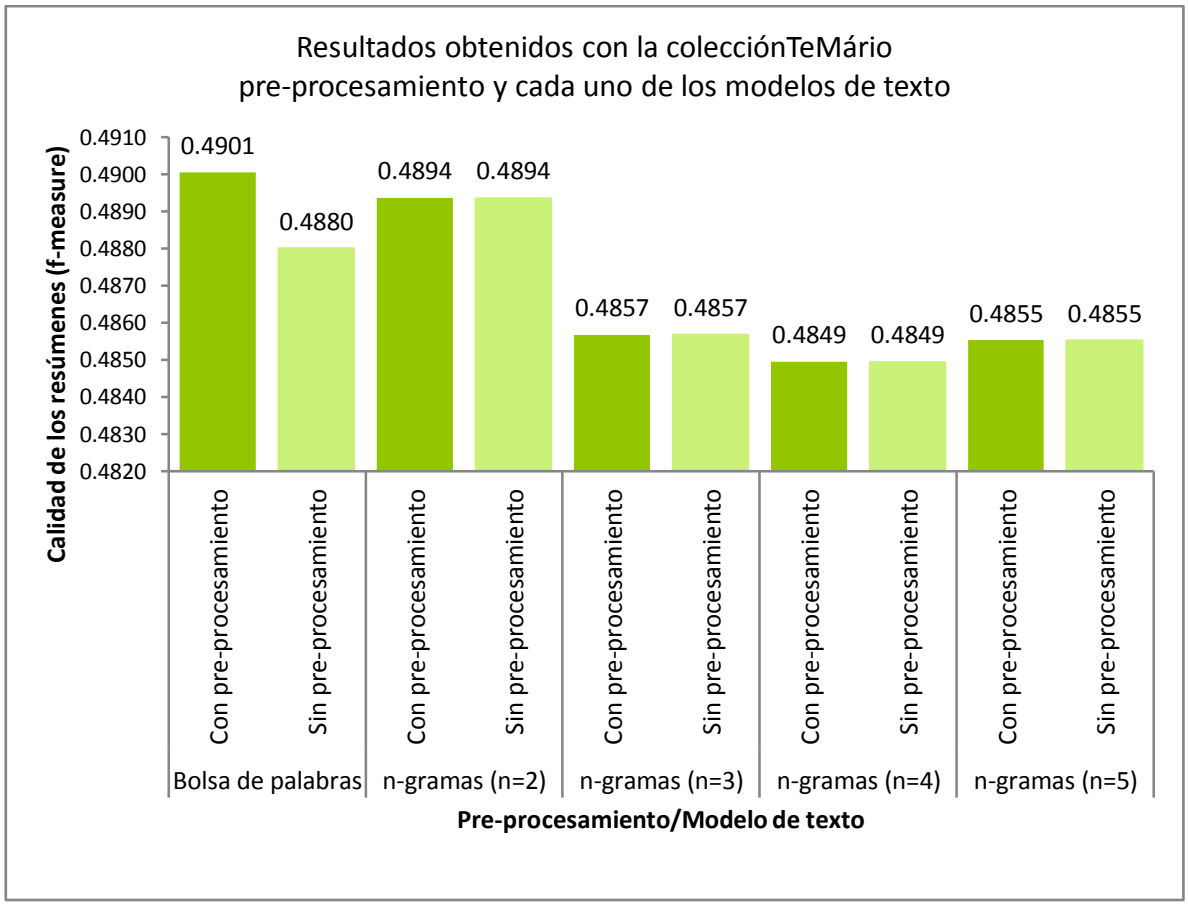

Figura 16. Resultados obtenidos con la colección en el lenguaje portugués con los parámetros pre-procesamiento y con el modelo de texto.

Como se puede observar los mejores resultados se obtuvieron pre-procesando la colección de documentos y utilizando el modelo de texto bolsa de palabras.

5.4.3 Importancia de las oraciones

En la figura 17, se muestran los resultados obtenidos con la colección de documentos TeMário para el lenguaje portugués en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

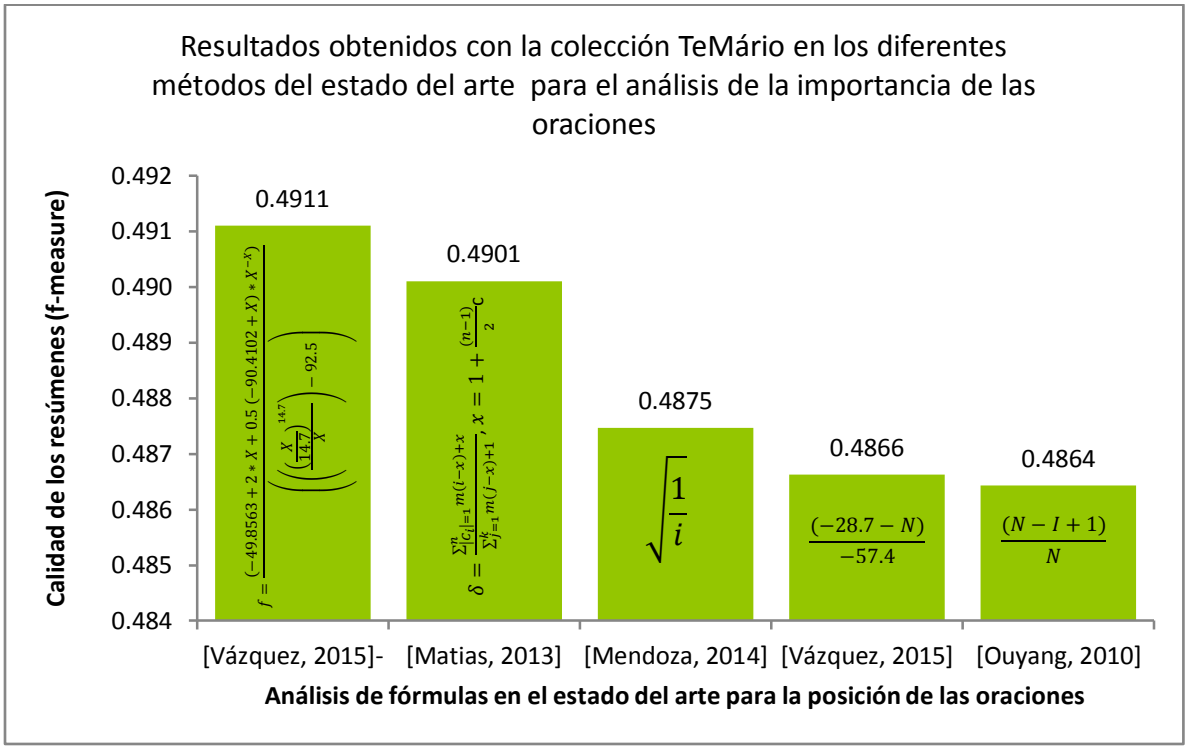

Figura 17. Resultados obtenidos con la colección en el lenguaje portugués en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

Como se puede observar en la gráfica anterior la fórmula propuesta en el trabajo de (Vázquez, 2015) es la que obtiene mejores resultados. Cabe mencionar, que para este experimento se ocuparon los mejores parámetros anteriores; bolsa de palabras y con pre-procesamiento.

5.4.4 Función de aptitud

En la figura 18, se muestran los resultados obtenidos con el lenguaje portugués con el ajuste de la función de aptitud.

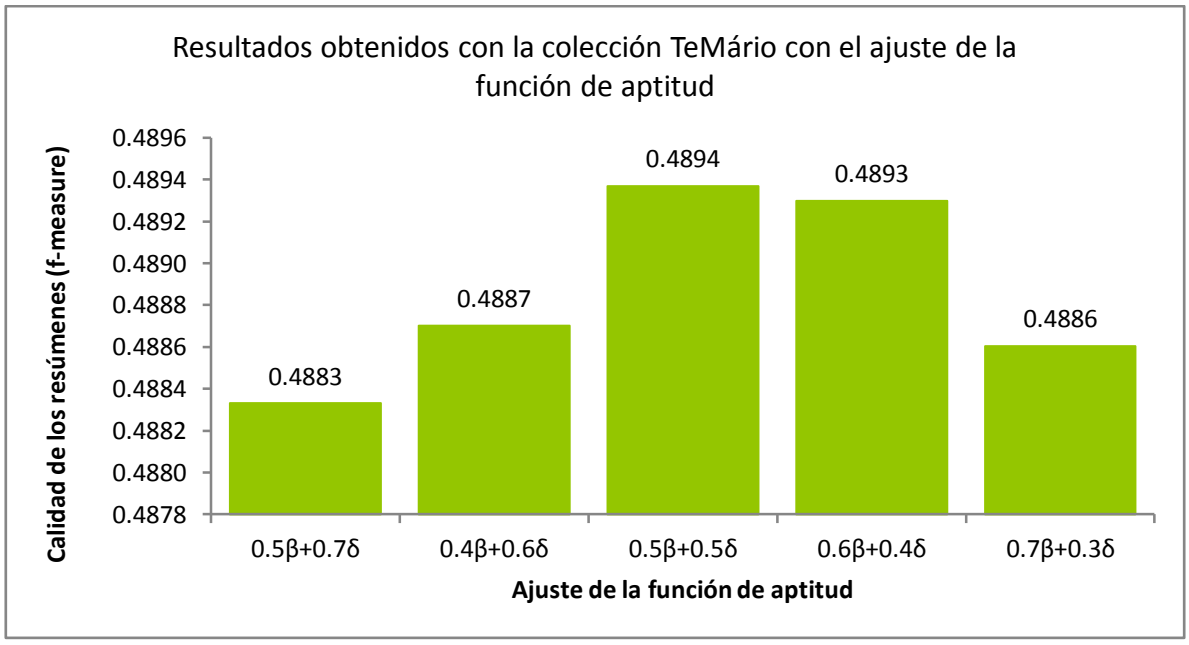

Figura 18. Resultados obtenidos con la colección en el lenguaje portugués con el ajuste de la función de aptitud.

Como se puede observar en la gráfica anterior los mejores resultados se obtuvieron manteniendo el equilibrio entre $\boldsymbol{\beta = 0.5}$ y $\boldsymbol{\delta = 0.5}$. Por lo que se puede decir que para el lenguaje portugués tanto la frecuencia de las oraciones como la posición de las oraciones tienen la misma importancia. Para este experimento los parámetros utilizados son: pre-procesamiento, bolsa de palabras y se utiliza la fórmula propuesta por (Vázquez, 2015).

5.4.5 Operador de selección

En el trabajo de (Matias, 2013) se utilizó el operador de selección ruleta. Todos los resultados mostrados en las gráficas anteriores se realizaron utilizando el operador ruleta. Sin embargo, se tenían antecedentes en el trabajo de (Vázquez, 2015) de que el operador de selección por torneo permitía mejorar los resultados del algoritmo genético, por lo que se probó en este trabajo obteniendo los resultados mostrado en la Figura 19.

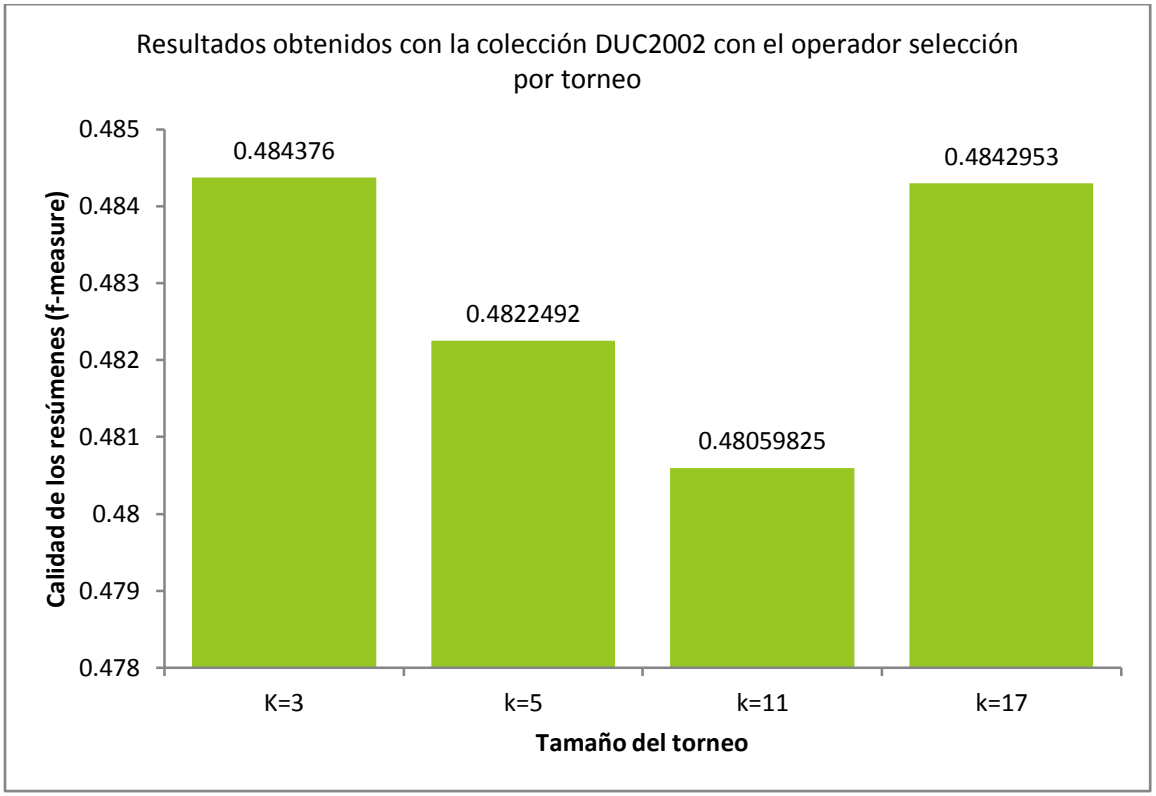

Figura 19. Resultados obtenidos con la colección en el lenguaje español con el operador de selección torneo.

En la gráfica anterior se puede observar que el resultado no supera a los anteriores donde se utilizó el operador de selección ruleta. Sin embargo, se muestran los resultados como evidencia que para el método de (Matias, 2013) el operador de selección por torneo no mejora la calidad de los resúmenes para el lenguaje portugués.

5.4.6 Comparación con los métodos del estado del arte y las herramientas comerciales

En la Figura 20, se muestran los resultados de los experimentos con el lenguaje portugués con las herramientas comerciales y los métodos del estado del arte. Los experimentos se realizaron utilizando la colección de documentos TeMário y se evaluaron con la herramienta ROUGE. Para la colección TeMário la extensión que se le puede dar a los resúmenes va de un 25 a 30 %. Los resultados de los experimentos mostrados se hicieron a una extensión de 30%.

En el estado del arte hay trabajos que prueban con la colección TeMário. Sin embargo, al tener un rango amplio para la longitud de los resúmenes, es más complicado determinar una medida base (baseline) estándar, por lo que para este trabajo se obtuvo la heurística baseline considerando una longitud del 30%.

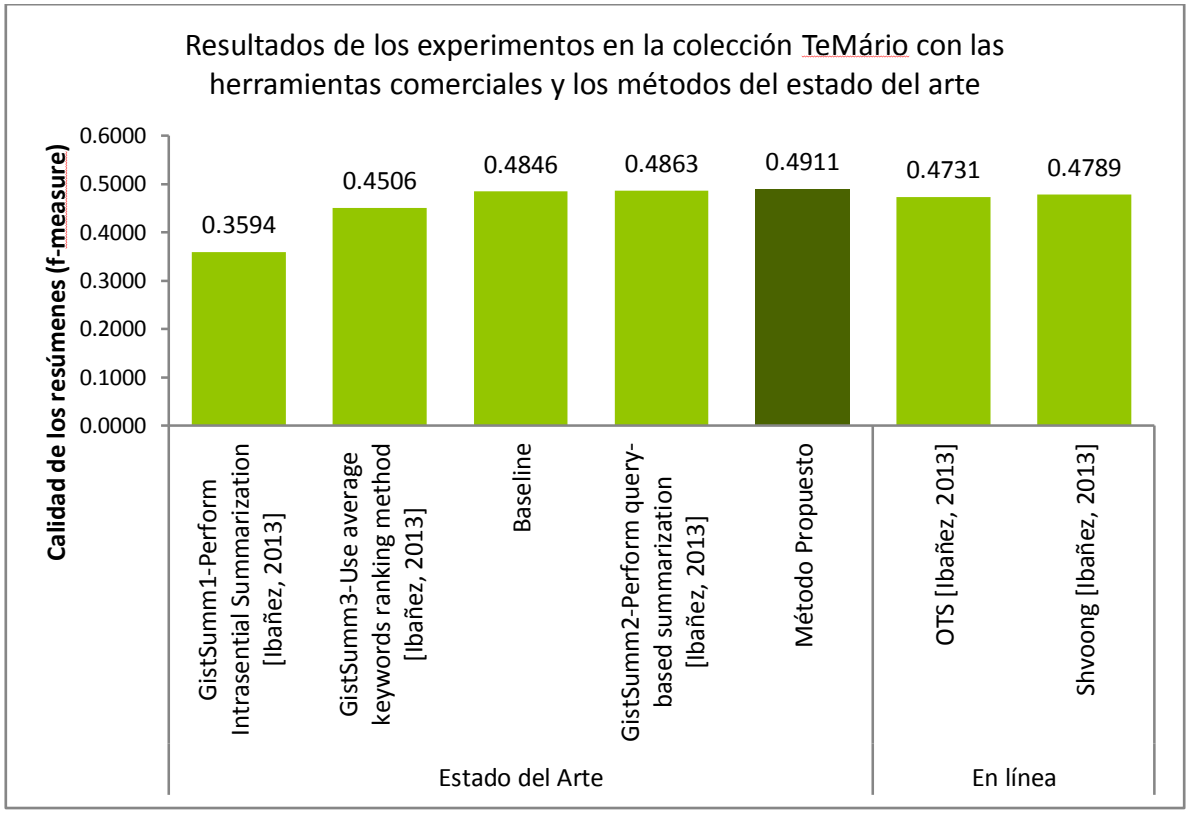

Figura 20. Resultados obtenidos con la colección en el lenguaje inglés con los métodos del estado del arte y las herramientas comerciales.

Como se puede observar el método propuesto supera a todas las herramientas comerciales en línea y obtiene los mejores resultados en el estado del arte.

En el estado del arte se encuentra el trabajo de (Mihalcea, 2005), en donde se prueba la colección TeMário. Los resultados obtenidos por el trabajo de Mihalcea superan a los obtenidos en este trabajo. Sin embargo, surge una problemática para comparar con este trabajo, debido a que el rango de extensión que se usa para esta colección no está bien definido, por lo que para verificar la extensión utilizada en Mihalcea se obtuvo el resultado de baseline para la colección TeMário en el rango establecido por la colección (25% al 30% de la extensión del documento). Pero habiendo probado con todas las extensiones posibles del 25% al 30% no se llegó al resultado de baseline obtenido por Mihalcea. Por lo que se piensa que la problemática está en la forma en que se están considerando las oraciones.

A continuación se muestra una tabla comparativa, entre los resultados de (Mihalcea, 2005) y los resultados del método propuestos en este trabajo. Baseline en el trabajo actual se calculó al 30%.

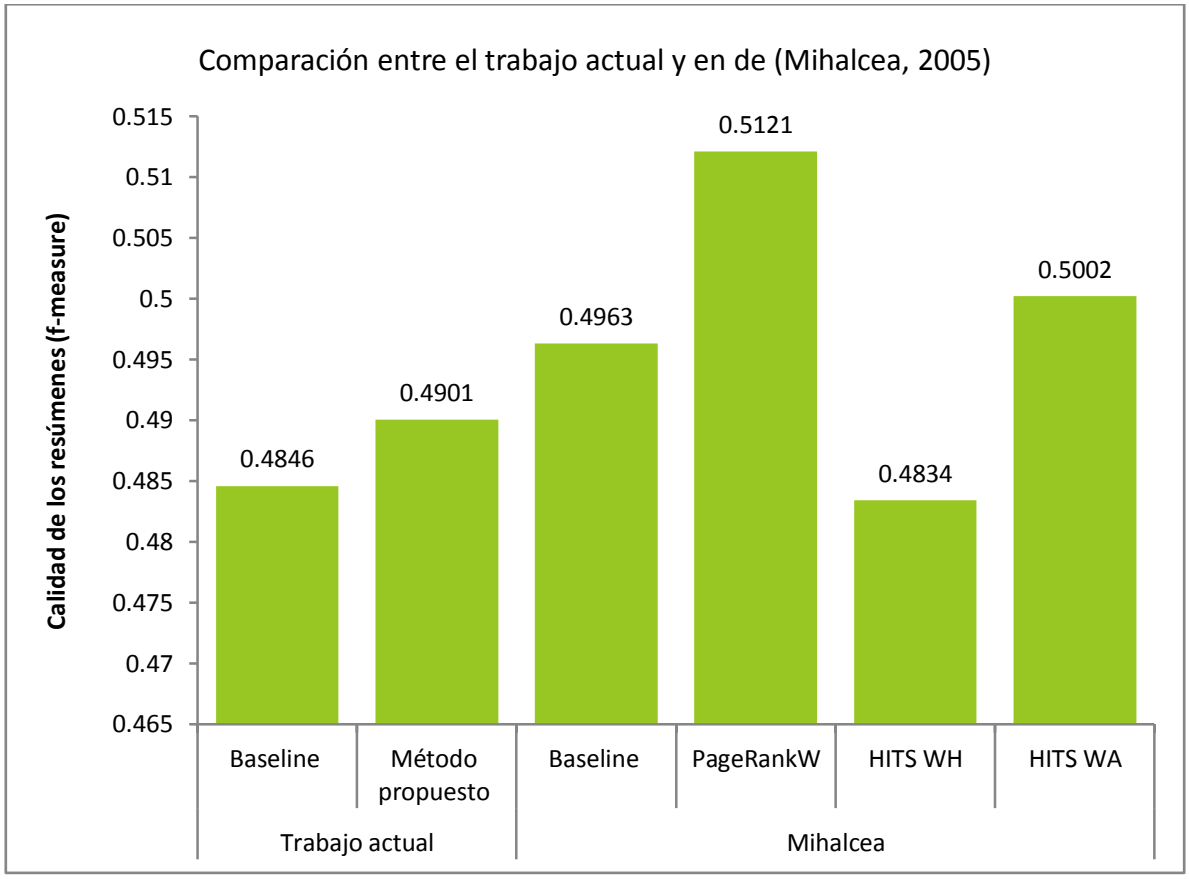

Figura 21. Comparación entre el trabajo actual y el de (Mihalcea, 2005).

Como se puede observar en la figura 21, los resultados de la heurística baseline son diferentes. También como se puede observar tanto el método presentado en este trabajo y el método de (Mihalcea, 2005) superan su baseline propuesto.

5.5 Resultados en lenguaje español (TER)

5.5.1 Resultados con el valor de la pendiente adecuado para cada modelo de texto

En la figura 22, se muestran los mejores resultados de cada modelo de texto y el valor de la pendiente que mejor se adapta a cada uno de ellos.

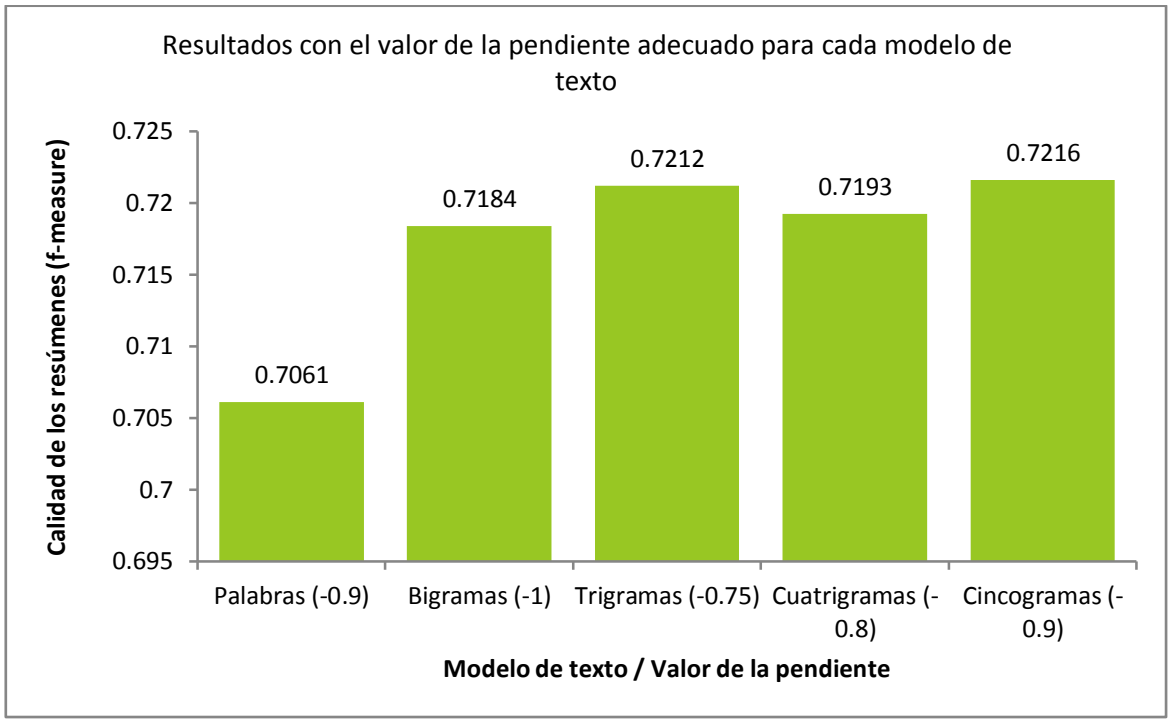

Figura 22. Resultados con el valor de la pendiente adecuado para cada modelo de texto.

Como se puede observar en la Figura 22, el mejor resultado 0.7216 se obtuvo con la pendiente 0.9. Este es la configuración que se considera para el siguiente experimento.

5.5.2 Pre-procesamiento y modelo de texto

En la figura 23, se muestran los resultados obtenidos con la colección de documentos TER del lenguaje español para los parámetros pre-procesamiento y modelo de texto.

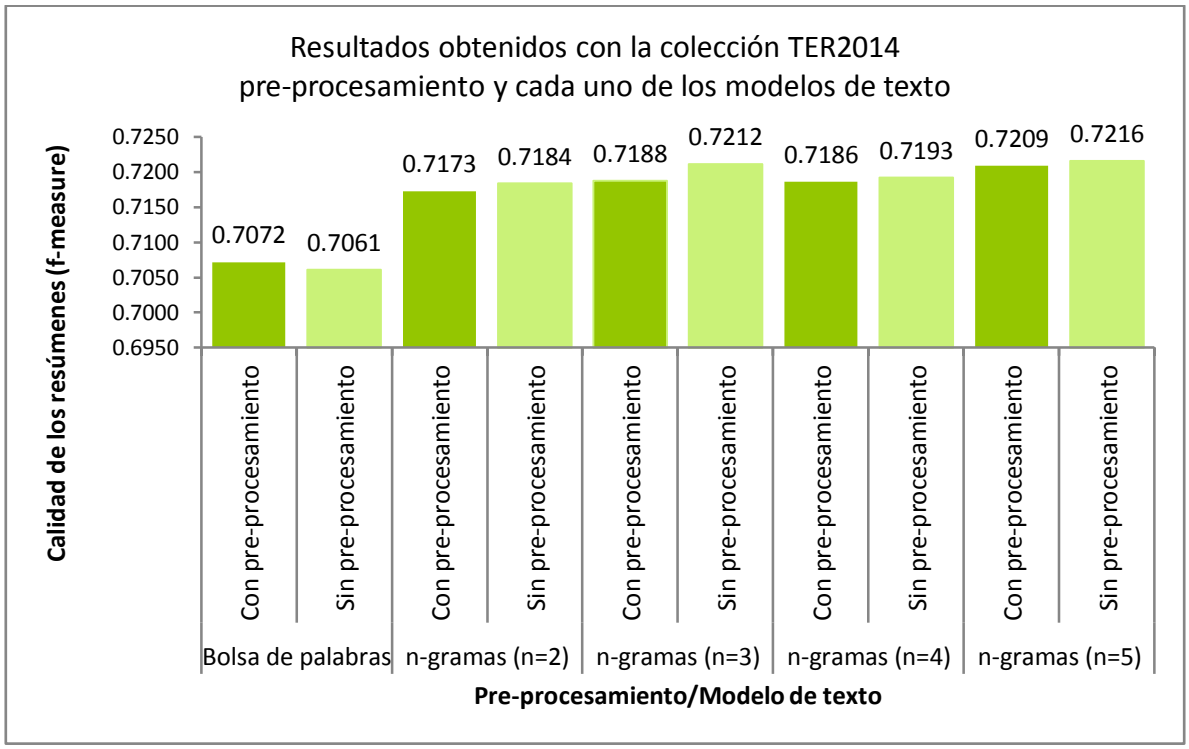

Figura 23. Resultados obtenidos con la colección en el lenguaje español con los parámetros pre-procesamiento y con el modelo de texto.

Como se puede observar los mejores resultados se obtuvieron sin pre-procesar la colección de documentos y utilizando el modelo de texto n-gramas con $n = 5$.

5.5.3 Importancia de las oraciones

En la figura 24, se muestran los resultados obtenidos con la colección de documentos TER para el lenguaje español en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

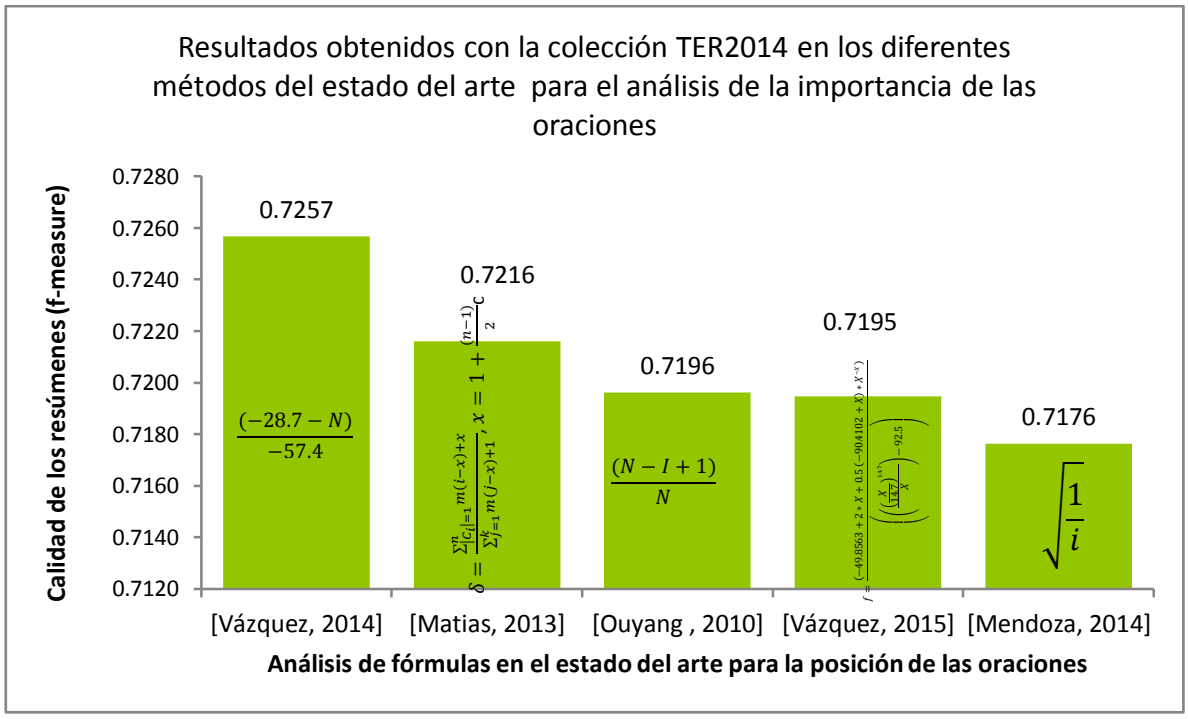

Figura 24. Resultados obtenidos con la colección en el lenguaje español en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

Como se puede observar en la gráfica anterior la fórmula propuesta en el trabajo de (Vázquez, 2015) es la que obtiene mejores resultados. Cabe mencionar, que para este experimento se ocuparon los mejores parámetros anteriores; n-gramas y sin pre-procesamiento.

5.5.4 Función de aptitud

En la figura 25, se muestran los resultados obtenidos con el lenguaje español con el ajuste de la función de aptitud.

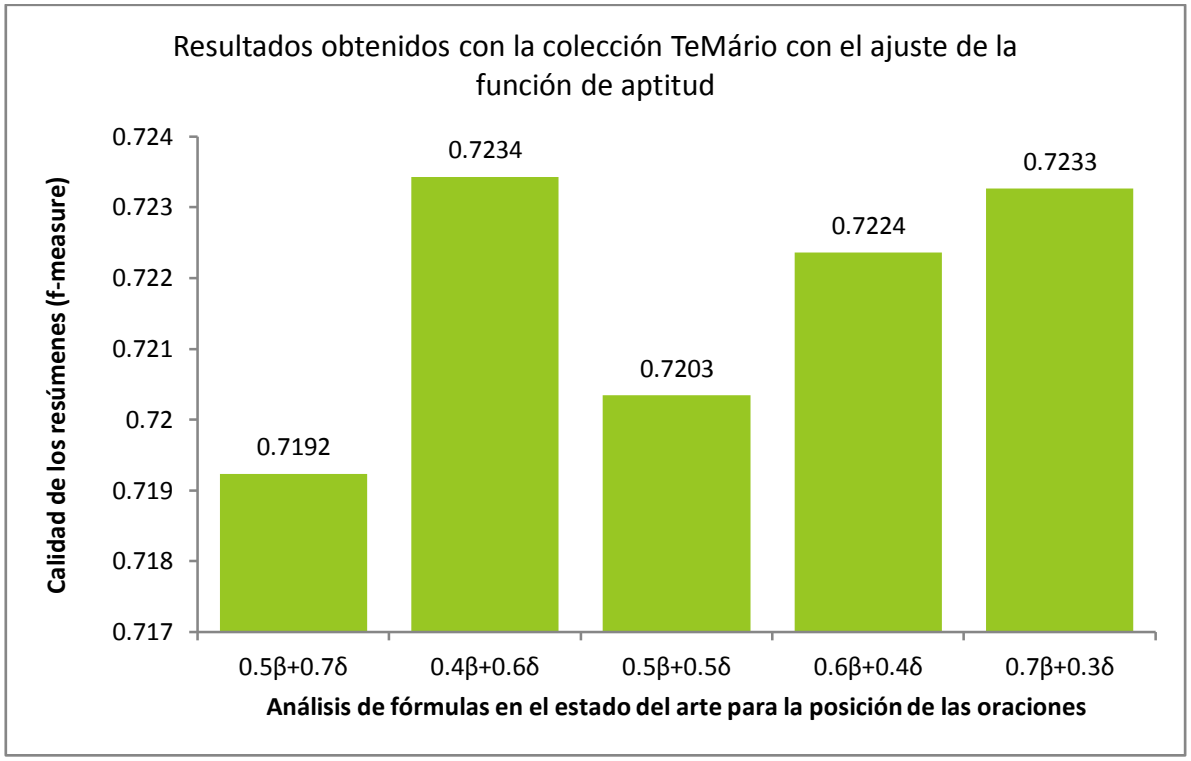

Figura 25. Resultados obtenidos con la colección en el lenguaje español con el ajuste de la función de aptitud.

Como se puede observar en la gráfica anterior los mejores resultados se obtuvieron con $\boldsymbol{\beta = 0.4}$ y $\boldsymbol{\delta = 0.6}$. Para el lenguaje español se puede ver que la importancia de la posición de las oraciones debe ser mayor, si se mantiene equilibrada la función los resultados bajan. Para este experimento los parámetros utilizados son: pre-procesamiento, bolsa de palabras y se utiliza la fórmula propuesta por (Vázquez, 2015).

.

5.5.5 Operador de selección

En el trabajo de (Matias, 2013) se utilizó el operador de selección ruleta. Todos los resultados mostrados en las gráficas anteriores se realizaron utilizando el operador ruleta. Sin embargo, se tenían antecedentes en el trabajo de (Vázquez, 2015) de que el operador de selección por torneo permitía mejorar los resultados del algoritmo genético, por lo que se probó en este trabajo obteniendo los resultados mostrado en la Figura 26.

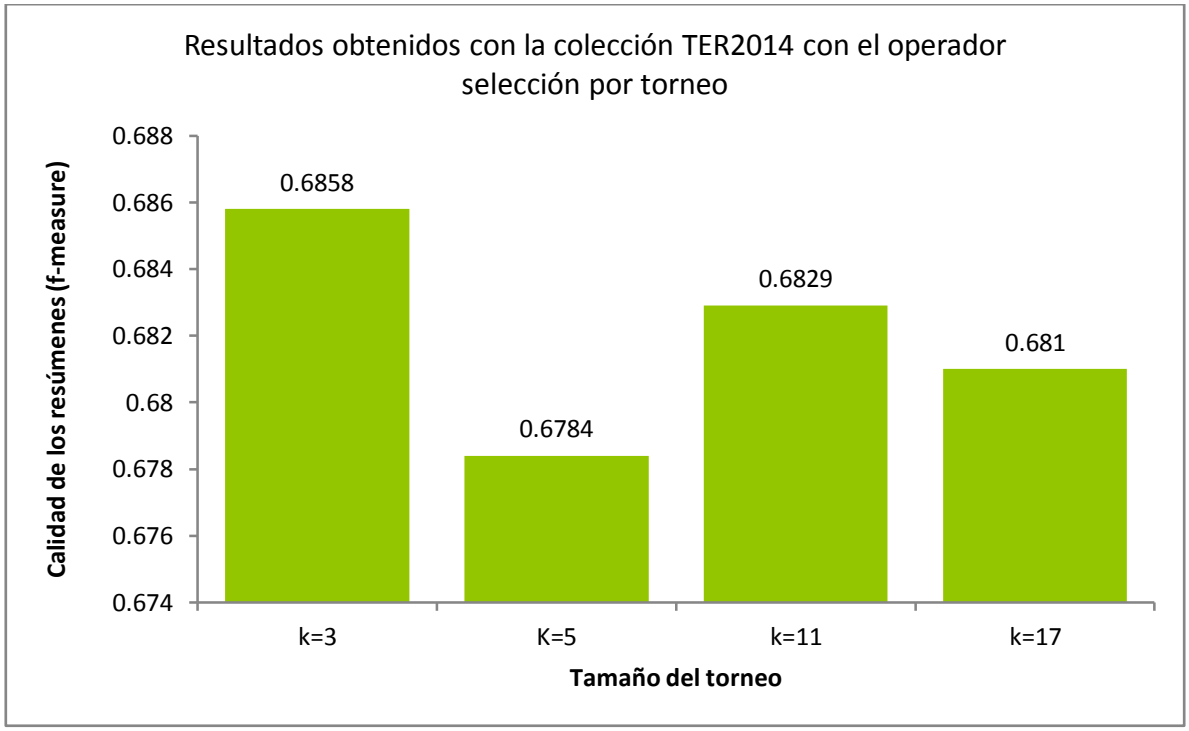

Figura 26. Resultados obtenidos con la colección en el lenguaje español con el operador de selección torneo.

En la gráfica anterior se puede observar que el resultado no supera a los anteriores donde se utilizó el operador de selección ruleta. Sin embargo, se muestran los resultados como evidencia que para el método de (Matias, 2013) el operador de selección por torneo no mejora la calidad de los resúmenes para el lenguaje español.

5.3.6 Comparación con los métodos del estado del arte y las herramientas comerciales

En la Figura 27, se muestran los resultados de los experimentos con el lenguaje español, con las herramientas comerciales y los métodos del estado del arte. Los experimentos se realizaron utilizando la colección de documentos TER y se evaluaron con la herramienta ROUGE.

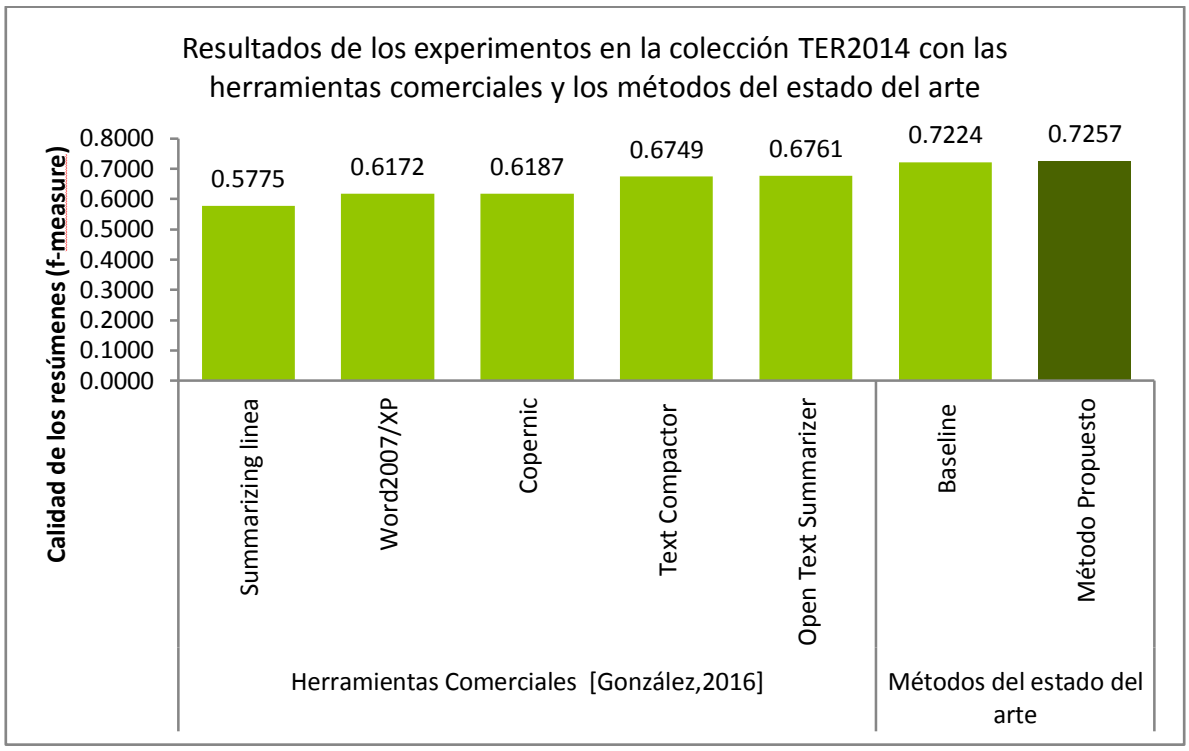

Figura 27. Resultados obtenidos con la colección en el lenguaje español con los parámetros pre-procesamiento y con el modelo de texto

Como se puede observar el método propuesto supera los resultados obtenidos con las herramientas comerciales instalables y en línea.

CAPÍTULO 6

Conclusiones y Trabajo Futuro

En este capítulo, se presentan las conclusiones generales del trabajo de investigación, así como las aportaciones principales.

6.1. Conclusiones

- Para saber que el método propuesto en (Matias, 2013) es independiente del lenguaje, se realizaron pruebas en tres colecciones de diferentes lenguajes (Inglés, Portugués y Español).
- Considerando los resultados obtenidos en los diferentes lenguajes se fueron ajustando sus parámetros.
- Los resultados obtenidos en todos los lenguajes fueron superados tanto para las herramientas comerciales como para los métodos del estado del arte por lo que se puede concluir que el método (Matias, 2013) es un método competitivo y robusto.

6.1.1 Conclusiones para el lenguaje inglés

En la tabla 12, se muestra la lista de parámetros modificados y el mejor valor obtenido de cada uno de ellos para el lenguaje inglés.

Tabla 12. Parámetros para el lenguaje inglés (DUC2002)

	Parámetros
Pre-procesamiento	Si
Modelo de texto	Bolsa de palabras
Importancia de las oraciones	[Vázquez,2015]
Función de aptitud	0.6β+0.4δ
Operador de selección	Ruleta

6.1.2 Conclusiones para el lenguaje portugués

En la tabla 13, se muestra la lista de parámetros modificados y el mejor valor obtenido de cada uno de ellos para el lenguaje portugués.

Tabla 13. Parámetros para el lenguaje portugués (TeMário).

Parámetros	
Pre-procesamiento	Si
Modelo de texto	Bolsa de palabras
Importancia de las oraciones	[Vázquez,2015]
Función de aptitud	0.5β+0.5δ
Operador de selección	Ruleta

6.1.1 Conclusiones para el lenguaje español

En la tabla 14, se muestra la lista de parámetros modificados y el mejor valor obtenido de cada uno de ellos para el lenguaje español.

Tabla 14. Parámetros para el lenguaje español (TER).

Parámetros	
Pre-procesamiento	No
Modelo de texto	n-gramas (*n*=5)
Importancia de las oraciones	[Vázquez,2015]
Función de aptitud	0.4β+0.6δ
Operador de selección	Ruleta

TER es un nuevo reto para la generación de resúmenes ya que el baseline es muy alto y el reto es superarlo.

6.2. Aportaciones

Una de las principales aportaciones de este trabajo es la creación de corpus TER (español mexicano) especial para resúmenes.

6.3 Trabajo futuro

- Agregar más parámetros al método
    - Longitud de las oraciones
    - Relación de las oraciones con el título
- Mejorar los resultados en el lenguaje español
- Poner a disposición de los investigadores el corpus TER para que lo puedan probar y de esta manera obtener valores que sirvan como referencia para la comparación de los resultados conseguidos con la realización del presente trabajo.
- Probar el método, utilizando n-gramas sintácticos propuestos por (Sidorov, 2014)

6.4 Publicaciones derivadas

Griselda Areli Matías Mendoza, Yulia Ledeneva, René García-Hernández. Evaluación de Herramientas Comerciales, Herramientas en Línea y Métodos del Estado del Arte para la Generación de Resúmenes de Textos para un solo Documento. Research in Computer Science. ISSN: 1870-4069, pp. 265-274, volumen 70, 2013.

Referencias.

Alfonseca, 2003	Alfonseca, E., & Rodríguez, P. (2003). Generating extracts with genetic algorithms. Springer-Verlag, 2633, 511-519.
Araujo, 2009	Araujo, L., & Cervigón, C. (2009). Un enfoque práctico. México D.F.: Alfaomega.
Arco, 2006	Arco, L., Bello, R., Mederos, J. & Pereza, Y. (2006). Agrupamiento de Documentos Textuales mediante Métodos Concatenados. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 10, pp. 43-53.
Berker, 2011	Berker, M. (2011). Using genetic algorithms with lexical chains for automatic text summarization. Thesis B.S., Computer Engineering.
Blair-Goldensohn, 2004	Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., Mckeown,k., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Siddharthan, A. & Siegelman, S. (2004). Columbia University at DUC 2004. In Proceedings of DUC'04, pp. 23–30.
Bouayad-Agha, 2009	Bouayad-Agha, N., G. Casamayor, G. Ferraro, S. Mille, V. Vidal, y L. Wanner. (2009). Improving the Comprehension of Legal Documentation: The Case of Patent Claims. En Proceedings of the International Conferenceon Artificial Intelligence and Law, pp. 78–87
Copernic Inc	Copernic Technologies Inc. (s.f.). 2000-2003, Marca registrada ©. Obtenido de Copernic Technologies Inc.: http://www.copernic.com/en/products/summarizer.
Da Cunha, 2008	Da Cunha Fanego, I. (2008). Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.
EcuRed, 2013	EcuRed. (2010). N-grama. 20 de mayo de 2013, de EcuRed. Conocimiento para todos y para todos Sitio web:

http://www.ecured.cu/index.php/N-grama

Edyburn, 2010 — Text Compactor © 2010-2014 Conocimiento by Design, Inc. Keith Edyburn [en línea] http://textcompactor.com/.

García, 2008 — García R., Montiel, R., Ledeneva, Y., Rendón, e., Gelbukh, A. & Cruz, R. (2008). Text Summarization by Senten ce Extraction Using Unsupervised Learning. 7° Conferencia Internacional Mexicana de Inteligencia Artificial (MICAI08); Notas de la conferencia de Inteligencia Artificial, Springer-Verlag, Vol 5317, pp133-143.

García, 2009 — García, R., Ledeneva, Y., Matias G., Hernández, A., Chávez, J., Gelbukh, A & Tapia, J. (2009). Comparing Commercial Tools and State-of-the-Art Methods for Generating Text Summaries. IEEE Computer Society Press, pp.92-96.

García, 2013 — García, R. & Ledeneva, Y. (2013). Single Extractive Text Summarization Based on a Genetic Algorithm. Springer Link, Lecture Notes in Computer Science, Vol. 7914, pp. 374-378.

González, 2016 — González, N. Evaluación de las herramientas comerciales de generación automática de resúmenes de texto para el idioma español. México: Tesis de licenciatura en proceso. Universidad Autónoma del Estado de México.

Hovy, 1999 — Hovy, E. y C. Lin (1999). Automated Text Summarization in SUMMARIST. En I. Mani y M. Maybury (eds.). Advances in Automatic Text Summarization. Cambridge: MIT. pp. 81-94.

INEGI, 2015 — INEGI. (2015). ESTADÍSTICAS A PROPÓSITO DEL… DÍA MUNDIAL DEL INTERNET (17 DE MAYO). Instituto Nacional de Estadística y Geografía, Vol 1, pp. 1-9.

Klingberg, 2009 — Klingberg, T. (2009). The Overflowing Brain Information Overload and the Limits of Working Memory. Stockholm: Oxford University Press Inc.

Kuri, 2007 — Kuri, Á., & Galaviz, J. (2007). Algoritmos Genéticos. México: Sociedad Mexicana de Inteligencia Artificial México.

Last, 2010	Last, M. & Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. NATO Science for Peace and Security Series - D: Information and Communication Security. Vol. 27: Web Intelligence and Security, pp. 207-237.
Ledeneva, 2008	Ledevena, Y. N. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. México, D.F.: Presentada en el Instituto Politécnico Nacional, para obtención del grado de Doctor.
Ledeneva, 2008a	Ledeneva, Y., Gelbukh, A. & García, R. (2008). Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. Research in Computing Science, Vol. 34, pp.163-174.
Ledeneva, 2008b	Ledeneva, Y., Gelbukh, A., García, R. Terms Derived from Frequent Sequences for Extractive Text Summarization. Springer-Verlag, ISBN 978-3-540-78134-9, ISSN 0302-9743, 2008. DOI 10.1007/978-3-540-78135-6_51. pp. 593-604
Ledeneva, 2011	Ledeneva, Y., García, R., Montiel, R., Cruz, R. & Gelbukh, A. (2011). EM Clustering Algorithm for Automatic Text Summarization. Springer-Verlag LNAI 7094, pp. 305-315.
Lin, 2004	Lin, C. (2004). ROUGE: A package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization of ACL.
Lo Cen, 2012	Lo Cen, J.. (2012). Procesamiento sobre texto. 21 de mayo de 2013, de Universidad de Costa Rica Facultad de Ingeniería Escuela de ciencias de la computación e Informática Sitio web: https://sites.google.com/site/ptriexamenfinal2012/about
Luhn, 1958	Luhn, H. (1958). The automatic creation of literature abstracts. IBM Journal of Research andDevelopment, Vol. 2, pp. 159-165
Márquez, 2010	Márquez G. (2010). Resúmenes automáticos: Enfoque extractivo y evaluación. 23 de mayo de 2013. De Escuela Politécnica Superior. Universidad Autónoma de Madrid. Sitio web: https://vimaco.files.wordpress.com/2010/05/resumenes-

automaticos.pdf

Matias, 2013 — Matias, G. (2013), Generación automática de resúmenes usando algoritmos genéticos, Presentada en la Unidad Académica Profesional Tianguistenco de la Universidad Autónoma del Estado de México, para obtención del Título de Ingeniera en Software.

Mei, 2015 — Mei, S., Guan, G., Wang, Z., Wan, S., He, M. & Degan, D.. (2015). Video summarization via minimum sparse reconstruction. Pattern Recognition, Vol. 48, pp. 522-533.

Mendoza, 2014 — Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., & León, E. (2014). *Extractive single-document summarization based on genetic operators and guided local search*. Vol. 41(No. 9), pp. 4158–4169.

Microsoft ® Office Word 2003 — Microsoft ® Office Word 2003 (11.8307.8221) SP3. (s.f.). Parte de Microsoft Office Professional Edition. Obtenido de 2003 Copyright © 1983-2003 Microsoft Corporation.

Microsoft ® Office Word 2007 — Microsoft ® Office Word 2007 (12.0.4518.1014) MSO . (s.f.). Parte de Microsoft Office Professional 2007 © 2006 Microsoft Corporation.

Mihalcea, 2005 — Mihalcea, R. & Taran, P.. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Vol. 1, pp. 602-607.

Módolo,2003 — Módolo, M. (2003). SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português. Master thesis. Departamento de Computação, UFSCar.

Montiel, 2009 — Montiel, R. (2009). Generación automática de resúmenes mediante aprendizaje no supervisado. Edo. de México: Presentada en el Instituto Tecnológico de Toluca, para obtención del Título de Ingeniero en Sistemas Computacionales.

Nichols, 2012	Nichols, J., Mahmud, J. & Drews, C.. (2012). Summarizing sporting events using twitter. IUI '12 Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, Vol.1, pp. 189-198
OTS, 2007	OTS. Open Text Summarizer. Página principal de OTS . [En línea] http://libots.sourceforge.net/.
Ouyang, 2010	Ouyang, Y., Li, W., Lu, Q., & Zhang, R. (2010). *A study on position information in document summarization*. pp. 919-927.
Parajes, 2006	Parajes G & Santos M. (2006). Inteligencia Artificial e Ingeniería del conocimiento. España, Madrid: Alfaomega Rama.
Pardo, 2003	Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal. June 26-27.
Patel, 2007	Patel, A., Siddiqui, T & Tiwary, U. (2007). A language independent approach to multilingual text summarization. Conference RIA2007, Pittsburgh PA, U.S.A., 123-132.
Pertinence, 2009	Pertinence. Pertinence Summarizer. Página principal de pertinence . [En línea] http://pertinence.net/index_en.html.
Plaza, 2010	Plaza, L. (2010). Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo. Madrid, España: Presentada en la Facultad de Informática de la Universidad Compútense de Madrid, para obtención del grado de Doctor.
Porter, 1980	Porter, M. (1980). An algorith, for suffix stripping. vol 40, pp. 211-218.
Qazvinian, 2008	Qazvinian , V., Sharif Hassanabadi, L., & Halavatu, R.. (2008). Summarising text with a genetic algorithm-based sentence

extraction. Int. J. Knowledge Management Studies, 1, pp. 426-444.

Qazvinian, 2013 — Qazvinian, V., Radev, D., Mohammad, S., Dorr, B.,, Zajic, D., Whidby, M. & Moon, T.. (2014). Generating Extractive Summaries of Scientific Paradigms. Journal Of Artificial Intelligence Research, Vol. 46, pp. 165-201.

Ramírez, 2007 — Kriscia Daviana Ramírez Benavides. (2007). Stemming-Lematización. 21 de mayo de 2013. Obtenido de Escuela de Ciencias de la Computación e Informática; Universidad de Costa Rica: Sitio web: http://www.ecci.ucr.ac.cr/~kramirez/RI/Material/Presentaciones/Stemming.pdf.

Saggion, 2008 — Saggion, H. (2008). SUMMA: A Robust and Adaptable Summarization Tool.Revue Traitement Automatique des Langues, Vol.49(2). pp103–125.

Saggion, 2011 — Saggion, H. Using SUMMA for Language Independent Summarization at TAC 2011. In Text Analysis Conference TAC 2011, USA.

Sidorov, 2014 — Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A. & Chanona-Hernández, L. (2014) *Syntactic N-grams as machine learning features for natural language processing*. Expert Systems with Applications, vol. 41, issue 3, pp. 853–860.

Suanmali, 2011 — Saunmali, L., Salim, N., & Mohammed, S. (2011). Genetic algorithm based sentence extraction for text summarization. International Journal of Innovative Computing, Vol. 1, pp. 2180-4370.

Summarizer, 2016 — Article Summarizer Online © 2016 Página principal de summarizing [En línea] http://www.summarizing.biz/best-summarizing-strategies/article-summarizer-online/.

Svhoong, 2005 — Svhoong. Svhoong Summarizer. Página principal de Svhoong. [En línea] http://es.shvoong.com/summarizer/.

Telléz, 2009 — Téllez, A., Montes, M. & Villaseñor, L.. (2009). Using Machine Learning for Extracting Information from Natural Disaster News

Reports. Computación y Sistemas. Instituto Politécnico Nacional México, vol. 13, pp. 33-44.

Tools4noobs, 2007 — Tools4noobs. Tools4noobs Summarizer. Página principal de tools4noobs. [En línea] http://www.tools4noobs.com/summarize/.

Vázquez, 2015 — Vázquez, E. Modelo de relevancia de la posición de las oraciones en resúmenes de texto, mediante regresión simbólica. México: Tesis de licenciatura; Universidad Autónoma del Estado de México.

Villatoro,2006 — Villatoro E. (2006). Generación automática de resúmenes de múltiples documentos. Puebla: Tesis de maestría; Instituto Nacional de Astrofísica, Óptica y Electrónica.

Wan — Wan, X. (2010). Towards a unified approach to simultaneous single-document and multi-document summarization. In proceeding of the 23rd international conference on computational linguistics, pp. 1137-1145.

Wang, 2013 — Wang, L. & Cardie, C. (2013). Domain-Independent Abstract Generation for Focused Meeting Summarization. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1395--1405.

Yahiaoui, 2003 — Yahiaoui, I., Merialdo, B., & Huet, B. (2003). Comparison of Multiepisode Video Summarization Algorithms. EURASIP Journal on Applied Signal Processing, 1, pp. 48-55.

Yang, 2011 — Yang, Z., Cai, K., Tang, J., Su, Z & Li, J.. (2011). Social context summarization. SIGIR '11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, Vol. 1, pp. 255-264.

Zhu, 2004 — Zhu, M. (2004), "Recall, precision, and average precision," Dept. Statistics Actuarial Sci., Univ. Waterloo, CA, Tech. Rep. 9.

Anexo 1.

Lista de palabras vacías para la colección en español

UN, UNA, UNAS, UNOS, UNO, SOBRE, TODO, TAMBIÉN, TRAS, OTRO, ALGÚN, ALGUNO, ALGUNA, ALGUNOS, ALGUNAS, SER, ES, SOY, ERES, SOMOS, SOIS, ESTOY, ESTA, ESTAMOS, ESTAIS, ESTAN, COMO, EN, PARA, ATRÁS, PORQUE, POR QUÉ, ESTADO, ESTABA, ANTE, ANTES, SIENDO, AMBOS, PERO, POR, PODER, PUEDE, PUEDO, PODEMOS, PODEIS, PUEDEN, FUI, FUE, FUIMOS, FUERON, HACER, HAGO, HACE, HACEMOS, HACEIS, HACEN, CADA, FIN, INCLUSO, PRIMERO, DESDE, CONSEGUIR, CONSIGO, CONSIGUE, CONSIGUES, CONSEGUIMOS, CONSIGUEN, IR, VOY, VA, VAMOS, VAIS, VAN, VAYA, GUENO, HA, TENER, TENGO, TIENE, TENEMOS, TENEIS, TIENEN, EL, LA, LO, LAS, LOS, SU, AQUÍ, MIO, TUYO, ELLOS, ELLAS, NOS, NOSOTROS, VOSOTROS, VOSOTRAS, SI, DENTRO, SOLO, SOLAMENTE, SABER, SABES, SABE, SABEMOS, SABEIS, SABEN, ULTIMO, LARGO, BASTANTE, HACES, MUCHOS, AQUELLOS, AQUELLAS, SUS, ENTONCES, TIEMPO, VERDAD, VERDADERO, VERDADERA, CIERTO, CIERTOS, CIERTA, CIERTAS, INTENTAR, INTENTO, INTENTA, INTENTAS, INTENTAMOS, INTENTAIS, INTENTAN, DOS, BAJO, ARRIBA, ENCIMA, USAR, USO, USAS, SA, USAMOS, USAIS, USAN, EMPLEAR, EMPLEO, EMPLEAS, EMPLEAN, AMPLEAMOS, EMPLEAIS, VALOR, MUY, ERA, ERAS, ERAMOS, ERAN, MODO, BIEN, CUAL, CUANDO, DONDE, MIENTRAS, QUIEN, CON, ENTRE, SIN, TRABAJO, TRABAJAR, TRABAJAS, TRABAJA, TRABAJAMOS, TRABAJAIS, TRABAJAN, PODRIA, PODRIAS, PODRIAMOS, PODRIAN, PODRIAIS, YO, AQUEL

Anexo 2.

Lista de palabras vacías para la colección en inglés

A, ABLE, ABOUT, ABOVE, ACCORDING, ACCORDINGLY, ACROSS, ACTUALLY, AFTER, AFTERWARDS, AGAIN, AGAINST, AIN'T, ALL, ALLOW, ALLOWS, ALMOST, ALONE, ALONG, ALREADY, ALSO, ALTHOUGH, ALWAYS, AM, AMONG, AMONGST, AN, AND, ANOTHER, ANY, ANYBODY, ANYHOW, ANYONE, ANYTHING, ANYWAY, ANYWAYS, ANYWHERE, APART, APPEAR, APPRECIATE, APPROPRIATE, ARE, AREN'T, AROUND, AS, ASIDE, ASK, ASKING, ASSOCIATED, AT, AVAILABLE, AWAY, AWFULLY, B, BE, BECAME, BECAUSE, BECOME, BECOMES, BECOMING, BEEN, BEFORE, BEFOREHAND, BEHIND, BEING, BELIEVE, BELOW ,BESIDE, BESIDES, BEST, BETTER, BETWEEN, BEYOND, BOTHBRIEF, BUT, BY,C, C'MON, C'S, CAME, CAN, CAN'T, CANNOT, CANT,, CAUSE, CAUSES, CERTAIN, CERTAINLY, CHANGES, CLEARLY, CO, COM, COME, COMES, CONCERNING, CONSEQUENTLY, CONSIDER, CONSIDERING, CONTAIN, CONTAINING, CONTAINS, CORRESPONDING, COULD, COULDN'T, COURSE, CURRENTLY, D, DEFINITELY, DESCRIBED, DESPITE, DID, DIDN'T, DIFFERENT, DO, DOES, DOESN'T, DOING, DON'T, DONE, DOWN ,DOWNWARDS, DURING, E, EACH ,EDU ,EG ,EIGHT ,EITHER, ELSE, ELSEWHERE, ENOUGH, ENTIRELY, ESPECIALLY, ET, ETC , EVEN, EVER ,EVERY ,EVERYBODY ,EVERYONE, EVERYTHING, EVERYWHERE, EX, ,EXACTLY, EXAMPLE, EXCEPT, F ,FAR ,FEW ,FIFTH ,FIRST , FIVE ,FOLLOWED ,FOLLOWING ,FOLLOWS ,FOR, FORMER, FORMERLY, FORTH, FOUR, FROM, FURTHER, FURTHERMORE ,G, GET ,GETS, GETTING, GIVEN, GIVES, GO, GOES , GOING, GONE, GOT, GOTTEN, GREETINGS ,H, HAD , HADN'T, HAPPENS, HARDLY, HAS, HASN'T, HAVE, HAVEN'T ,HAVING ,HE ,HE'S, HELLO, HELP, HENCE ,HER, HERE, HERE'S, HEREAFTER ,HEREBY ,HEREIN ,HEREUPON , HERS, HERSELF ,HI ,HIM, HIMSELF, HIS, HITHER, HOPEFULLY, HOW, HOWBEIT, HOWEVER, I, I'D, I'LL, I'M, I'VE, IE ,IF, IGNORED, IMMEDIATE, IN, INASMUCH, INC, INC., INDEED, INDICATE, INDICATED, INDICATES, INNER, INSOFAR, INSTEAD, INTO, INWARD, IS, ISN'T, IT, IT'D, IT'LL, IT'S, ITS, ITSELF, J, JUST, K, KEEP, KEEPS, KEPT, KNOW, KNOWS, KNOWN, L, LAST,, LATELY ,LATER,

LATTER, LATTERLY, LEAST, LESS, LEST, LET, LET'S, LIKE, LIKED, LIKELY, LITTLE, LOOK, LOOKING, LOOKS, LTD, M, MAINLY, MANY, MAY, MAYBE, ME, MEAN, MEANWHILE, MERELY, MIGHT, MORE, MOREOVER, MOST, MOSTLY, MUCH, MUST, MY, MYSELF, N, NAME, NAMELY, ND, NEAR, NEARLY, NECESSARY, NEED, NEEDS, NEITHER, NEVER, NEVERTHELESS, NEW, NEXT, NINE, NO, NOBODY, NON, NONE, NOONE, NOR, NORMALLY, NOT, NOTHING, NOVEL, NOW, NOWHERE, O, OBVIOUSLY, OF, OFF, OFTEN, OH, OK, OKAY, OLD, ON, ONCE, ONE, ONES, ONLY, ONTO, OR, OTHER, OTHERS, OTHERWISE, OUGHT, OUR, OURS, OURSELVES, OUT, OUTSIDE, OVER, OVERALL, OWN, P, PARTICULAR, PARTICULARLY, PER, PERHAPS, PLACED, PLEASE, PLUS, POSSIBLE, PRESUMABLY, PROBABLY, PROVIDES, Q, QUE, QUITE, QV, R, RATHER, RD, RE, REALLY, REASONABLY, REGARDING, REGARDLESS, REGARDS, RELATIVELY, RESPECTIVELY, RIGHT, S, SAID, SAME, SAW, SAY, SAYING, SAYS, SECOND, SECONDLY, SEE, SEEING, SEEM, SEEMED, SEEMING, SEEMS, SEEN, SELF,SELVES, SENSIBLE, SENT, SERIOUS, SERIOUSLY, SEVEN, SEVERAL, SHALL, SHE, SHOULD, SHOULDN'T, SINCE, SIX, SO, SOME, SOMEBODY,, SOMEHOW, SOMEONE, SOMETHING, SOMETIME, SOMETIMES, SOMEWHAT, SOMEWHERE, SOON, SORRY,, SPECIFIED, SPECIFY, SPECIFYING, STILL, SUB, SUCH, SUP, SURE, T, T'S, TAKE, TAKEN,, TELL, TENDS, TH, THAN, THANK, THANKS, THANX, THAT, THAT'S, THAT'S, THE, THEIR, THEIRS, THEM, THEMSELVES, THEN, THENCE, THERE, THERE'S, THEREAFTER, THEREBY, THEREFORE, THEREIN, THERES, THEREUPON, THESE, THEY, THEY'D, THEY'LL, THEY'RE, THEY'VE, THINK,, THIRD, THIS, THOROUGH, THOROUGHLY, THOSE, THOUGH, THREE, THROUGH, THROUGHOUT, THRU, THUS, TO, TOGETHER, TOO, TOOK, TOWARD, TOWARDS, TRIED, TRIES, TRULY, TRY, TRYING, TWICE, TWO, U, UN, UNDER, UNFORTUNATELY, UNLESS, UNLIKELY, UNTIL, UNTO, UP, UPON, US, USE, USED, USEFUL, USES, USING, USUALLY, UUCP, V, VALUE, VARIOUS, VERY, VIA, VIZ, VS, W, WANT, WANTS, WAS, WASN'T, WAY, WE, WE'D, WE'LL, WE'RE, WE'VE, WELCOME,, WELL, WENT, WERE, WEREN'T, WHAT, WHAT'S, WHATEVER, WHEN, WHENCE, WHENEVER, WHERE, WHERE'S, WHEREAFTER, WHEREAS, WHEREBY, WHEREIN, WHEREUPON, WHEREVER, WHETHER, WHICH, WHILE, WHITHER, WHO, WHO'S, WHOEVER, WHOLE, WHOM, WHOSE, WHY, WILL, WILLING, WISH, WITH, WITHIN, WITHOUT, WON'T, WONDER, WOULD, WOULDN'T, X, Y, YES, YET, YOU, YOU'D, YOU'LL, YOU'RE, YOU'VE, YOUR, YOURS, YOURSELF, YOURSELVES, Z, ZERO

Anexo 3.

Lista de palabras vacías para la colección en portugués

DE, A, O, QUE, E, DO, DA, EM, UM, PARA, COM, NÃ£O, UMA, OS, NO, SE, NA, POR, MAIS, AS, DOS, COMO, MAS, AO, ELE, DAS, Ã, SEU, SUA, OU, QUANDO, MUITO, NOS, JÃ¡, EU, TAMBÃ©M, SÃ³, PELO, PELA, ATÃ©, ISSO, ELA, ENTRE, DEPOIS, SEM, MESMO, AOS, SEUS, QUEM, NAS, ME, ESSE, ELES, VOCÃª, ESSA, NUM, NEM, SUAS, MEU, Ã S, MINHA, NUMA, PELOS, ELAS, QUAL, NÃ³S, LHE, DELES, ESSAS, ESSES, PELAS, ESTE, DELE, TU,TE ,VOCÃªS, VOS, LHES , MEUS, MINHAS, TEU, TUA, TEUS, TUAS, NOSSO, NOSSA, NOSSOS, NOSSAS, DELA, DELAS, ESTA, ESTES, ESTAS, AQUELE, AQUELA, AQUELES, AQUELAS, ISTO, AQUILO, ESTOU, ESTÃ¡, ESTAMOS, ESTÃ£O, ESTIVE, ESTEVE, ESTIVEMOS, ESTIVERAM, ESTAVA, ESTÃ¡VAMOS, ESTAVAM, ESTIVERA, ESTIVÃ©RAMOS, ESTEJA, ESTEJAMOS, ESTEJAM, ESTIVESSE, ESTIVÃ©SSEMOS, ESTIVESSEM, ESTIVER, ESTIVERMOS, ESTIVEREM, HEI, HÃ¡, HAVEMOS, HÃ£O, HOUVE, HOUVEMOS, HOUVERAM, HOUVERA, HOUVÃ©RAMOS, HAJA, HAJAMOS, HAJAM, HOUVESSE, HOUVÃ©SSEMOS, HOUVESSEM, HOUVER, HOUVERMOS, HOUVEREM, HOUVEREI, HOUVERÃ¡, HOUVEREMOS, HOUVERÃ£O, HOUVERIA, HOUVERÃAMOS, HOUVERIAM, SOU, SOMOS, SÃ£O, ERA, Ã©RAMOS, ERAM, FUI, FOI, FOMOS, FORAM, FORA, FÃ´RAMOS, SEJA, SEJAMOS, SEJAM, FOSSE, FÃ´SSEMOS, FOSSEM, FOR, FORMOS, FOREM, SEREI, SERÃ¡, SEREMOS, SERÃ£O, SERIA, SERÃAMOS, SERIAM, TENHO, TEM, TEMOS, TÃ©M, TINHA, TÃNHAMOS, TINHAM, TIVE, TEVE, TIVEMOS, TIVERAM, TIVERA, TIVÃ©RAMOS, TENHA, TENHAMOS, TENHAM, TIVESSE, TIVÃ©SSEMOS, TIVESSEM, TIVER, TIVERMOS, TIVEREM, TEREI, TERÃ¡, TEREMOS, TERÃ£O, TERIA, TERÃAMOS, TERIAM

Anexo 4.

Documentación de la colección en español para procesamiento del lenguaje natural

Contenido

Introducción

Según el Diccionario Manual de la Lengua Española (Corpus, (n.d.)) un corpus es un conjunto extenso de textos de diversas clases, ordenados y clasificados que sirven como base de una investigación.

El corpus generado, tiene como fin poder servir de apoyo para el área del procesamiento del lenguaje natural en el idioma español.

Desarrollo del CORPUS.

El CORPUS está compuesto por noticias en español, del periódico "Crónica". Las noticias fueron descargadas de la página oficial de éste periódico www.cronica.com.mx.

Actualmente se cuentan con las noticias de los años 2012, 2013 y 2014

Los archivos descargados estaban en formato html, en la Figura 1, se muestra un ejemplo de noticia en formato página web.

Figura 1. Ejemplo de noticia en formato .html

En la Figura 2, se muestra un ejemplo de noticia en código html.

```
<! http://www.cronica.com.mx/notas/2014/825232.html >
 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link rel="shortcut icon" type="image/x-icon" href="../../img/favicon.ico" />
<link rel="stylesheet" type="text/css" href="../../css/style.css" />
<script type="text/javascript" src="https://apis.google.com/js/plusone.js">
  {lang: 'es'}
</script>
<meta property="fb:app_id" content="306656926115103"/>
<div id="fb-root"></div>
<!-- CODIGO DEL FACEBOOK-->
<script>(function(d, s, id) {
  var js, fjs = d.getElementsByTagName(s)[0];
  if (d.getElementById(id)) return;
  js = d.createElement(s); js.id = id;
  js.src = "//connect.facebook.net/es_LA/all.js#xfbml=1&appId=306656926115103";
  fjs.parentNode.insertBefore(js, fjs);
}(document, 'script', 'facebook-jssdk'));</script>
<!--FINALIZA CODIGO DEL FACEBOOK-->
<script type="text/javascript" src="../../js/letra.js"></script>
<script type="text/javascript" src="../../js/jquery.js"></script>
<script type='text/javascript' src="../../js/carousel.js"></script>
<script type='text/javascript' src="../../js/init.js"></script>
<script type="text/javascript">
jQuery.noConflict();
                                        jQuery(document).ready(function() {
                                                jQuery("#su-fpcarousel_50a2b85990490").css("display","block");
                                                jQuery("#su-fpcarousel_50a2b85990490").jcarousel({
                                                        auto: 5,
                                                        scroll: 1,
                                                        wrap: "last",
                                                        animation: 1900,
                                                        initCallback: mycarousel_initCallback
                                                 });
                                        });
function generateTabs(element) {
    jQuery(function($) {
        jQuery("."+element+" .tab[id^=tab_menu]").click(function() {
            var currentDiv=$(this);
            jQuery("."+element+" .tab[id^=tab_menu]").removeClass("selected");
            currentDiv.addClass("selected");
            var index=currentDiv.attr("id").split("tab_menu_")[1];
            jQuery("."+element+"-container .tabcontent").css('display','none');
            jQuery("."+element+"-container #"+element+"_tab_content_"+index).fadeIn();
        });
    });
}
generateTabs('divTabs');
                                </script>
<script type="text/javascript">
jQuery.noConflict();
                                        jQuery(document).ready(function() {
                                                jQuery("#su-fpcarousel_50a2b85990490x").css("display","block");
                                                jQuery("#su-fpcarousel_50a2b85990490x").jcarousel({
                                                        auto: 5,
                                                        scroll: 1,
                                                        wrap: "last",
                                                        animation: 1900,
                                                        initCallback: mycarousel_initCallback
```

Figura 2. Ejemplo de noticia en código html

Analizando la página de las noticias se pudo determinar que las partes importantes son:

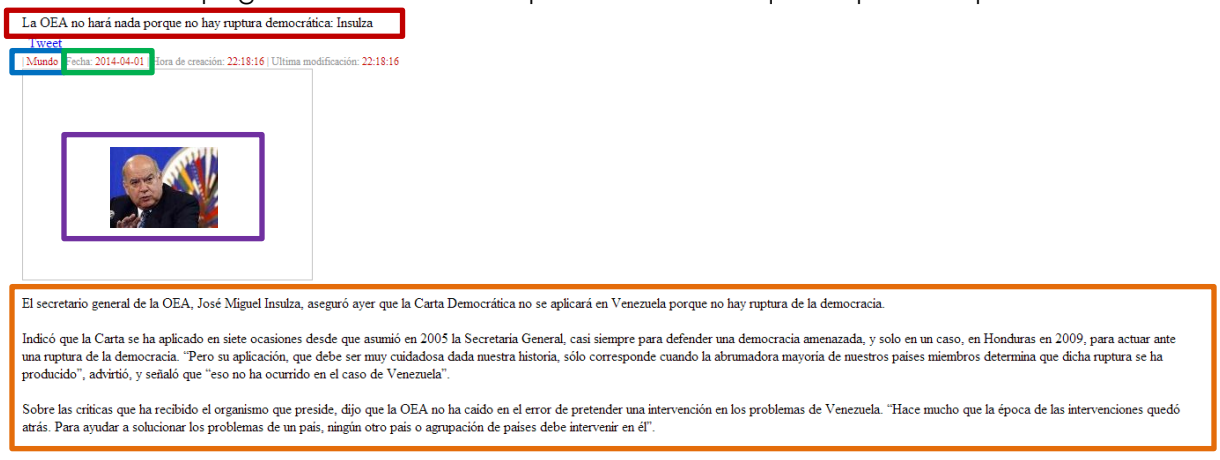

Figura 3. Partes importantes de la noticia.

El título

La OEA no hará nada porque no hay ruptura democrática: Insulza

La categoría

| Mundo |

La fecha

| Fecha: 2014-04-01 |

La imagen

El texto

El secretario general de la OEA, José Miguel Insulza, aseguró ayer que la Carta Democrática no se aplicará en Venezuela porque no hay ruptura de la democracia.

Indicó que la Carta se ha aplicado en siete ocasiones desde que asumió en 2005 la Secretaría General, casi siempre para defender una democracia amenazada, y solo en un caso, en Honduras en 2009, para actuar ante una ruptura de la democracia. "Pero su aplicación, que debe ser muy cuidadosa dada nuestra historia, sólo corresponde cuando la abrumadora mayoría de nuestros países miembros determina que dicha ruptura se ha producido", advirtió, y señaló que "eso no ha ocurrido en el caso de Venezuela".

Sobre las críticas que ha recibido el organismo que preside, dijo que la OEA no ha caído en el error de pretender una intervención en los problemas de Venezuela. "Hace mucho que la época de las intervenciones quedó atrás. Para ayudar a solucionar los problemas de un país, ningún otro país o agrupación de países debe intervenir en él".

Adicional a las partes que componen la noticia, también se considero el nombre del archivo que contiene la noticia, como clave para posteriormente poder ser identificada.

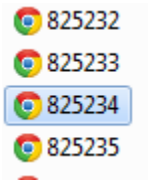

Figura 4. Ejemplo con los nombre de los archivos que contienen las noticias.

Para el proceso de limpieza se utilizó el programa "0007__FORMATEXT", el cual aplica expresiones regulares en los textos para poder usar o editar nuevos formatos. Este programa usa un archivo de entrada en donde la primera línea corresponde a la expresión regular (ER) a aplicar y la segunda línea es el nuevo formato por el que se quiere remplazar la primera línea o ER. Este programa permite tener muchos pares de patrones (ER-nuevo formato). Las expresiones regulares deben estar formadas en base a la sintaxis utilizada por java.

La sintaxis del programa "0007__FORMATEXT" es:

-IF str -SUBF str [-ONESTR) [-REVERSE) [-FIRST) [-UPPERCASE | -LOWERCASE)

-IF	Unformat input text file
-SUBF	Input text file with the patterns
-ONESTR	Reads all the lines of the input text file as one string, by default each line is read it by one string
-REVERSE	Generates a text file with patterns in reverse order as NewFormat-ER, which can be used to get the original text file

-FIRST The pattern is applied only to the first occurrence in the text
-UPPERCASE Converts all the out characters in upper case
-LOWERCASE Converts all the out characters in lower case

La sintaxis utilizada para la limpieza de los archivos utilizando el programa FORMATEXT es:

java -jar dist\FORMATEXT.jar -ONESTR -IF Noticias\828765.html -SUBF limpiar.txt >Clean\828765.txt

El archivo con los patrones para limpiar el texto se llama limpiar.txt y contiene lo siguiente.

(?m)(.*)(\r\n)
$1
.*\<! http://www.cronica.com.mx/notas/2014/([0-9\-)+)\.html >.*\<title>La Cr..nica de Hoy \|\s*(.*)\</title>.*\(\s+|.+)\s+\|\s+([^<)+)\.*\|\s+Fecha:\s+([0-9\-)+)\(.*\<div\s+class="(slider|img_nota)">\s*(\ |\s*)*<img\s+src="([a-zA-Z0-9\./\-_)+)".*/>)?.*\<span\s+class=\"texto\">(\s+\ \s*(.+)\ )?\s+\.*
{$1}{$2}{$4}{$5}{$9}{$10}
\ |\ |\
|\</br>|\
|â€œ|â€¦| |\|\|\ |\ |\|\

ñ
ñ
á
á
é
é
í
í
ó
ó
ú
ú
“
“
”
”
ç
ç
Ã\?
Á
‘
‘
’
’
„
„
"
"
&

&

⁄
/
Á
Á
É
É
Í
Í
Ó
Ó
Ú
Ú
Ñ
Ñ
·
·
Ã¡
á
Ã©
é
Ã-
í
Ã³
ó
Ãº
ú
Ã±
ñ
Ã\?
Á
Ã‰
É
Ã
Í
Ã\“
Ó
Ãš

Ú
Í¼
ü
Í§
ç
Â

Íš
Ú
Í“
Ó
Íœ
Ü
Í‘
Ñ
€™
'
\â\€\?

—
—
¿
¿
ä
ä
ë
ë
ï
ï
ö
ö
ü
ü
Ä
Ä
Ë
Ë
Ï
Ï
Ö

Ö
Ü
Ü
…
…
¡
¡
©
©
«
«
¬
¬
®
®
¯
¯
°
°
±
±
»
»
¼
¼
½
½
¾
¾
×
\×
÷
÷
'
‘
̀
̀
́
́
̂

^
̃
~
̄
-
̠
_
̋
"
´
´
–
\–
À
À
È
È
Ì
Ì
Ò
Ò
Ù
Ù
à
à
è
è
ì
ì
ò
ò
ù
ù
ç
ç
˜
˜
ˆ
ˆ
‹

‹
›
›
€
€
™
™
•
•
â
â
ê
ê
î
î
ô
ô
û
û
Â
Â
Ê
Ê
Î
Î
Ô
Ô
Û
Û
Ã
Ã
ã
ã
\<iframe.*\>

\</iframe>

\

<p style="text-align: center;">

\.\.\/\.\.\/
http://www.cronica.com.mx/
<input type="text" />

\<a href.*\>

\<p.*\>

\<span.*\>

\<font.*\>

\<div.*\>

\<br clear="all" />

LRM|OVR|evn|ovr|wg3|IGG|EVN

\<wbr>|\</wbr>|\|\

\^{|\}

\s+

A continuación se explican cada uno de los patrones, en donde la primera línea es el patrón a buscar y la segunda línea es el patrón a realizar.

(?m)(.)(\r\n)* En esta línea se indica que el texto se leerá en una sola línea con *?m*, y se quitaran

$1	todos los tabuladores y enters con *(\r\n)* El modificado se guardara para su posterior uso

A continuación se muestra el patrón donde se indican las partes del texto que quieren guardar (El título, el texto, la imagen, etc).

.\<! http://www.cronica.com.mx/notas/2014/([0-9\-)+)\.html >.*\<title>La Cr..nica de Hoy \|\s*(.*)\</title>.*\(\s+|.+)\s+\|\s+([^<)+)\.*\|\s+Fecha:\s+([0-9\-)+)\(.*\<div\s+class="(slider|img_nota)">\s*(\ |\s*)*<img\s+src="([a-zA-Z0-9\./\-_)+)".*/>)?.*\<span\s+class=\"texto\">(\s+\ \s*(.+)\ )?\s+\.*

En seguida se muestra desglosada la línea anterior con cada una de las partes que se están guardando del texto.

Nombre del archivo

```
<! http://www.cronica.com.mx/notas/2014/825232.html >
```

Patrón utilizado	**Descripción**
.\<! http://www.cronica.com.mx/notas/2014/([0-9\-)+)\.html >*	Con este patrón se obtiene el nombre del archivo.

Se obtiene

{825232}

Título del documento

```
<title>La CrÃ³nica de Hoy | UN CATALÃ?N GOBERNARÃ? FRANCIA Y UNA ANDALUZA, PARíS</title>
```

Patrón utilizado	**Descripción**
.\<title>La Cr..nica de Hoy |\s*(.*)\</title>*	Con este patrón se obtiene el título de la noticia.

Se obtiene

{UN CATALÃ?N GOBERNARÃ? FRANCIA Y UNA ANDALUZA, PARíS}

Categoría de la noticia

```
<span style="font-size:12px; color:#999;">  |
                <font color="#CC0000">Mundo</font>
```

Ó

```
<span style="font-size:12px; color:#999;">Notimex  |
                <font color="#CC0000">Negocios</font>
```

Patrón utilizado	Descripción
.*\(\s+|.+)\s+|\s+([^<]+)\	Con este patrón se obtiene la categoría de la noticia.

Se obtiene

{Mundo}

Fecha de la noticia

```
                    Fecha:
<font color="#CC0000">2014-04-01</font>
```

Patrón utilizado	Descripción
.*|\s+Fecha:\s+([0-9\-]+)\	Con este patrón se obtiene la fecha de la noticia.

Se obtiene

{2014-04-01}

Imagen de la noticia

```
| <div class="slider">
  <div>
        <img src="../../nimagenes/3/2014-03-31_10-03-47___3164.jpeg" alt=""/>
      </div>
```

ó

```
<div class="img_nota">
<img src="../../nimagenes/5/2014-04-01_12-04-08___3590.jpeg" width="330" height="240" />
```

Patrón utilizado	Descripción
(.*\<div\s+class="(slider|img_nota)">\s*(\ |\s*)*<img\s+src="([a-zA-Z0-9\./\-_]+)".*/>)?	Con este patrón se obtiene la imagen de la noticia.

Se obtiene

{…/nimagenes/3/2014-03-31_10-03-47___3164.jpeg}

Texto de la noticia

```
<p>
      Manuel Valls, de padre catal&aacute;n y nacido en Barcelona, fue nombrado ayer por el presidente de Francia, Fran&ccedil;ois Hollande, nuevo primer ministro,
tras la debacle sufrida por el Partido Socialista en las elecciones municipales del domingo, donde s&oacute;lo lav&oacute; su orgullo con la victoria de la socialista,
Anne Hidalgo, nacida en San Fernando de C&aacute;diz (Andaluc&iacute;a), y que se convierte, adem&aacute;s, en la primer mujer que gobernar&aacute; la capital gala, al
vencer a la candidata conservadora Nathalie Kosciusko-Morizet. La decisi&oacute;n de Hollande de poner al frente de su gobierno al ministro del Interior, considerado
un &ldquo;duro&rdquo; dentro del socialismo franc&eacute;s, ya que de &eacute;l fueron medidas muy pol&eacute;micas como la expulsi&oacute;n de gitanos, responde a un
deseo de mostrar un Ejecutivo m&aacute;s agresivo que levante su popularidad. Como an&eacute;cdota, un t&iacute;o abuelo de Valls, Manuel Valls i Gorina, fue el
compositor del himno del FC Barcelona, de donde le viene la afici&oacute;n por el club blaugrana que cultiva en la actualidad.</p>
```

Patrón utilizado	Descripción
.*\<span\s+class=\"texto\">(\s+\ \s*(.+)\ )?\s+\.*	Con este patrón se obtiene el texto de la noticia.

Se obtiene

{ Manuel Valls, de padre catalán y nacido en Barcelona, fue nombrado ayer por el presidente de Francia, François Hollande, nuevo primer ministro, tras la debacle sufrida por el Partido Socialista en las elecciones municipales del domingo, donde sólo lavó su orgullo con la victoria de la socialista, Anne Hidalgo, nacida en San Fernando de Cádiz (Andalucía), y que se convierte, además, en la primer mujer que gobernará la capital gala, al vencer a la candidata conservadora Nathalie Kosciusko-Morizet. La decisión de Hollande de poner al frente de su gobierno al ministro del Interior, considerado un “duro” dentro del socialismo francés, ya que de él fueron medidas muy polémicas como la expulsión de gitanos, responde a un deseo de mostrar un Ejecutivo más agresivo que levante su popularidad. Como anécdota, un tío abuelo de Valls, Manuel Valls i Gorina, fue el compositor del himno del FC Barcelona, de donde le viene la afición por el club blaugrana que cultiva en la actualidad.}

Para poder guardar el texto que se debe seleccionar el grupo correcto para poder obtener un texto prácticamente limpio.

.\<! http://www.cronica.com.mx/notas/2014/**([0-9\-)+)**\.html >.*\<title>La Cr..nica de Hoy \|\s***(.*)**\</title>.*\(\s+|.+)\s+\|\s+**([^<)+)**\.*\|\s+Fecha:\s+**([0-9\-)+)**\(.*\<div\s+class="(slider|img_nota)">\s*(\ |\s*)*<img\s+src="**([a-zA-Z0-9\./\-_)+)**".*/>)?.*\<span\s+class=\"texto\">(\s+\ \s***(.+)**\ )?\s+\.*

{$1}{$2}{$4}{$5}{$9}{$10}

El resultado de este proceso sería el siguiente.

```
{825232}{UN CATALÃ?N GOBERNARÃ? FRANCIA Y UNA ANDALUZA, PARÃ?S}{Mundo}{2014-04-
01}{../../nimagenes/3/2014-03-31_10-03-47___3164.jpeg}{Manuel Valls, de padre
catal&aacute;n y nacido en Barcelona, fue nombrado ayer por el presidente de
Francia, Fran&ccedil;ois Hollande, nuevo primer ministro, tras la debacle
sufrida por el Partido Socialista en las elecciones municipales del domingo,
donde s&oacute;lo lav&oacute; su orgullo con la victoria de la socialista, Anne
Hidalgo, nacida en San Fernando de C&aacute;diz (Andaluc&iacute;a), y que se
convierte, adem&aacute;s, en la primer mujer que gobernar&aacute; la capital
gala, al vencer a la candidata conservadora Nathalie Kosciusko-Morizet. La
decisi&oacute;n de Hollande de poner al frente de su gobierno al ministro del
Interior, considerado un &ldquo;duro&rdquo; dentro del socialismo
franc&eacute;s, ya que de &eacute;l fueron medidas muy pol&eacute;micas como la
expulsi&oacute;n de gitanos, responde a un deseo de mostrar un Ejecutivo
m&aacute;s agresivo que levante su popularidad. Como an&eacute;cdota, un
t&iacute;o abuelo de Valls, Manuel Valls i Gorina, fue el compositor del himno
del FC Barcelona, de donde le viene la afici&oacute;n por el club blaugrana que
cultiva en la actualidad.}
```

El siguiente paso para obtener un texto limpio es quitar los tags que aún se encuentren presentes en el texto, así como cambiar los códigos html en texto, como por ejemplo los acentos.

A continuación se presentan algunos ejemplos.

\ |\ |\ |\</br>|\ |â€œ|â€¦| |\\|\\|\ |\ |\|\	En esta línea se buscan los tags y se remplazan con la siguiente línea Esta línea simplemente es un espacio en blanco

Otro ejemplo es al utilizar palabras con acentos, en ocasiones el texto se compone con palabras como las que se muestran a continuación

Código en Java Script	Se remplaza por el carácter
Ã\?	Á
Ã‰	É
Ã	Í
Ã\“	Ó
Ãš	Ú
Ã¡	á
Ã©	é
Ã-	í
Ã³	ó
Ãº	ú

Cuando se obtiene la dirección de la imagen de la noticia, la ruta no está completa ({…/nimagenes/3/2014-03-31_10-03-47___3164.jpeg}), sin embargo se puede utilizar un patrón para obtener la dirección completa.

Patrón utilizado	**Se remplaza por**
\.\.\/\.\.\/	http://www.cronica.com.mx/

Se obtiene

http://www.cronica.com.mx/ nimagenes/3/2014-03-31_10-03-47

En el texto se pueden encontrar ligas hacia otras páginas o código para dar formato al texto y a continuación se muestran algunos ejemplos para eliminarlas.

Patrón utilizado	**Descripción**

\<a href.*\>	Al aplicar estos patrones se eliminan los tags que contengan al principio <a href> y <font>
\<font.*\>	

Limpiar toda la colección

Lo explicado anteriormente solo es para un archivo, sin embargo puede realizarse de manera automática la limpieza para toda la colección de archivo. A continuación se explica cómo se realiza este proceso.
Se tienen dos carpetas:

Carpeta	Descripción
Noticias	En esta carpeta se tienen las noticias en archivos .html para ser limpiadas
Clean	En esta carpeta se guardan los archivos limpios

Teniendo los archivos a limpiar en la carpeta de Noticias se procede a realizar el siguiente archivo.

File.bat

```
dir /B Noticias > files.txt
```

Este archivo crea otro archivo llamado files.txt el cual contiene una lista con los nombre de los archivos que están en la carpeta de Noticias.
Posteriormente se genera un archivo .bat que permite generar de manera automática un archivo para poder limpiar todos los archivos contenidos en la carpeta Noticias.

Prueba_ce.bat

```
java -jar dist\FORMATEXT.jar -FIRST -IF files.txt -SUBF main.txt > final.bat
```

Esta sentencia ocupa el archivo files.txt (el cual ya está generado y contiene los nombre de los archivos a limpiar), main.txt y final.bat. A continuación se describe cada uno.
El archivo main.txt es el que tiene la sentencia ocupada para ir limpiando cada uno de los archivos. Sin embargo, esta ajuntada con variables que permiten guardar esta sentencia con cada uno de los archivos a limpiar.

main.txt

```
(?m)(.*).html
java -jar dist\\FORMATEXT.jar -ONESTR -IF Noticias\\$1.html -SUBF limpiar.txt >Clean\\$1.txt
```

En el archivo final.bat se guardan cada una de las sentencia para limpiar los archivos de la carpeta Noticias.

final.bat

```
java -jar dist\FORMATEXT.jar -ONESTR -IF Noticias\828765.html -SUBF limpiar.txt >Clean\828765.txt
java -jar dist\FORMATEXT.jar -ONESTR -IF Noticias\828766.html -SUBF limpiar.txt >Clean\828766.txt
java -jar dist\FORMATEXT.jar -ONESTR -IF Noticias\828767.html -SUBF limpiar.txt >Clean\828767.txt
java -jar dist\FORMATEXT.jar -ONESTR -IF Noticias\828768.html -SUBF limpiar.txt >Clean\828768.txt
…
```

Para agilizar el proceso y no tener que ejecutar cada uno de los archivos .bat por separado se realizó un archivo .bat principal.

1.bat

```
call file.bat
call prueba_ce.bat
call final.bat
```

Este archivo permite que la limpieza de los archivo se realice de forma automática.

Tablas de código

ISO 8859-1 characters

Carácter	Tag HTML
¡	&iexcl;
¢	&cent;
£	&pound;
¤	&curren;
¥	&yen;
¦	&brvbar;
§	&sect;
¨	&uml;
©	&copy;
ª	&ordf;
«	&laquo;
¬	&not;

®	®
¯	¯
°	°
±	±
²	²
³	³
´	´
µ	µ
¶	¶
·	·
¸	¸
¹	¹
º	º
»	»
¼	¼
½	½
¾	¾
¿	¿
À	À
Á	Á
Â	Â
Ã	Ã
Ä	Ä
Å	Å
Æ	Æ
Ç	Ç
È	È
É	É
Ê	Ê
Ë	Ë
Ì	Ì
Í	Í
Î	Î
Ï	Ï
Ð	Ð
Ñ	Ñ
Ò	Ò
Ó	Ó
Ô	Ô

Õ	Õ
Ö	Ö
×	×
Ø	Ø
Ù	Ù
Ú	Ú
Û	Û
Ü	Ü
Ý	Ý
Þ	Þ
ß	ß
à	à
á	á
â	â
ã	ã
ä	ä
å	å
æ	æ
ç	ç
è	è
é	é
ê	ê
ë	ë
ì	ì
í	í
î	î
ï	ï
ð	ð
ñ	ñ
ò	ò
ó	ó
ô	ô
õ	õ
ö	ö
÷	÷
ø	ø
ù	ù
ú	ú
û	û
ü	ü

ý	ý
þ	þ
ÿ	ÿ

Math symbols

Carácter	Tag HTML
ƒ	ƒ

Flechas

Carácter	Tag HTML
←	←
↑	↑
→	→
↓	↓
↔	↔
↵	↵
⇐	⇐
⇑	⇑
⇒	⇒
⇓	⇓
⇔	⇔

Operadores matemáticos

Carácter	Tag HTML
∀	∀
∂	∂
∃	∃
∅	∅
∇	∇
∈	∈
∉	∉
∋	∋
∏	∏
∑	∑
−	−
*	∗
√	√
∝	∝
∞	∞
∠	∠

∧	∧
∨	∨
∩	∩
∪	∪
∫	∫
∴	∴
∼	∼
≅	≅
≈	≈
≠	≠
≡	≡
≤	≤
≥	≥
⊂	⊂
⊃	⊃
⊄	⊄
⊆	⊆
⊇	⊇
⊕	⊕
⊗	⊗
⊥	⊥
⋅	⋅

Puntuación general

Carácter	**Tag HTML**
•	•
…	…
′	′
″	″
‾	‾
⁄	⁄

Varias técnicas

Carácter	**Tag HTML**
⌈	⌈
⌉	⌉
⌊	⌊
⌋	⌋
⟨	⟨
⟩	⟩

Formas geométricas

Carácter	Tag HTML
◊	◊

Varios símbolos

Carácter	Tag HTML
♠	♠
♣	♣
♥	♥
♦	♦

Símbolos de letras

Carácter	Tag HTML
℘	℘
ℑ	ℑ
ℜ	ℜ
™	™
ℵ	ℵ

Letras Guiegas

Carácter	Tag HTML
Α	Α
Β	Β
Γ	Γ
Δ	Δ
Ε	Ε
Ζ	Ζ
Η	Η
Θ	Θ
Ι	Ι
Κ	Κ
Λ	Λ
Μ	Μ
Ν	Ν
Ξ	Ξ
Ο	Ο
Π	Π
Ρ	Ρ
Σ	Σ
Τ	Τ

Υ	Υ
Φ	Φ
Χ	Χ
Ψ	Ψ
Ω	Ω
α	α
β	β
γ	γ
δ	δ
ε	ε
ζ	ζ
η	η
θ	θ
ι	ι
κ	κ
λ	λ
μ	μ
ν	ν
ξ	ξ
ο	ο
π	π
ρ	ρ
ς	ς
σ	σ
τ	τ
υ	υ
φ	φ
χ	χ
ψ	ψ
ω	ω
ϑ	ϑ
ϒ	ϒ
ϖ	ϖ

Caracteres especiales para HTML

Controles y latín básico.

Carácter	Tag HTML
"	"
&	&
<	<

>	>

Latín extendido

Carácter	Tag HTML
Œ	Œ
œ	œ
Š	Š
š	š
Ÿ	Ÿ

Separación de modificación de cartas

Carácter	Tag HTML
ˆ	ˆ
˜	˜

Puntuación general

	‌
	‍
	‎
	‏
–	–
—	—
‘	‘
’	’
‚	‚
“	“
”	”
„	„
†	†
‡	‡
‰	‰
‹	‹
›	›
€	€

http://tunes.org/wiki/html_20special_20characters_20and_20symbols.html

Expresiones regulares en java.

Una expresión regular es un patrón que describe a una cadena de caracteres.

	Significado	**Ejemplo**	**Resultado**
****	Marca de carácter especial	/\$ftp/	Busca la palabra $ftp
^	Comienzo de una línea	/^-/	Líneas que comienzan por -
$	Final de una línea	/s$/	Líneas que terminan por s
.	Cualquier carácter (menos salto de línea)	/\b.\b/	Palabras de una sola letra
|	Indica opciones	/(L|l|f|)ocal/	Busca Local, local, focal
()	Agrupar caracteres	/(vocal)/	Busca vocal
[)	Conjunto de caracteres opcionales	/escrib[aoe)/	Vale escriba, escribo, escribe
*****	Repetir 0 o más veces	/l*234/	Valen 234, 1234, 11234...
+	Repetir 1 o más veces	/a+mar/	Valen amar, aamar, aaamar...
?	1 o 0 veces	/a?mar/	Valen amar, mar.
{n}	Exactamente n veces	/p{2}sado/	Vale ppsado
{n,}	Al menos n veces	/(m){2}ala/	Vale mmala, mmmala....
{m,n}	entre m y n veces	/tal{1,3}a/	Vale tala, talla, tallla
\b	Principio o final de palabra	/\bver\b/	Encuentra ver en "*ver de*", pero no en "*verde*"
\B	Frontera entre no-palabras	/\Bver\B/	Empareja ver con "*Valverde*" pero no con "*verde*"
\d	Un dígito	/[A-Z)\d/	No falla en "A4"
\D	Alfabético (no dígito)	/[A-Z)\D/	Fallaría en "A4"
\O	Carácter nulo		
\t	Caracter ASCII 9 (tabulador)		
\f	Salto de página		
\n	Salto de línea		
\w	Cualquier alfanumérico, [a-zA-Z0-9_)	/\w+/	Encuentra *frase* en "*frase*.", pero no el . (punto).

\W	Opuesto a \w ([^a-zA-Z0-9_))	/\W/	Hallaría sólo el punto (.)
\s	Carácter tipo espacio (como tab)	/\sSi\s/	Encuentra *Si* en *"Digo Si ",* pero no en *"Digo Sientate*"
\S	Opuesto a \s		
\cX	Carácter de control X	\c9	El tabulador
\oNN	Carácter octal NN		
\xhh	El hexadecimal hh	/\x41/	Encuentra la *A* (ASCII Hex41) en *"letra A*"

El signo de dólar "**$**"	Representa el final de la cadena de caracteres o el final de la línea, si se utiliza el modo multi-línea. No representa un carácter en especial sino una posición. Si se utiliza la expresión regular "\.$" el motor encontrará todos los lugares donde un punto finalice la línea, lo que es útil para avanzar entre párrafos.

Anexo 5.

Documentación de la colección TER

El corpus en español para resúmenes fue creado a partir del corpus de noticias obtenido del periódico CRÓNICA. Se seleccionaron de manera aleatoria 20 noticias de cada una de las categorías proporcionadas por el periódico.

1. Academia (20 archivos)
2. Bienestar (20 archivos)
3. Ciudad (20 archivos)
4. Cultura (20 archivos)
5. Deportes (20 archivos)
6. Espectáculos (20 archivos)
7. Estados (20 archivos)
8. Mundo (20 archivos)
9. Nacional (20 archivos)
10. Negocios (20 archivos)
11. Opinión (20 archivos)
12. Sociedad (20 archivos)

Las consideraciones para la selección fueron:

- Que las noticias tuvieran como longitud más de 100 palabras.
- Que fueran de diferentes longitudes
- Todas las noticias seleccionadas para formar el corpus fueron noticias del mes de abril de 2015

Consideraciones para colocar el nombre a las noticias.

Para nombrar a cada uno de los archivos se siguieron las siguientes consideraciones.

- Se consideró un número consecutivo (1-20)
- Se tomaron dos letras para la categoría
- Se consideró la fecha de la noticia
- La clave de la noticia

Según las categorías a continuación se muestra las letras que se consideraron para formar parte del nombre.

- Academia AC
- Bienestar BI
- Ciudad CI
- Cultura CU
- Deportes DE
- Espectáculos ES

- Estados ED
- Mundo MU
- Nacional NA
- Negocios NE
- Opinión OP
- Sociedad SO

A continuación se muestra el siguiente ejemplo.

Una vez construido el corpus de noticias en español, se construyeron para cada archivo dos resúmenes creados por dos expertos.
Las consideraciones tomadas para seleccionar un experto fueron:

- Nacionalidad mexicana
- Educación mínima de licenciatura

Para que el experto creara el resumen se le proporcionó una hoja de instrucciones que debía seguir, junto con un ejemplo. (se anexa hoja de instrucciones)

Creación del corpus en español mexicano para la generación de resúmenes

Nombre del experto: _____

Instrucciones para formar el resumen:

- Lea la noticia completamente.
- Seleccione las oraciones que consideres más importantes para formar el resumen de la noticia.
- De las oraciones seleccionadas:
  - Sume el número de palabras, el resumen debe ser mayor a 100 palabras.
  - Ordénelas de acuerdo a la importancia que usted considere deben tener.

Ejemplo para formar el resumen

Se muestra el título, la clave de la noticia, el texto a resumir, un apartado donde el experto colocará el orden de las oraciones de acuerdo a lo que el considere, finalment un apartado para colocar el número total de palabras que contendrá el resumen. Para todos los casos, el título y clave de la noticia no se considerarán como parte d texto a resumir, por lo que solo se debe resumir el texto incluido en la tabla.
La tabla está constituida de la siguiente manera, tiene el número de oraciones así como el número de palabras por oración. En el siguiente ejemplo, se han seleccionado oraciones (1,2 y 8) dando un total de 119 palabras. Sin embargo sólo es un posible resumen, ya que alguna otra persona pudo haber seleccionado por ejemplo la oraciones 1,2 y 3 dando un total de 115 palabras lo cual está dentro del rango de palabras mínimas. En otro caso, alguna otra persona pudo haber seleccionado la oraciones 1, 2, 4, 5 y 7 dando un total de 182 palabras, considerando que el mínimo de palabras es de 100 se pudiera considerar correcto. Sin embargo, si analizamos caso podemos observar que se pudieran considerar menor número de oraciones y aun así conseguir tener el mínimo de palabras. Por lo que se debe considerar cumpl con el mínimo de palabras seleccionando sólo el número de oraciones importantes que superen esta cantidad. En el apartado donde se coloca la importancia de la oraciones, el experto según las oraciones que ha seleccionado debe ordenarlas colocando un número que será consecutivo según la importancia dada.

Título de la noticia: Real Madrid estrena el año trabajando en un ambiente distendido
Clave de la noticia: 01DE010112_624998

Oración		Palabras por oración	Texto por oración	Importancia de la oración dada por el experto
1	●	54	La plantilla del Real Madrid estrenó el año con una sesión de entrenamiento vespertina, desde las 16:00 horas y a puerta cerrada en la ciudad deportiva de Valdebebas, para preparar el partido de ida de octavos de Copa contra el Málaga, que se jugará el martes a las 22:00 horas en el Santiago Bernabeu	2
2	●	48	Fue una sesión de trabajo a puerta cerrada, aunque el club anuncia en su página web que el ambiente fue "cordial y distendido" entre los 21 jugadores del primer equipo que el técnico Jose Mourinho tiene disponibles, a los que se sumaron los canteranos Nacho, Morata y Mejías	3
3	○	13	Los lesionados Sergio Ramos y Di María continuan con sus procesos de recuperacion	
4	○	28	Los jugadores blancos se ejercitaron el año y se ejercitaron por penúltima ocasión antes de afrontar el partido contra el Malaga del chileno Manuel Pellegrini, ex técnico madridista	
5	○	34	Los habituales estiramientos y ejercicios de potenciacion muscular a las ordenes del preparador físico Rui Faria dieron inicio a un entrenamiento en el que el esférico volvió a ser protagonista, según informa el club	
6	○	15	Concluida la parte física, los jugadores blancos llevaron a cabo el trabajo con el balon	
7	○	18	No saltaron al césped los lesionados Sergio Ramos y Di María, que continuan con sus procesos de recuperacion	
8	●	17	Este lunes, Mourinho dirigirá un entrenamiento vespertino en el estadio Santiago Bernabeu, a partir de las 18:00.	1

∑Total 119 La suma debe ser mayor a 100 palabras

Ilustración 1. Hoja de instrucciones proporcionadas al experto para generar el resumen.

Una vez creado el resumen por el experto se nombraron los archivos de resumen de la siguiente manera.

Resumen Experto 1

Nombre: Selene Vargas Flores

Clave de experto: LX

SUM_01AC010414_825278_LX

Prefijo — Nombre de la noticia — Calve del

Resumen Experto 2

Nombre: Yanet Hernández Casimiro

Clave de experto: RX

SUM_01AC010414_825278_RX

Se anexa el archivo con los nombres de los expertos y las claves asignadas a cada uno de ellos.

Clave de experto	Nombre del expero	Noticias Asignadas
A	Artemio Becerril García	Sociedad 01-10 Negocios 01-10
B	Betsabé Alarcón Reyes	Nacional 16-20 Opinión 11-20 Negocios 11-20 Estados 09-20
C	Chriatian Ruiz Ugalde	Negocios 11-20
D	Eder	Ciudad 11-20 Estados 01-08
E	Griselda Areli Matias Mendoza	Nacional 01-15 Estados 09-20 Mundo 11-20 Bienestar 01-10
F	Miguel Ángel García Calderón	FA- Miguel FB-Ema Mendoza Soto FC-Ulises García Calderc Cuidad 01-10 Espectaculos 01-10 Estados 01-08 Sociedad 11-20 Negocios 01-10
G	Nancy Nagay González González	Cultural 07-20 Bienestar 01-10
H	Nayely Osorio de Jesús	Mundo 1-10 Academia 11-20
I	Néstor López Alarcón	Bienestar 11-20 Cultural 07-20
J	Rafael Cruz Reyes	Opinión 01-10 Mundo 01-20 Depoprtes 01-20 Ciudad 11-20 Academia 11-20 Cultura 01-07
K	René Arnulfo García Hernández	Cultura 02-07
L	Selene Vargas Flores	Academia 01-10 Sociedad 11-20
M	Gustavo Ignacio Cejudo Hernández	Nacional 01-20

N	Alan Serrano León	Ciudad 01-10
O	Marcela Camacho Ávila	Depoprtes 01-20
P	Mario Ibañez	Opinión 01-20
R	Yanet Hernández Casimiro	Academia 01-10
S	Alejandra Rios	Espectaculos 01-20
T	Jovani Armeaga García	Bienestar 11-20
U	Andrea Maria Ester Fernández Mendoza	Espectaculos 11-20
W	Karol Moreno Villa	Sociedad 01-10

A	Mario Ibañez López
B	Mario-Alexander Aikmen López Becerril
C	Mario-Jaime Cruz Meza
D	Mario-Carlos Andrés Rodríguez Placios
E	Mario-Eduardo Israel González García
F	Mario-Victor Antonio Dimas López
G	Mario-Ulises Santana Juárez
H	Mario-Sergio Oliveros Martínez
I	Mario-Carlos Alberto Fuentes González
J	Mario-Edgar Humberto Alvarez Carmona
K	Mario-Maria de los Angeles Monroy Lara
L	Mario-Isaac Adrian Alvarez Ortiz
M	Mario-Alexander Aikmen López Becerril
N	Mario-Luis Omar Escamilla Martínez
O	Mario-Christian Ruben López García
P	Mario-Iván Hernández Esquivel
Q	Mario-Alberto Durán Hernández
R	Mario-Ulises Uriel Alvarez Vilchis
S	Mario-Raúl Palma Camacho