UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

Evaluación de herramientas comerciales y métodos del estado del arte para la generación de resúmenes en idioma ruso

Tesis
Para Obtener el Título de
Ingeniera en Software

Que Presenta

Jessica Maribel Rojas Sánchez

Directora:
Dra. Yulia Nikolaevna Ledeneva

TIANGUISTENCO, MÉX. JUNIO 2016

UAEM | Universidad Autónoma del Estado de México

UAP TIANGUISTENCO
Unidad Académica Profesional Tianguistenco

El comité revisor designado por la Subdirección Académica de la Unidad Académica Profesional Tianguistenco de la Universidad Autónoma del Estado de México, aprobó la tesis: **EVALUACIÓN DE HERRAMIENTAS COMERCIALES Y MÉTODOS DEL ESTADO DEL ARTE PARA LA GENERACIÓN DE RESÚMENES EN IDIOMA RUSO** y autorizó la impresión de la misma de la C. **JESSICA MARIBEL ROJAS SÁNCHEZ** el día **23 de Mayo de 2016**.

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

Revisor
Mtro. Rafael Cruz Reyes

Revisor
Dr. José Luis Tapia Fabela

Asesor
Dra. Yulia Nikolaevna Ledeneva

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

Subdirección Académica
UAP
Tianguistenco

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

Dedico este trabajo a:

Mis dos ángeles de la vida, mis padres: "Rosa y Carlos" a ti papi por forjar mi camino por dirigirme al sendero correcto, a ti mami por ser una guerrera por darme el apoyo, la comprensión, la libertad… simplemente por darme tu vida a cambio de mi felicidad a ustedes por todo su amor infinito ¡Siempre juntos, hú!

A mí querido hermano por siempre creer en mí, por ayudarme a vencer el miedo Ing. Juan Carlos.

A mis abuelos porque gracias a la dicha del señor hoy puedo gozar de sus abrazos, por las bendiciones y los ricos desayunos que nunca faltaron antes de salir de casa, porque eres mi segunda mamá, a ti abuelita Elia.

A mi príncipe Eduar, mi sobrino por ser la chispa de alegría que llego a darle a la familia, por recordarme día a día lo fácil que es sonreír.

Principalmente a Dios.

Agradecimientos

Especial agradecimiento a mi directora de tesis Dra. Yulia Nikolaevna por haberme brindado la oportunidad de trabajar en conjunto, por compartirme de sus conocimientos, por su apoyo personal, así como también por su paciencia para guiarme durante todo el desarrollo de la tesis.

A mis maestros que con su amor a la educación y su paciencia hoy estoy escribiendo estas líneas a: Mtra. Gloria Ortega por todo el apoyo durante la carrera. Mtra. Patricia Hurtado por la motivación a la titulación. Mtra. Griselda por el apoyo durante el desarrollo de esta tesis. Mtro. Rafael Cruz por compartirme de su inapreciable aprendizaje, por todo el tiempo brindado así como de su paciencia, por darme el ejemplo del aprendizaje autónomo y diario, por la motivación no solo académico sino personal. Al grupo de seminario por los consejos, correcciones, aportaciones etc. sin duda alguna una forma dinámica para un buen aprendizaje.

Y sin duda alguna a mis compañeros y amigos (Diana, Osmar, Armando, Moisés, Edgar, Eder, Gaby, Yanet, Miguel, Alan, Lic. Stephany y Dr. Mario) por su apoyo académico, por los convivios que fortalecieron nuestros lazos de amistad. A mi amigo Abraham por coincidir en este viaje. Por ser la familia que yo elegí para acompañarme en una parte de mi vida.

"No puedo hacer un resumen de mi vida, porque está conformada por varias épocas y circunstancias, libros, amistades y pleitos, y eso, sólo admite resúmenes parciales."

-Carlos Monsiváis

Resumen

Con el crecimiento exponencial de la información, la tarea de su comprensión se vuelve más difícil. Una forma de ayudar en esta tarea, es mediante la generación automática de resúmenes de texto. El objetivo de estos resúmenes, es presentar el contenido en una versión más corta que la del texto original, y que permita al usuario comprender más rápido el gran volumen de información.

Los métodos de integración de textos se pueden clasificar en extractivos y abstractivos. Los resúmenes abstractivos utilizan métodos lingüísticos para examinar e interpretar el texto y luego encontrar nuevos conceptos y expresiones, para describir mejor la generación de un nuevo texto más corto, que trasmite la información más importante del documento original Debido a esto, tales resúmenes hasta hoy los realizan los seres humanos. Por su parte, un resumen extractivo está compuesto con fragmentos del documento original. El principal objetivo de la generación de resúmenes extractivos es la selección automática de frases (frases, oraciones o párrafos) de texto que reflejarían de manera adecuada el contenido del documento. Estos últimos son los que se utilizan para la generación automática de resúmenes.

Dado que los trabajos tomados en cuenta del estado del arte, operan con corpus en idiomas inglés, portugués y español. Surge el interés de conocer la calidad de las herramientas comerciales si se utiliza en otro idioma. En el presente trabajo de tesis se realiza la evaluación de las herramientas comerciales y los métodos del estado del arte para la generación automática de resúmenes en el idioma ruso.

Para la selección de las herramientas comerciales se analizaron cuáles de ellas están a disposición, y a su vez realizan resúmenes en idioma ruso. Las herramientas comerciales seleccionadas en línea fueron: Open Text Summarizer, Text Compactor, Tools4noobs, T-Conspectus. Las herramientas comerciales instalables elegidas fueron: Microsoft Word 2003 y Microsoft Word 2007, en estas dos últimas herramientas se probaron en los sistemas operativos

Windows: VISTA, XP, 7 Ultimate y 8. La selección de los métodos del estado del arte, se tomaron en cuenta los trabajos más relacionados de los cuales se tiene mayor información.

La aportación de este trabajo es la creación del corpus en el idioma ruso que lleva por el nombre *TEXTRUSS*. Está compuesto de 242 noticias contenidas en 11 categorías, con la finalidad de tener mayor diversidad de temas. Tal corpus se utiliza para generar resúmenes tanto como para las herramientas comerciales como para los métodos del estado del arte. La segunda aportación consiste en realizar la evaluación de herramientas comerciales y métodos del estado de arte.

Para la evaluación de los resúmenes se utilizó el sistema ROUGE el cual mide la similitud y determina la calidad de un resumen automático. Se muestran los resultados obtenidos.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.

Antecedentes.

Durante las últimas décadas se ha presentado un crecimiento en la cantidad de información almacenada en medios electrónicos. Se estima que el 20 por ciento de la información electrónica de una empresa está almacenado en bases de datos y el 80% está de manera textual [Nea 02], la cual es posible acceder de manera fácil a los objetos y registros de dicha información. Debido al crecimiento de la información en bases de datos, y a los beneficios que se pueden obtener a partir de ésta, ha cobrado gran interés el análisis de la información contenida en éstas. Mismo que ha sido abordado por el área de investigación "*Descubrimiento de Conocimiento en Bases de Datos*", que ha sido definida por [Fayyad 96] como:

> *"El descubrimiento de conocimiento en bases de datos es el proceso no trivial de identificar patrones en datos los cuales deben de ser válidos, novedosos, potencialmente útiles y entendibles."*

De acuerdo con las investigaciones, el volumen de información publicada en la Web se estima de 167 terabytes, mientras que la *Deep Web,* o también llamada red profunda, es de 400 a 450 veces más grande, por lo tanto entre 66,800 y 91,850 terabytes es el volumen de información [Lyman 16].

1.1 Relevancia de la cantidad de información

Actualmente la información textual dentro de Internet ha ido incrementando constantemente, en un 40% anual en este universo digital [Gantz 14]. El aumento acelerado de la información, ha provocado que se incremente la importancia de mejorar el desempeño de herramientas y métodos que resuman y presenten información relevante para el usuario para su mayor comprensión.

Los resúmenes son útiles ya que nos ayudan a facilitar tareas de lectura, por ejemplo, en artículos científicos, libros, películas, artículos de noticias, resúmenes de información para empresarios, etc. Los resúmenes generados a partir de estos documentos ayudan al usuario a discernir, entre varios documentos de la colección, sobre el mismo tema que le interesan y los que no. Los sistemas de generación automática de resúmenes se utilizan, por las personas tanto por grandes empresas.

1.2 Definición de resumen

Un resumen según [Lloret 08] es un texto que se genera a partir de uno o más textos, que contienen la información más significativa y que no es más extenso que la mitad de los textos originales, manteniendo las ideas principales. Además un buen resumen también debe de tener coherencia, la extensión de éste, depende de las necesidades del usuario y del tamaño del documento original siendo esto lo más flexible para definir al usuario que tan

amplio lo requiere [García 09a]. De manera que un resumen debería de contener explícitamente la información que el usuario está buscando. Sin embargo, actualmente los grandes sistemas de generación de resúmenes no han alcanzado este nivel.

1.3 Tipos de resúmenes

En el área de Procesamiento del Lenguaje Natural (PLN), los métodos para la construcción de resúmenes de manera automática se caracterizan por su contenido en resúmenes indicativos y resúmenes informativos [Hovy 99a], donde los primeros sólo aportan una idea general sobre el contenido real del documento origen, y los segundos son los que dan al usuario una versión corta del contenido real del documento original. Sin embargo, los resúmenes pueden ser clasificados también por la forma en que son construidos, éstos son: en resúmenes abstractivos y resúmenes extractivos no solo para un único documento, sino también para una colección de documentos [Lin 97], [García 09a].

Los resúmenes abstractivos se basan en un análisis semántico del documento esperando llegar a entenderse a nivel de frase. Este análisis es de mayor profundidad que el extractivo, permite obtener una representación del contenido del texto que puede utilizarse, junto a técnicas de generación de lenguaje natural, para confeccionar el resumen. Para las fases de análisis y generación se requiere gran cantidad de conocimiento del dominio, por lo que este tipo de sistemas solo es aplicable en ámbitos muy concretos y conocidos [Montiel 09a].

Los resúmenes extractivos inician con una selección de frases o párrafos a partir del texto original [Alguliev 05]. Por lo general son presentados al usuario en el mismo orden que el documento original, es decir una copia del texto original con la mayoría de las frases omitidas. La mayor ventaja de este enfoque es que resulta muy robusto y es fácilmente aplicable a contextos de propósito general, ya que su independencia del dominio, e incluso del género, de los documentos es muy alta [Maña 03]. Además, los resúmenes extractivos se caracterizan por ser el producto de un análisis superficial del texto fuente, no profundizando más allá del nivel sintáctico; de esta manera se puede trabajar de manera independiente de lenguaje y dominio.

Una de las características significativas de los resúmenes extractivos es la incorporación de métodos de aprendizaje automático, para lograr identificar con mayor precisión los atributos más importantes del texto a resumir.

Los resúmenes extractivos al utilizar reglas matemáticas permiten que se puedan aplicar a cualquier texto no importando su lenguaje y se pueden extraer oraciones, frases o párrafos [Ledeneva 08b]. Un resumen extractivo solo decide, para cada oración si es o no incluida en el resumen, lo que no hace necesario la comprensión del documento, además de alcanzar una mayor eficiencia al no utilizar un costoso análisis lingüístico [Ledeneva 08b].

En esta tesis, se generan los resúmenes extractivos que obtienen solamente las oraciones importantes.

1.4 Herramientas comerciales

Generalmente los humanos realizan resúmenes de tipo abstractivo, mientras que la mayoría de las herramientas comerciales y los métodos del estado del arte generan resúmenes de forma extractiva. Las herramientas comerciales permiten la Generación de Resúmenes Extractivos Individuales (GREI).

Las herramientas comerciales se clasifican en dos tipos:

- ✓ Las herramientas en línea, es decir, que están disponibles en internet, son alojadas en páginas web. Algunas herramientas en línea son: *Tool4noobs* [T4NS 15], *T-Conspectus* [Conspectus 15], *Open Text Summarizer* [OTS 15].

- ✓ Las herramientas instalables, son ejecutadas en el ordenador (.exe) y no necesitan internet para su funcionamiento, unas de ellas son gratuitas, y otras tienen un costo. Algunas herramientas instalables son: *Copernic Summarizer* [Copernic 15] y *Microsoft Office Word* [MOW 15].

Algo que tienen en común es que el método con el cual trabajan no es publicado. Actualmente existen varias herramientas comerciales que nos ayudan en la generación de resúmenes automáticos. En esta tesis las herramientas comerciales consideradas para analizar y comparar se mencionan a continuación:

- *Open Text Summarizer* [OTS 15]
- *Text Compactor* [TextCompactor 15]
- *Tools4noobs* [T4NS 15]
- *T-Conspectus* [Conspectus 15]
- Sistema operativo *Windows* 7 paquetería de *Microsoft Word* 2003 [MOW 15]
- Sistema operativo *Windows* 7 paquetería de *Microsoft Word* 2007 [MOW 15]
- Sistema operativo *Windows* VISTA paquetería de *Microsoft Word* 2003 [MOW 15]
- Sistema operativo *Windows* VISTA paquetería de *Microsoft Word* 2007 [MOW 15]
- Sistema operativo *Windows* XP paquetería de *Microsoft Word* 2003 [MOW 15]
- Sistema operativo *Windows* XP paquetería de *Microsoft Word* 2007 [MOW 15]
- Sistema operativo *Windows* 8 paquetería de *Microsoft Word* 2003 [MOW 15]
- Sistema operativo *Windows* 8 paquetería de *Microsoft Word* 2007 [MOW 15]

1.5 Métodos del estado del arte

En este trabajo los métodos del estado de arte considerados para analizar y comparar se mencionan a continuación:

-Comparación de los modelos de bolsa de palabras y *n*-gramas (utilizando algoritmo genético) [Matias 13], [Matias 16].
-*1st Best Method, 2nd Best Method, 3rd Best Method, MFS, 4th Best Method, 5th Best Method, baseline* [Ledeneva 11]
-*n*-gramas, ponderación booleana [García 08]
-SFMs, bolsa de palabras y bigramas [García 09]
-*n*-gramas, bolsa de palabras, y SFMs [Montiel 09]

-Algoritmo *PageRank* [Mihalcea 04]

1.6 Motivación y posibles aplicaciones

La motivación por la cual surge esta tesis es el interés de conocer: ¿Cómo es que funcionan los generadores automáticos de resúmenes?, ¿Cómo es que varía el desarrollo de estos mismos, aplicándole otro idioma? Es por eso, que se elige el idioma ruso, ya que a diferencia del idioma de los trabajos del estado del arte ya realizados, el idioma ruso contiene símbolos cirílicos.

Otro aspecto de motivación es aportar algo nuevo a las investigaciones y a los trabajos ya realizados, es por eso que en el desarrollo de esta tesis también se apoya a tener un estado de arte más amplio.

Se dice que una herramienta es un apoyo a las tareas a desarrollar, ya que ayuda a hacerlo menos difícil, o mejor aún, las herramientas automáticas son las que trabajan por sí mismas, es decir, hacen la mayor parte del trabajo, tal es el caso de las herramientas generadoras de resúmenes automáticas, las cuales sirven como apoyo, ya sea a empresas que trabajan con grandes cantidades de información o a los usuarios que trabajan con tareas más simples.

Los resúmenes generados automáticamente pueden utilizarse como sustitutos de los documentos originales o simplemente como referencia en la selección de documentos para una lectura más profunda. La primera de estas aplicaciones la encontramos, por ejemplo, en la generación de resúmenes sobre noticias periodísticas, donde es posible encontrar sistemas completamente operativos como [NewsBlaster 16] de la Universidad de Columbia. Este sistema combina la generación de resúmenes con un sistema de personalización de noticias, permitiendo al usuario acceder únicamente a aquellas que le interesen según indique su modelo de usuario [Díaz 05].

La generación automática de resúmenes, también es de gran utilidad como paso intermedio en otras tareas de PLN, por ejemplo, se ha demostrado que el uso de resúmenes en lugar de los documentos originales, en tareas de recuperación y categorización de información

produce un ahorro de tiempo, sin pérdida significativa de efectividad [Mani 01], [Saggion 10].

La presente aportación de investigación, puede ser aplicada a la contribución de mejorar un sistema de generación de resúmenes automáticos. Así como, demostrar la calidad del contenido de resúmenes generados, tanto de las herramientas comerciales como para los métodos del estado del arte, para el idioma ruso.

1.7 Planteamiento del problema

Hoy en día se tiene conocimiento del avance significativo de las herramientas y métodos del estado del arte para la generación de resúmenes extractivos, que están disponibles y se cuenta además con diversos corpus en diferentes idiomas.

De acuerdo a los métodos del estado de arte elegidos para trabajar se muestra que solo se ha trabajado con corpus en idioma inglés, portugués e español, es por ello que se desea conocer la calidad de los resúmenes para el idioma ruso, empleando herramientas de generación de resúmenes en herramientas comerciales.

Con ello tenemos la pregunta de investigación:

¿Cuál es la calidad de los resúmenes generados por las herramientas comerciales actuales y los métodos del estado del arte para el idioma ruso?

1.8 Objetivos

1.8.1 Objetivo general

Conocer la calidad del desempeño de los resúmenes, entre las herramientas comerciales y los métodos del estado del arte, cuando se les aplica un corpus en el idioma ruso.

1.8.2 Objetivos particulares

- Describir los métodos del estado del arte de los trabajos relacionados.
- Verificar si aún están a disposición las herramientas comerciales instalables y en línea (*Copernic*, *Open Text Summarizer*, *Microsoft Office Word 2007*, *Microsoft Office Word 2003* en sistema operativo *de Windows 7 y Windows XP, Text Compactor, Online Summarize tools4noobs, Shvoong summarizer*).
- Buscar herramientas comerciales que se especialicen en el idioma ruso.
- Examinar las herramientas comerciales instalables y en línea cómo es que realizan el proceso de resúmenes (principalmente, los resúmenes en idioma ruso).
- Comparar la validación y coherencia del resumen generado.
- Establecer los parámetros para cada herramienta, para evaluarlos iguales.
- Realizar los resúmenes con cada herramienta.
- Evaluar cual herramienta realiza una mejor calidad en el idioma ruso.
- Analizar los resultados.

1.9 Hipótesis

Si se aplica el corpus desarrollado para este trabajo, a las herramientas comerciales y a los métodos del estado del arte revisados, será posible conocer su desempeño, al ser evaluadas con el sistema *ROUGE*.

1.10 Delimitación del problema

- El método probado no se concentra en el tiempo de resolución, si no en la calidad del resumen.
- Con los resultados obtenidos se espera ver la comparación de la calidad, más no se desarrollará ningún sistema.

- El método utilizado solo se prueba con el corpus *TEXTRUSS*.
- No se desarrolla ningún método, ya que se pretende evaluar con los métodos del estado del arte.
- En esta tesis, no se emplea lista de palabras vacías, ni *stemming*.

1.11 Estructura de la tesis

En este primer capítulo, se describe la relevancia de la cantidad de información, se explica que es un resumen, los tipos de resúmenes que existen y se define con cuál de ellos se trabaja. También se habla de cómo es que los resúmenes son un apoyo para la resolución de la cantidad exponencial de información, mediante herramientas comerciales, así mismo se describen las herramientas existentes, con las cuales se trabaja. Se menciona que es método del estado del arte. Y por último, se presenta la motivación y posibles aplicaciones, el planteamiento de problema, los objetivos, la hipótesis, de delimitación del problema de esta tesis.

En el segundo capítulo, se expone al lector en que área de investigación está situado en el presente trabajo. Se describe de manera breve algunas de las aplicaciones del área. Se mencionan los conceptos básicos que se manejaran a lo largo de la investigación. También se presenta los pasos para la creación de un resumen extractivo, describiendo cada uno de ellos, mismos que en algunos de los trabajos del estado del arte se implementan. Además se explican las técnicas para la generación automática de resúmenes. Por último, se presenta el método de evaluación que se utiliza en esta tesis.

En el tercer capítulo, se describe algunos de los trabajos al estado del arte relacionados a la generación automática de resúmenes, cabe mencionar, que estos se describen de acuerdo al idioma del trabajo, comenzando por el idioma inglés, portugués e español.

En el cuarto capítulo, se explica la metodología de trabajo que se desarrolla para llevar acabo el trabajo de investigación.

En el quinto capítulo, se describe a detalle el corpus *TEXTRUSS*, el cual se creó. Se mencionan los parámetros establecidos para cada herramienta comercial. También se describe la serie de pasos efectuados para la implementación del método que se tomó del estado del arte. Por último se muestran los resultados de las herramientas comerciales y el método del estado del arte, así como su análisis.

En el sexto capítulo, se presentan las conclusiones de la tesis. Finalmente, se presentan los anexos correspondientes.

CAPÍTULO 2.

Marco Teórico

El presente capítulo tiene como objetivo presentar al lector los conceptos más importantes de la tarea de generación automática de resúmenes de texto. En principio se habla del procesamiento de lenguaje natural, con el fin de situar al lector en que área de investigación nos encontramos. Se menciona un esquema el cuál usan la mayoría de los sistemas y métodos que se involucran en el PLN. Así como también, se menciona de manera breve las tareas y aplicaciones del área. Continuando con la definición de resumen y de los factores que intervienen en su elaboración, siguiendo con la clasificación de los resúmenes. Se describen las fases de la arquitectura general de la generación automática de resúmenes, así como las técnicas. Finalmente, se describe la herramienta de evaluación *ROUGE*.

2.1 Procesamiento del Lenguaje Natural

El exceso de información conlleva un problema, la sobrecarga, provocando que no haya tiempo para leerlo todo, sin embargo, es necesario tomar decisiones críticas basadas en la información disponible. En este contexto, surge la necesidad de desarrollar sistemas que resuman automáticamente los contenidos y por consiguiente situarlos en el dominio del Procesamiento del Lenguaje Natural (PLN).

Por PLN se entiende la habilidad de la máquina para procesar la información relacionada, no simplemente las letras o los sonidos del lenguaje. La ciencia que estudia el lenguaje humano se llama lingüística para las computadoras, tiene el objetivo de dotar a las computadoras con la capacidad de entender el lenguaje humano.

[Gelbukh 10] dice que la lingüística computacional, es su etapa actual de desarrollo, es principalmente una rama de las tecnologías de aprendizaje automático, una parte de la inteligencia artificial y la estadística.

El esquema general de la mayoría de los sistemas y métodos que involucran el procesamiento de lenguaje enseñan a la computadora a tratar el procesamiento de textos, es decir, enseñarle el camino de aprendizaje de la estructura de textos y de su formalización [Gelbukh 06].

Esta tesis solo está enfocada al PLN, relacionado a textos escritos.

2.2 Aplicaciones del PLN

El PLN tiene un gran número de aplicaciones prácticas en muchas áreas de la vida cotidiana del ser humano, algunas de ellas han sido plasmadas en películas de ciencia ficción, en donde las personas pueden hablar con las máquinas (o sea robots), que persona no quisiera poder obtener un androide como C-3PO (personaje ficticio del universo de la Guerra de las Galaxias, diseñado en un principio para tareas del hogar, quién después, tuvo el uso de

traductor, dominando más de siete millones de formas de comunicación) [StarWars 16] y quizá no estemos muy lejos de lograrlo, con los avances pequeños pero significativos logros tecnológicos en las aplicaciones de las tecnologías del PLN.

Las aplicaciones del PLN son las siguientes:

- Extracción de frases clave usando patrones léxicos en artículos científicos [Hernández 16].
- Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos [Vargas 16].
- Detección de fragmentos de texto como candidato a hipervínculo [Camacho 15].
- Comparación de medidas de similitud en cadenas textuales, para la detección de plagio en tareas escolarees [Armeaga 15].

2.3 Generación automática de resúmenes

Un resumen es una transformación reductiva de un texto fuente a un texto resumen por reducción de su contenido mediante selección y/o generalización de lo que es importante en el texto fuente [Sparck 99b]. La generación automática de resúmenes se define como el proceso de extraer la información más importante de una fuente (o de varias) para producir una versión abreviada destinada a un usuario.

La generación automática de resúmenes, se realiza mediante la extracción del texto más representativo del documento original. El texto del documento está dividido en fragmentos (oraciones, párrafos, etc.), los fragmentos elegidos no sufren modificación respecto del texto original y son colocados en el nuevo documento en el mismo orden de su selección, formando así el resumen. Para identificar los fragmentos clave puede considerarse la estructura del texto y la longitud del texto; en este último nos aporta la forma en que se puede dividir; en los textos largos como libros, los fragmentos a considerar podrían ser los párrafos, en cambio para textos más cortos sería suficiente considerar las oraciones, tal es el caso de esta tesis.

2.3.1 Definición de resumen

Según [Sparck 99a], un resumen consiste en la transformación de un texto mediante la reducción de su contenido, ya sea por su selección o por generalización de lo que se considera importante. La elaboración de un resumen debe abordarse teniendo en mente las características del texto a resumir. Se han propuesto diferentes clasificaciones, para la generación automática de resúmenes [Sparck 99a], [Hovy 99a], [Nenkova y McKeown 11], [Lloret y Palmar 12]. Sin embargo, los factores predominantes en tales clasificaciones, son el medio de información (textos, imágenes, videos o voz), la entrada (un documento o múltiples documentos), la salida (extractivo o abstractivo), el propósito (genérico, personalizado, enfocado a una consulta, indicativo o informativo) y la cantidad de lenguas (monolingüe o multilingüe). En esta tesis, se consideran especialmente dos factores: el tipo de resumen que se genera en la entrada es de un solo documento y en la salida es de tipo extractivo.

Un resumen escrito es un texto que trasmite la información de otro texto de manera abreviada. Hacer resúmenes es una técnica de estudio fundamental: exige una lectura atenta y comprensiva para identificar la información más importante incluida en el documento original. Para la creación de un resumen elaborado por un humano, existen múltiples técnicas, por lo cual no existe un método establecido para poder realizarlos, sin embargo, el principal objetivo de un resumen es extraer las características más importantes del texto original de un documento y plasmarlo en un nuevo documento, en un texto más pequeño al texto original sin olvidar la representación sintética y objetivo del documento original.

De acuerdo con [Maqueo 98], se propone algunos pasos para realizar un resumen elaborado por un humano los cuales son:

1. Leer con atención un texto.
2. Separar en bloques de ideas.
3. Subrayar las ideas principales.

4. Redactar el resumen enlazando las ideas principales con las relaciones correspondientes.

Este tipo de resúmenes son llamados abstractivos.

2.3.2 Clasificación de resúmenes

Los resúmenes se pueden clasificar por su propósito [Márquez 10], de acuerdo a su entrada [Márquez 10] y por su estrategia de condensación Enrique Alfonseca et. al [Alfonseca 03], Ladda Saunmali et. al [Saunmali 11], Mine Berker [Berker 11], Iria Da Cunha [Cunha 08] y Romyna Montiel [Montiel 09] (Ver Figura 1).

Figura 1. Clasificación de la generación automática de resúmenes.

El objetivo final de un Texto Automático Resumido (por sus siglas en inglés, *ATS*), es determinar cuáles técnicas producen un resumen similar al del ser humano, actualmente hay varias

herramientas comerciales que nos ayudan en la generación automática de resúmenes, entre ellas: *Open Text Summarizer, Text Compactor, Herramienta Tools4noobs, herramienta T-Conspectus, Microsoft Office Word 2007* (en sistemas operativos: *7, VISTA, XP* y *8*) y *Microsoft Office Word 2003* (en sistemas operativos: *7, VISTA, XP* y *8*).

2.3.3 Métodos de resúmenes

De acuerdo con [Mani 01] se puede obtener dos enfoques principales en la clasificación de técnicas para la generación automática de resúmenes:

a) Distinguir entre técnicas que generan resúmenes mediante extracción y técnicas que generan resúmenes mediante abstracción, siendo por tanto el factor discriminante la existencia o no de un proceso de reescritura del resumen, utilizando técnicas de generación de lenguaje, a partir de una representación intermedia de la información contenida en el documento a resumir.

b) Distinguir, en función de la profundidad del análisis acometido y del conocimiento empleado, entre enfoques superficiales, enfoques basados en la estructura del discurso y enfoques en profundidad.

En este trabajo, se aplica la primera clasificación, ya que de acuerdo con en al análisis del trabajo, deseamos saber la calidad de la generación automática de resúmenes extractivos, mediante herramientas comerciales y métodos del estado del arte. A continuación describiremos a detalle de que trata cada uno de los dos enfoques, orientándonos más en el primer enfoque [Plaza 10]. Por último, se mencionan unos trabajos que se han desarrollado bajo estas dos técnicas.

- Las técnicas de extracción:

Generan resúmenes compuestos íntegramente por material del documento original.

-Primero: La fase de análisis, se limita a la extracción de segmentos clave del texto.

-Segundo: La fase de síntesis, se dedican a eliminar la incoherencia y la redundancia e incluso a resolver referencias anafóricas (gramática).

- Las técnicas de abstracción:

Generan resúmenes, que incluyen contenidos que no están presentes explícitamente en el texto de entrada.

-Primero: La fases de análisis, construye una representación semántica del texto fuente, mediante la identificación de conceptos genéricos y relaciones entre ellos, generalmente haciendo uso de alguna plantilla o esquema que marca la información que se considera importante de acuerdo al contexto.

-Segundo: La fase de síntesis, Implica el uso de generación del lenguaje natural para reescribir el texto que conforma al texto general.

Desventaja: trata únicamente técnicas aplicadas a dominios muy acotados.
Enfoques superficiales y enfoques discursivos
Realizan resúmenes mediante extracción.
Enfoques profundos
Utilizan técnicas de abstracción y reescritura del texto.

Uno de los problemas que existe en la generación automática de resúmenes, es identificar bajo la independencia mencionada la información más importante en el documento original, es por ello que existen diversos enfoques utilizados en la generación automática de resúmenes, como lo son los métodos y las técnicas. Algunos trabajos que utilizan técnicas por: selección de frase, basados en palabras clave y ubicación de texto (Acero 2001), Frecuencia de palabras e índice estadísticas (Cuhnna 2007), grado de importancia de las oraciones (Lee 2006), similitud de la oración (Márquez 2007), análisis lingüístico de las frases

(Matias 203). Algunos trabajos, el resumen está sujeto a la adaptación del perfil del usuario de sus requerimientos.

2.4 Pasos para la creación de un resumen extractivo

[Ledeneva 08b] propone cuatro pasos para la creación de un resumen extractivo es decir, genera un resumen de texto mediante la extracción de algunas frases del texto, mediante un enfoque independiente del dominio y del idioma, basado en el método estadístico de un solo documento para la generación automática de resúmenes (Ver Figura 2).

Muestran experimentalmente que las frases frecuentes maximales y las palabras que son partes de bigramas que se repiten más de una vez en el texto son buenos términos para describir el contenido del documento. También demuestran que la frecuencia del término como término peso da buenos resultados.

El algoritmo que proponen, consiste en una secuencia estándar de pasos, los cuales se mencionan a continuación:

- Selección de un término: Decide qué características se van a utilizar para describir las frases.
- Término de ponderación: Decide cómo se va a calcular la importancia de cada característica.
- Sentencia de ponderación: Decide cómo se va a combinar en la medida de la importancia de la oración, en la importancia de las características.
- Selección de la oración: Decide que frases son seleccionadas para el resumen.

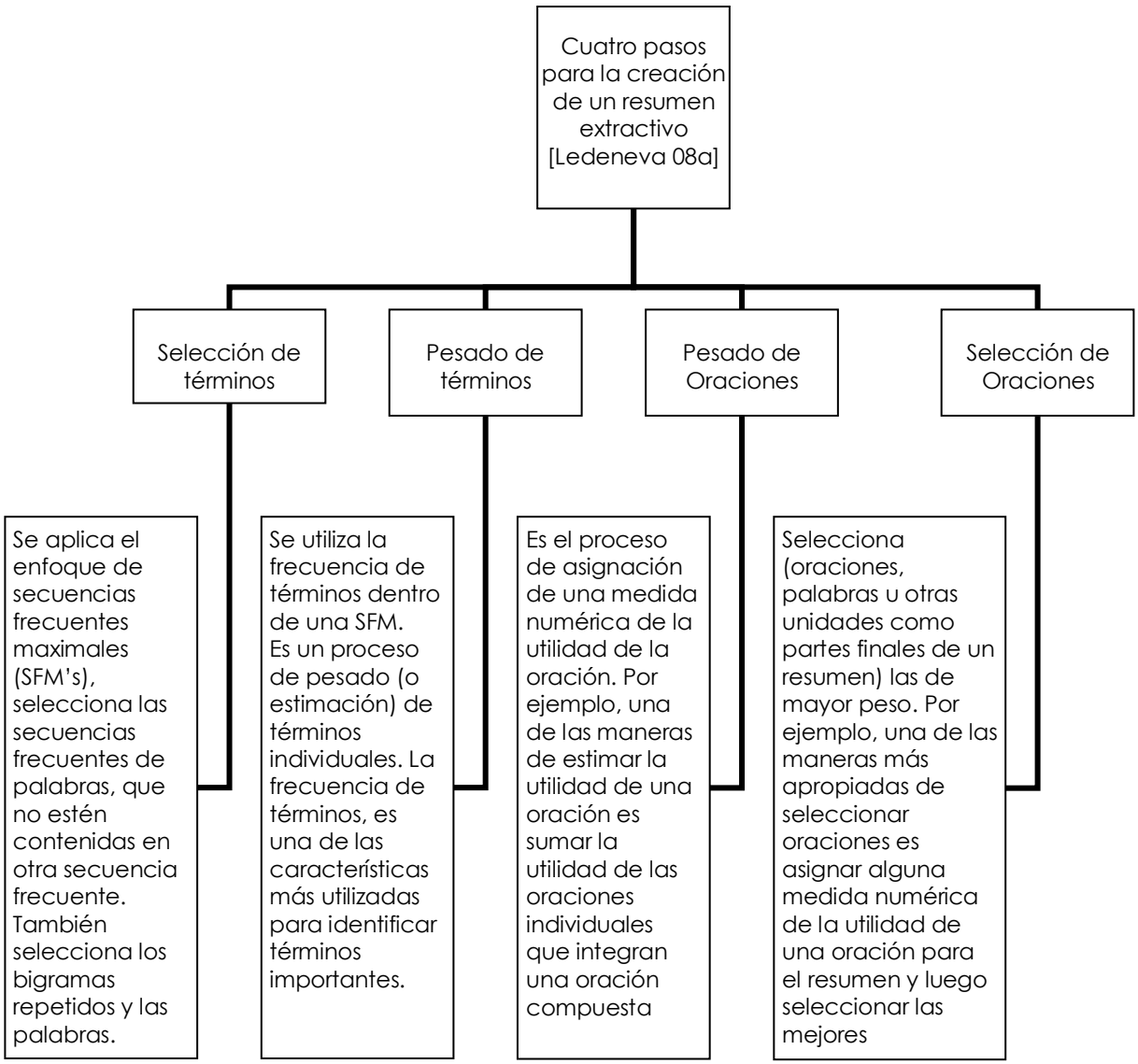


Figura 2. Identificación de cuatro pasos para componer un resumen de texto extractivo.

La generación de resúmenes, se realiza con la selección de piezas de texto (frases, oraciones, párrafos, etc.) a partir del texto original, por lo general estos se presentan al usuario en el mismo orden de selección, es decir, una copia del texto original con la mayoría de las partes de texto omitido.

Una manera de seleccionar las frases apropiadas es asignar alguna medida numérica de utilidad a una oración del resumen y luego seleccionar las mejores oraciones. El proceso de asignación de estos pesos de utilidad se llama ponderación de la oración. Una de las maneras de estimar la utilidad de una oración es sumar los pesos de utilidad de las partes individuales de los cuales consta la oración. El proceso de la estimación de los términos individuales se llama ponderación del término. Para esto, uno debe decidir lo que los términos son: por ejemplo, puede ser palabras. Decidir que partes contarán como términos es la tarea de selección de términos [Ledeneva 08b].

2.5 Método de evaluación

En esta tesis se utiliza *ROUGE* que es una herramienta que realiza las tareas de evaluación de resúmenes de forma automática. Fue creado por Lin y Hovy [Lin04] este sistema mide la similitud y determina la calidad de un resumen automático, es decir, compara un resumen modelo con un resumen candidato (o varios) y evalúa cuál de ellos es de mayor calidad.

Se eligió la herramienta ROUGE como evaluador de los resúmenes generados automáticamente por las herramientas comerciales y los métodos del estado del arte, ya que en los trabajos del estado del arte evalúan con ROUGE, logrando con esto una evaluación equitativa en los resultados.

Uno de los parámetros que contienen la herramienta de evaluación ROUGE, es que en la evaluación de sus resúmenes tiene definida la longitud de 100 palabras, por tal los resúmenes con los cuales se comparan deben de tener la misma longitud de palabras para ser equitativa la evaluación.

ROUGE incluye varias medidas automáticas que miden la similitud entre los resúmenes. Las medidas cuentan el número de unidades diferentes como *n*-gramas, secuencias de palabras y pares de palabras entre el resumen generado automáticamente para después ser evaluadas con los resúmenes ideales creados por los seres humanos.

En esta tesis la evaluación se realiza a través de las siguientes medidas (Ver Figuras 3, 4, 5).

- Precisión (P): Refleja cuantas de las oraciones extraídas por el sistema fueron correctas.

$$P=\frac{(correct)}{\#(correct + wrong)}$$

Figura 3. Ecuación de la medida Recuerdo [Lin 04].

- Recuerdo (R): Refleja cuantas oraciones correctas falló el sistema

$$R=\frac{(correct)}{\#(correct + missed)}$$

Figura 4. Ecuación de la medida Precisión [Lin 04].

- *F-Measure*

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

Figura 5. Ecuación de la medida F-measure [Lin 04].

Donde "*correct*" es el número de oraciones extraídas por el sistema y por el humano, "*wrong*" es el número de oraciones extraídas por el sistema pero no por el ser humano y "*missed*" es el número de oraciones extraídas por el ser humano, pero no por el sistema.

CAPÍTULO 3

Estado del Arte

En este capítulo, se explican algunos de los trabajos relacionados para la generación automática de resúmenes. Cabe mencionar, que las subsecciones están divididas por el trabajo a presentar. Se presentan los trabajos implementados para varios idiomas: inglés, portugués e español. Se describen los resultados obtenidos.

3.1 Trabajos relacionados

A continuación se presentan los trabajos relacionados a la generación automática de resúmenes para diferentes idiomas: inglés, portugués e español.

3.1.1 Generación automática de resúmenes usando algoritmos genéticos

En el trabajo de [Matias 13] propuso un método para la generación de resúmenes de un solo documento independiente del lenguaje. La técnica que desarrolla es mediante un algoritmo genético que utiliza el modelo de texto *n*-gramas. El corpus utilizado está en el idioma inglés. Su desarrollo fue por módulos, creados de forma independiente.

En su trabajo se realiza la comparación de bolsa de palabras y *n*-gramas, como modelos de representación de textos, para la generación automática de resúmenes.

Los parámetros propuestos a utilizar en el algoritmo genético:

- Función de aptitud: método de bolsa de palabras.
- Selección: por ruleta.
- Mutación: por intercambio.
- Condición de parada: número máximo de generación.
- Precisión: sumatoria de frecuencias de los *n*-gramas del texto original.
- Recuerdo: sumatoria de frecuencias de los *n*-gramas diferentes del resumen.

Para realizar la comparación del método propuesto se probaron las siguientes herramientas comerciales (*Microsoft Word* 2007 y 2003; *Copernic Summarizer*) y los métodos del estado del arte (*Clustering with SFMs k-means, método propuesto con algoritmo genético, SFMs-EMs, SFMs (1 best+first), baseline, SFMs (k-best), Text Rank, Baseline: random*).

Los resúmenes creados fueron evaluados con la herramienta *ROUGE,* demostrando que el método propuesto por el algoritmo genético, es mejor que las herramientas comerciales (Ver Figura 6). Además que los métodos del estado de arte superaron a las herramientas comerciales.

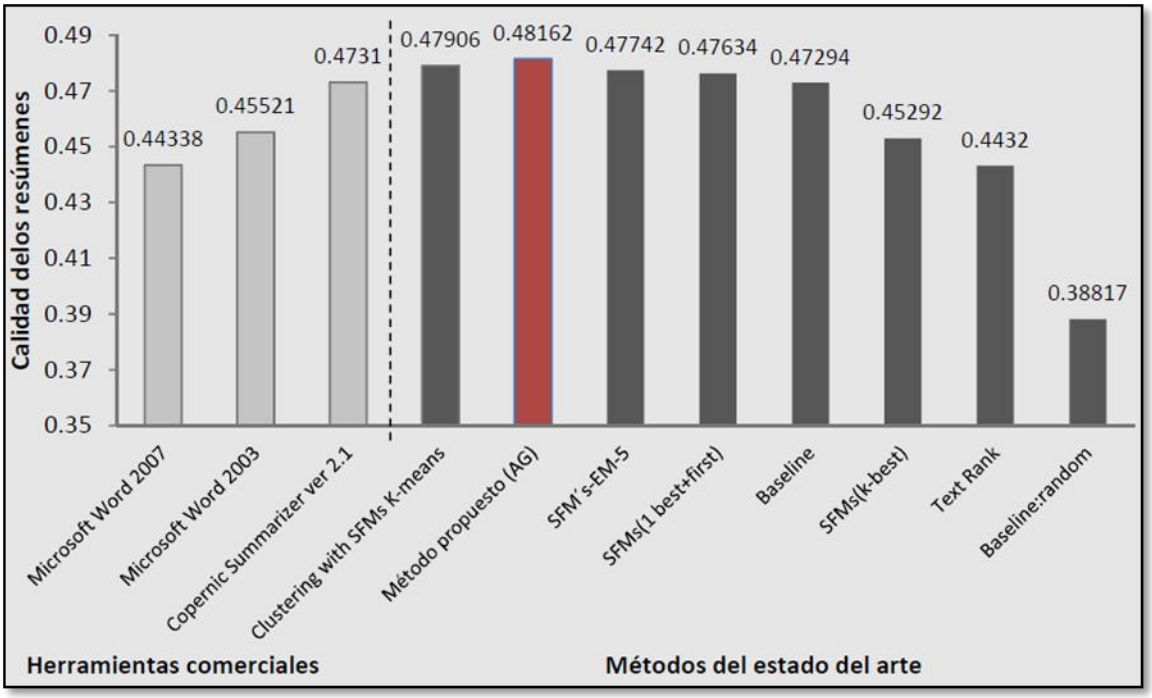


Figura 6. Resultados obtenidos con ROUGE para la colección de resúmenes obtenidos por las herramientas comerciales y los métodos del estado de arte [Matias 13].

3.1.2 Comparación de las herramientas comerciales y los métodos del estado del arte para multi-documentos

En el trabajo de [Ledeneva 11] se compararon dos tipos de métodos para la generación de resúmenes de varios documentos: las herramientas comerciales (instalables y en línea) y los métodos del estado del arte. Utilizando la colección DUC-2002 que contiene 567 artículos de noticias de diferentes longitudes en idioma inglés; la colección es sobre tecnología, alimentos, política, finanzas, etc. Para cada documento de la colección se proporcionan dos resúmenes generados por dos expertos humanos con una longitud mínima de 100 palabras. La herramienta que se utilizó para la comparación automática de resúmenes es ROUGE.

Se tomaron en cuenta los parámetros de tamaño y formato del resumen. El parámetro de tamaño del resumen depende del tamaño del documento original. El parámetro del formato depende de cómo se le presenta al usuario, en este caso se eligió mediante frases claves y de acuerdo al contexto con el cual se presenta en el documento original. Las herramientas comerciales (instalables y en línea) y los métodos del estado del arte que se evaluaron fueron las siguientes:

Herramientas instalables:

- *Copernic*
- *WORD 2007-XP*
- *WORD 2003-XP*
- *WORD 2007-7*
- *WORD 2003-7*
- *WORD 2007-VISTA*
- *WORD 2003-VISTA*

Herramientas en línea:

- *Tool4noobs*
- *Pertinence*
- *Svhoong*
- *OTS*

Métodos del estado del arte:

- *1st Best Method*
- *2nd best Method*
- *3rd best Method*
- *MFS*
- *4th best Method*
- *5th best Method*
- *Baseline*

Se comprobó que la herramienta instalable *Copernic Summarizer* obtiene los mejores resultados. Mientras que la herramienta en línea *Shvoong* es la que obtiene mayor calidad de los resúmenes en línea (Ver Figura 7). Mediante esta comparación lograron ver que los resultados de la herramienta comercial instalable *Microsoft Office Word* son incoherentes, ya que la calidad de los resúmenes depende del sistema operativo. La evaluación se realizó mediante 18 métodos del estado del arte y herramientas comerciales, demostrando que los métodos son mejores que las herramientas comerciales.

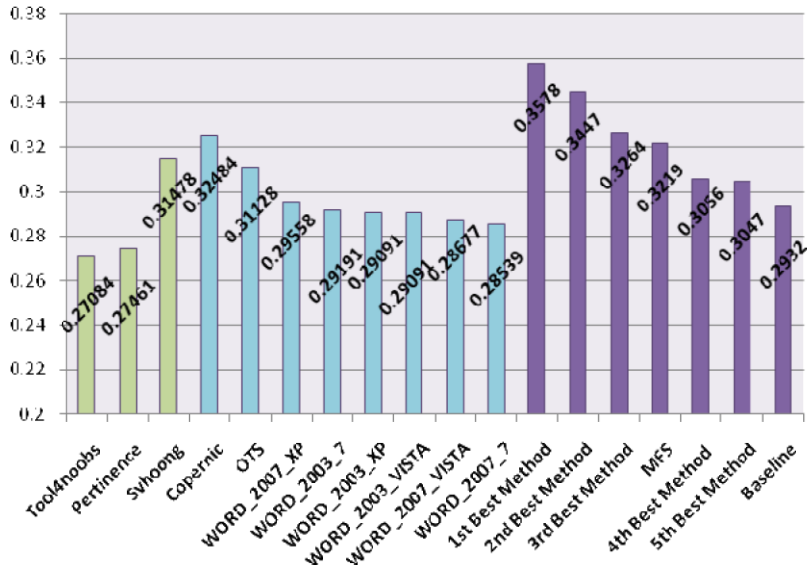


Figura 7. Los resultados de la colección de documentos obtenidos por las herramientas comerciales y el de los métodos de estado de arte [Ledeneva 11].

3.1.3 Resúmenes extractivos utilizando aprendizaje no supervisado

[García 08] propone un algoritmo de aprendizaje no supervisado, que sustituye los pasos de ponderación de oraciones y etapa de selección de oraciones propuestos en [Ledeneva 08]. Ayudando a detectar automáticamente los grupos de oraciones similares de los cuales se selecciona la oración más representativa y con ellas ir formando el resumen extractivo independiente del idioma y del contexto. Se utilizó la colección DUC-2002.

Los pasos generales que se propone en este trabajo son:

1. Selección de términos: Utiliza los n-gramas
2. Ponderación de términos: Utiliza ponderación booleana aquí se modela la presencia o ausencia de un término en el documento, definido como:
    - 2.1 Frecuencia de los términos (TF):
    - 2.2 Frecuencia inversa de documentos (IDF):
    - 2.3 TF-IDF

En la Figura 8, se puede ver que el mejor resultado se obtuvo con n-gramas de tamaño 10, con el pesado de IDF, lo cual da solución al problema de distinguir entre los dos documentos del mismo vocabulario.

Selección de	Pesado de términos			
términos	BOOL	TF	IDF	TFIDF
1 grama	0.47264	0.47439	0.47387	0.47298
2 gramas	0.47445	0.47436	0.47519	0.47517
3 gramas	0.47683	0.47684	0.47673	0.47673
4 gramas	0.47638	0.47638	0.47661	0.47658
5 gramas	0.47689	0.47689	0.47658	0.47658
6 gramas	0.47729	0.47725	0.47762	0.47762
7 gramas	0.47721	0.47738	0.47713	0.47739
8 gramas	0.4786	0.47826	0.47818	0.47808
9 gramas	0.47845	0.47845	0.47773	0.47868
10 gramas	0.47803	0.47842	0.47906	0.47854
11 gramas	0.47753	0.47642	0.47599	0.47599

Figura 8. Evaluación de algoritmo no supervisado para generar resúmenes de 100 palabras [Ledeneva 08].

3.1.4 Comparación de tres modelos de texto para la generación automática de resúmenes

En el trabajo [García 09] se comparan tres modelos de textos para la generación automática de resúmenes de un solo documento. El primer modelo es basado en SFMs, el segundo modelo es con bolsa de palabras o 1-gramas, y el tercer modelo con bigramas o 2-gramas.

En la realización de sus pruebas utilizaron la colección de documentos estándar DUC-2002 la cual contiene 567 noticias sobre diversos temas y de diferentes longitudes. Utilizando la herramienta *ROUGE* para evaluar automáticamente los resúmenes generados. Para la comparación de los tres modelos se utilizó el método propuesto por [García 08], el cual es independiente del dominio y del lenguaje.

El algoritmo utilizado para este trabajo consiste de los siguientes pasos:

- Pre-procesamiento:
  Elimina palabras vacías (aplicación del algoritmo *Stemming)*.

- Selección de términos:
  Se elige el modelo para representar el texto (bolsa de palabras, *n-gramas* o SFMs).

- Pesado de términos:
  Se elige uno de los cuatro métodos para calcular la importancia de cada término: BOOL, TF, IDF y TF-IDF

- Agrupamiento de oraciones:
  Se realiza la selección de centroides para cada grupo ya sea aleatoriamente o por medio del *baseline*.

- Selección de oraciones:
  Se selecciona la oración más cerca al centroide del grupo para ir formando el resumen.

Como se puede observar (Ver Figura 9), el modelo de bigramas con pesado booleano obtuvo los mejores resultados a comparación de otros modelos. También mostraron que el pre-procesamiento de los documentos no es tan relevante en la generación automática de resúmenes extractivos.

Modelo de texto		Pesado de términos			
		BOOL	TF	IDF	TF-IDF
CP	BDP	0.4726	0.4743	0.4738	0.4729
	2-gramas	0.4744	0.4743	**0.4751**	**0.4751**
	SFM's	0.4702	0.4686	0.4705	0.4698
SP	BDP	0.4729	0.4721	0.4719	0.4737
	2-gramas	**0.4757**	0.4752	0.4753	0.4752
	SFM's	0.4687	0.4683	0.4707	0.4736

Figura 9. Evaluación con los tres modelos [García 09].

3.1.5 Generación automática de resúmenes mediante aprendizaje no supervisado

Así mismo la técnica de pre-procesamiento que utiliza [Montiel 09] es una representación y agrupamiento de documentos combinando varias técnicas de minería de texto, demostrando que el mejor resumen es con el modelo de *n*-gramas, *baseline* y sin pre-procesamiento, y aplicando el pesado de términos booleano.

Su método combina técnicas de pre-procesamiento y representación de textos. En la técnica de pre-procesamiento aplica la extracción de palabras poco relevantes para el contenido del texto y en la técnica de representación de los textos, se apoya de la extracción de secuencias de palabras, comparando tres modelos: bolsa de palabras, n-gramas y secuencia de palabras, basándose en la frecuencia en que aparecen en el documento. Para ayudar un poco al llenado del resumen se realiza un agrupamiento de oraciones similares y de cada grupo se selecciona la mejor oración para formar el resumen.

Como se observa en la gráfica (Ver Figura 10), se exponen los valores de *F-measure* obtenidos con todos los modelos de representación de textos aplicados, mostrando que los resúmenes obtenidos con los modelos donde se utilizó *baseline* fueron de buena calidad, pero aquellos que se obtuvieron con el modelo *de n-gramas* superaron a los generados con bolsa de palabras y SFMs. Sin embargo, el mejor valor de *F-measure* alcanzado con el modelo de *n*-gramas, demostró que la combinación de *baseline* con el algoritmo de *k*-medias y la eliminación del pre-procesamiento ha permitido obtener los mejores resúmenes.

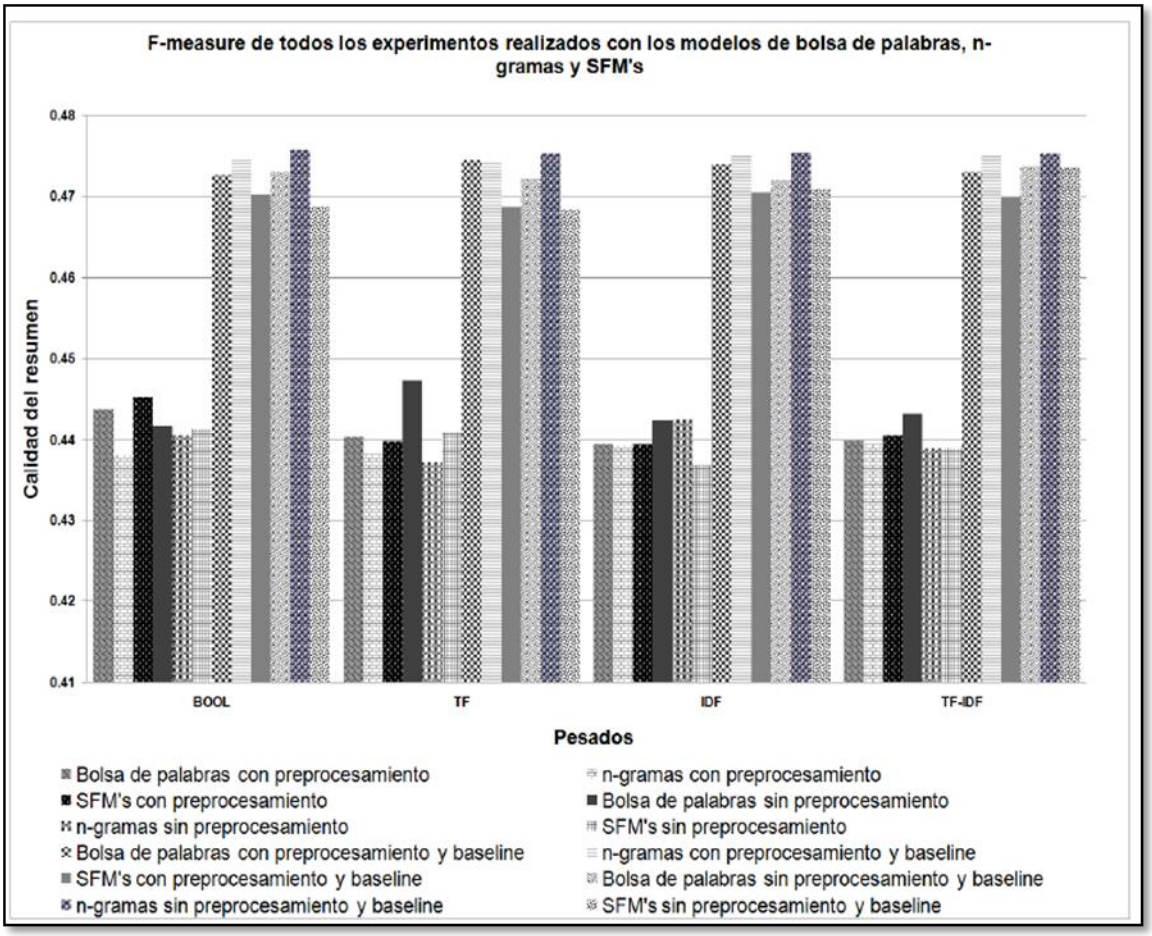


Figura 10. Evaluación mediante ROUGE para la generación automática de resúmenes mediante aprendizaje no supervisado [Montiel 09].

3.1.6 Generación automática de resúmenes de múltiples documentos

En el trabajo de [Villatoro 06] se implementan dos arquitecturas para la generación de resúmenes automáticos para múltiples documentos, mediante la clasificación a nivel de oraciones. La primera arquitectura, genera el resumen por cada documento, basándose en técnicas de aprendizaje supervisado, selecciona la información más relevante de cada documento dentro de la colección inicial usando secuencias de palabras. Los resúmenes generados son utilizados como colección de entrada al segundo módulo en el que se desarrolla bajo un esquema de un algoritmo de agrupamiento (*clustering*), el cual organiza la información por sub-temas, eliminando redundancias y controlando los niveles de comprensión, aplicando una técnica de aprendizaje no supervisado.

Para la realización de las pruebas utiliza tres corpus diferentes, el primer corpus está en idioma español mexicano conformado por 300 noticias sobre desastres, el segundo corpus llamado CAST conformado por 164 artículos de noticias de temas variados y, por último, el corpus DUC2003 que contiene 660 documentos en idioma inglés.

Para la construcción de los resúmenes [Villatoro 06] emplea una metodología en la cual los documentos de entrada tienen que estar relacionados temáticamente. Los resúmenes son construidos a base de extractos del documento original y para su selección utiliza un modo de representación de las oraciones.

Sus experimentos fueron divididos en dos etapas. Primero, evaluaron la etapa supervisada por medio de la estrategia de validación cruzada (*cross fold validation*) y además se muestran los resultados del clasificador al tratar con un conjunto de datos no vistos.

Segundo, dado que la salida de la segunda etapa es el resumen multi-documento, ésta es evaluada por medio de *ROUGE*.

Se consideraron tres diferentes medidas de similitud para formar la matriz de similitudes y para definir el valor del umbral.

[Villatoro 06] comprobó que usando un umbral alto o fuerte con la medida de similitud (DICE) permite al sistema formar resúmenes de mayor calidad. El mejor resultado obtenido de *F-measure* es 0.40981. Creando un resumen que coincide hasta un 40% de su contenido comparados con los resúmenes ideales creados por humanos.

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

Otros trabajos como la técnica que propone [Ibáñez 13] es la generación automática de resúmenes mediante herramientas comerciales aplicado al corpus TeMário en el idioma portugués. Se realiza la comparación de herramientas en línea (*Shvoong, Tools4noobs, GistSumm y Open Text Summarizer*) y herramientas instalables (*Microsoft Office Word 2003* y *Microsoft Office Word 2007*, en sistemas operativos *Windows*: *XP*, *VISTA*, *7 Ultimate* y *8*). Demostrando que en la comparación de las herramientas comerciales, la herramienta en línea *GistSumm* en su 3ra opción (realización de resumen basado en consultas) es la que realiza con mayor calidad los resúmenes automáticos para el idioma portugués (Ver Figura 11).

Su metodología de desarrollo que utilizó fue:

1.- Selección de las herramientas comerciales para la generación de resúmenes automáticos.
2.- Descripción de los métodos del estado del arte.
3.- Descripción de la colección de documentos a evaluar (TeMário).
4.- Descripción del sistema para la evaluación de los resúmenes (*ROUGE*).
5.- Selección de las opciones de las herramientas comerciales para la generación de resúmenes automáticos.
6.- Proceso de generación de los resúmenes automáticos
- *Microsoft Office Word 2003* (en los sistemas operativos *XP*, *VISTA*, *7* y *8*)
- *Microsoft Office Word 2007* (en los sistemas operativos *XP*, *VISTA*, *7* y *8*)
- Herramienta *GistSumm*
- *Shvoong*
- *Tools4noobs*
- *OTS*

7.- Proceso de comparación de los resúmenes generados por las herramientas y métodos.
8.- Presentación de resultados.

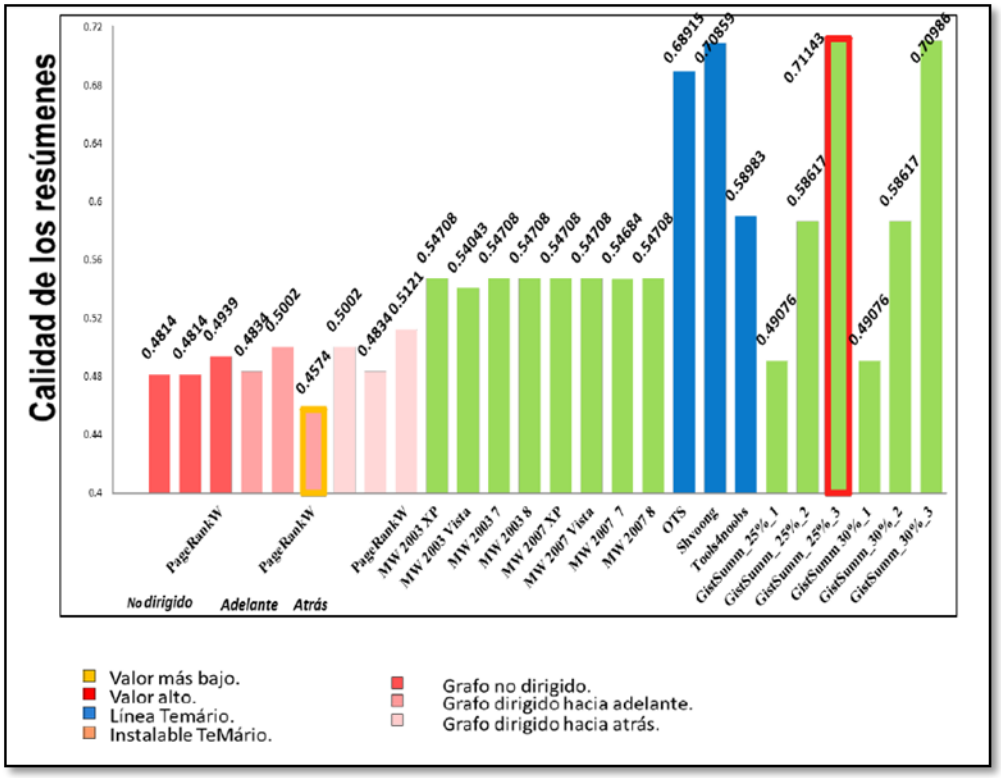


Figura 11. Resultados de evaluación mediante ROUGE para la comparación de las herramientas comerciales para la generación automática de resúmenes de textos con el corpus TeMário [Ibáñez 13].

3.1.8 Generación automática de resúmenes independientes del lenguaje

El trabajo [Matias 16] está enfocado a la generación automática de resúmenes independientes del lenguaje, se propuso superar la calidad de [Matias 13] además, se realiza la comparación de herramientas comerciales y métodos del estado de arte.

Se emplean tres corpus de diferente idioma: el corpus en idioma portugués el cuál utiliza la colección TeMário, el corpus en idioma inglés el cual contiene la colección DUC2002 y el corpus en idioma español TER el cual fue creado en el desarrollo de ese trabajo especialmente para el uso de resúmenes.

La metodología que utiliza está dividida en cuatro etapas, las cuales se describen a continuación:

- Colecciones:
  - -DUC2002 (inglés)
  - -TeMário (portugués)
  - -TER2014 (español)
- Parámetros:
  - -Pre-procesamiento
  - -Modelo de texto
  - -Importancia de las oraciones
  - -Función de aptitud
  - -Operador de selección
- Pruebas
- Evaluación

A continuación se muestra los parámetros con los cuales se evaluó cada uno de los corpus:

Con la colección en idioma inglés el método propuesto por [Matias 16] supero los resultados de las herramientas comerciales y los métodos del estado del arte (Ver Figura 12). Los parametros establecidos para este corpus son los siguientes:

-Pre-procesamiento: si
-Modelo de texto: bolsa de palabras
-Importancia de las oraciones: [Vázquez 15]
-Función de aptitud: $0.6\beta+0.4\delta$
-Operador de selección: ruleta

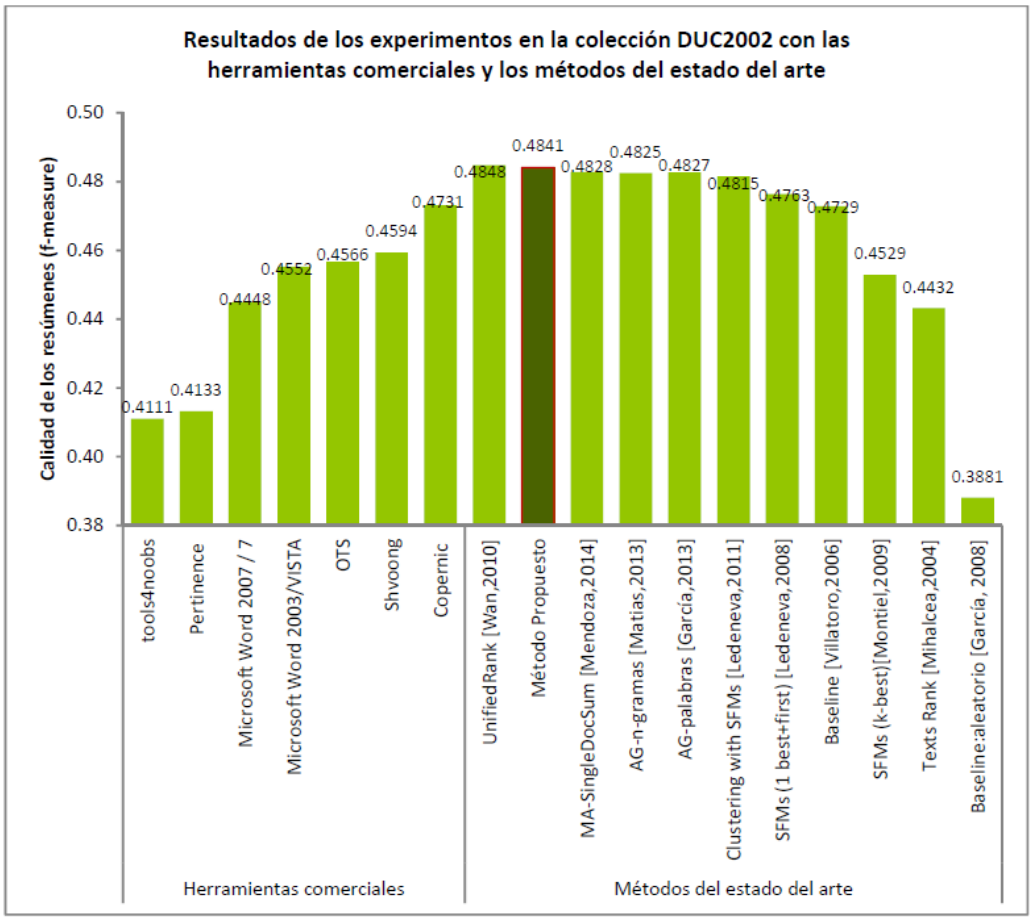


Figura 12. Resultados de herramienta ROUGE, con la colección DUC2002 con las herramientas comerciales y los métodos del estado del arte [Matias 16].

El método propuesto por [Matias 16] supera a las herramientas comerciales y a los métodos del estado de arte con la colección en idioma portuges (Ver Figura 13). Los parámetros para esta colección quedaron de la siguiente forma:

-Pre-procesamiento: si
-Modelo de texto: bolsa de palabras
-Importancia de las oraciones: [Vázquez 15]
-Función de aptitud: $0.5\beta+0.5\delta$
-Operador de selección: ruleta

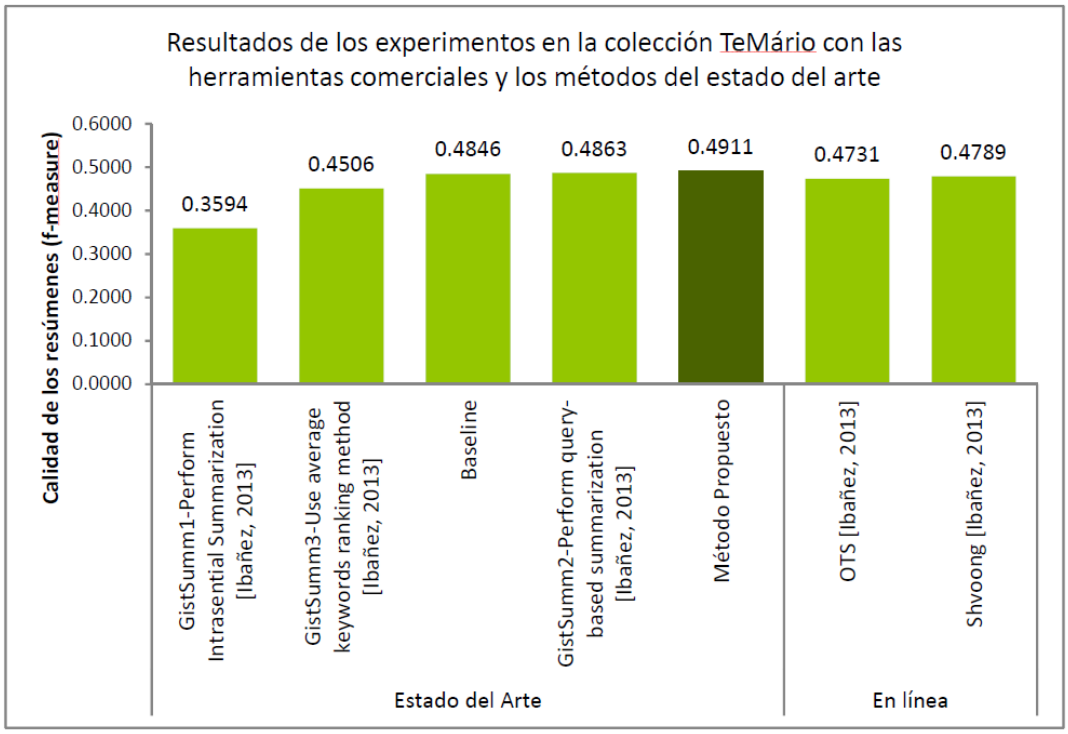


Figura 13. Resultados de herramienta ROUGE, con la colección TeMario con las herramientas comerciales y los métodos del estado del arte [Matias 16].

Al igual que las otras colecciones de diferente idioma, con la colección en idioma español es superada por el método propuesto por [Matias 16]. Los parámetros establecidos para esta prueba son:

-Pre-procesamiento: no

-Modelo de texto: n-gramas (n=5)

-Importancia de las oraciones: [Vázquez 15]

-Función de aptitud: $0.4\beta+0.6\delta$

-Operador de selección: ruleta

Con los parámetros mostrados anteriormente se puede ver que en el trabajo de [Matias 16] supera tanto a las herramientas comerciales como a los métodos del estado del arte en los tres corpus de diferente idioma. Cabe destacar, que este método [Matias 16] se toma como referencia para comprobar la calidad de las herramientas comerciales con el corpus *TEXTRUSS*.

CAPÍTULO 4

Metodología de trabajo

En este capítulo, se presenta la metodología propuesta para la evaluación de las herramientas comerciales y los métodos del estado de arte para el corpus *TEXTRUSS*. Se describe cada etapa de la metodología propuesta.

4.1 Metodología de trabajo

En este capítulo, se proponen las etapas que se tiene que llevar a cabo para la comparación de las herramientas comerciales y los métodos del estado de arte para el corpus *TEXTRUSS*. Estas etapas componen la metodología propuesta que consta de 6 fases que se muestran en la Figura 14.

4.1.1 Fase 1. Creación del corpus

Primero, se realizó la búsqueda de un sitio web de noticias adecuado a las siguientes características:

- Los artículos tienen que ser de diferentes tamaños.
- Los artículos tienen que ser de diferentes temáticas.
- Tiene que contener las partes importantes del texto marcadas que pudieran ser consideradas como un resumen de la noticia.
- Los resúmenes marcados tienen que contener más de 100 palabras.
- Los artículos completos tienen que cumplir con una longitud mayor a la de los resúmenes generados.
- Los artículos que son de recetas de comida, no se toman en cuenta ya que en la mayoría no contienen la descripción sino los nombres de ingredientes y su peso.

Una vez encontrado el sitio web en el idioma ruso que se llama gazeta.ru [Gazeta.ru 15], se realizó la descarga de los artículos conformados por once categorías. Cada categoría contiene 22 artículos de diferentes longitudes. Una vez teniendo los artículos descargados se seleccionó el resumen basado en las partes marcadas por el experto que escribió la noticia.

Segundo, la organización del corpus *TEXTRUSS* tiene que responder a las siguientes características:

- La estructura de cada documento de la noticia tiene que marcarse con las etiquetas correspondientes.
- La organización del corpus *TEXTRUSS* tiene que realizarse a través de varios formatos que se requieren ser de la siguiente manera:

1. Con título: Contiene el artículo completo que lo compone: título de la noticia, una muy breve descripción de lo que trata la noticia (abarca máximo dos renglones), el autor de la fotografía, fecha de publicación de la noticia, autor de la noticia, resumen de la noticia, noticia, etc.
2. Sólo resumen de la noticia: Contiene el resumen de la noticia.
3. Sin título: Contiene la noticia con el resumen, se omiten partes del encabezado como son el autor de la fotografía, fecha de publicación de la noticia, descripción de la noticia, etc.

- Para asignar nombre representativo a cada uno de los archivos, se definió el siguiente formato:
  Numerodelarchivo | primerasdosletrasdelacategoria | fechadeconsulta | guionbajo | clavedeURLdelanoticia. Ejemplo 01AU170815_7692628
- En total son 11 carpetas nombradas de acuerdo a las 11 categorías correspondientes.
- Dentro de cada carpeta se encuentran 22 artículos correspondientes a la categoría.

Tercero, para la realización de las pruebas, la organización del corpus es la siguiente:

- Se tiene la carpeta nombrada "Corpus *TEXTRUSS*", la cual contiene el corpus *TEXTRUSS*.
- Otra carpeta nombrada "Métodos", en la cual contiene los archivos modificados del corpus *TEXTRUSS* para ser procesados con los métodos del estado del arte.
- Otra carpeta con el nombre "Herramientas", la cual contiene todos los resúmenes de las herramientas comerciales (instalables y en línea) realizados bajo los parámetros establecidos.

Figura 14. Metodología de trabajo.

4.1.2 Fase 2. Selección de herramientas comerciales y métodos del estado del arte

En esta fase, se eligen los resúmenes con cada una de las herramientas comerciales y los métodos del estado de arte que cumplieron con los requerimientos especificados en la sección anterior.

4.1.3 Fase 3. Determinación de parámetros

Tomando en cuenta los trabajos del estado del arte, se realiza la búsqueda de las herramientas comerciales (instalables y en línea) que están disponibles para su descarga e instalación y las páginas web que estuvieran disponibles. Una vez determinadas las herramientas en línea se realiza las investigación de cada herramienta como por ejemplo: si realizan resúmenes en idioma ruso, si el resumen que elabora cada herramienta es mayor a 100 palabras, si la coherencia del resumen estaba bien estructurados por lo tanto era coherente, etc. Una vez que la herramienta cumple con los requerimientos establecidos se diseña una fórmula para que el resumen resultante fuera mayor de 100 palabras. Esta fórmula se aplicaba en el umbral de cada una de las herramientas.

4.1.4 Fase 4. Realización de resúmenes

Una vez obtenidos los parámetros se realizaron los resúmenes, tanto para las herramientas comerciales seleccionadas como para los métodos del estado del arte.

4.1.5 Fase 5. Evaluación

Una vez teniendo los resúmenes de las herramientas comerciales y los métodos del estado del arte, se realizó la evaluación con cada una de estas mediante la herramienta *ROUGE* [Lin 04].

4.1.6 Fase 6. Comparación de los métodos del estado de arte y las herramientas comerciales

En esta última fase se realizó el análisis de la evaluación de cada herramienta, así como el de los métodos del estado del arte.

4.1.7 Fase 7. Conclusiones

En esta fase, se presentan las conclusiones de la tesis. Para poder determinar cuál de las herramientas comerciales y los métodos del estado del arte realizan los resúmenes con mayor calidad para el corpus TEXTRUSS.

CAPÍTULO 5

Resultados

En este capítulo, siguen los pasos de la metodología propuesta. Se describen la creación y la organización del corpus *TEXTRUSS*. Se detallan los parámetros y la selección de herramientas comerciales y los métodos del estado del arte. Los resultados que se obtuvieron para las herramientas comerciales y los métodos del estado del arte. Finalmente, se efectúo la comparación de las herramientas comerciales y los métodos del estado del arte.

5.1. Corpus TEXTRUSS

5.1.1. Creación del corpus TEXTRUSS

El corpus está compuesto por artículos de noticias con su resumen, realizados por un humano experto en el idioma ruso. Las noticias fueron descargadas del portal de noticias gazeta.ru [Gazeta.ru 15]. El corpus es de diferentes dominios y contiene 11 categorías de la siguiente manera:

1. ПОЛИТИКА (POLÍTICA)
2. БИЗНЕС (NEGOCIOS)
3. ОБЩЕСТВО (COMPAÑIA)
4. МНЕНИЯ (CRITICAS)
5. КУЛЬТУРА (CULTURA)
6. НАУКА (CIENCIA)
7. ТЕХНОЛОГИИ (TECNOLOGÍA)
8. НЕДВИЖИМОСТЬ (BIENES INMUEBLES)
9. АВТО (AUTO)
10. СТИЛЬ ЖИЗНИ (ESTILO DE VIDA)
11. СПОРТ (DEPORTES)

De cada categoría se obtuvieron 22 artículos. En total la colección contiene 242 artículos.
Las partes de la estructura de cada artículo son las siguientes (Ver Figura 15):

«Пятый элемент» покажется вам короткометражкой» Título de la noticia.

Люк Бессон снимет научно-фантастический фильм Referencia de la noticia.

Фотография: Ли Джин-человек Autor de la fotografía

Виктория Сеничкина 13.07.2015, 16:01 Fecha de publicación de la noticia | autor de la noticia

Люк Бессон рассказал о том, что приступит к съемкам нового научно-фантастического фильма «Валериан и город тысячи планет», и показал эскизы к будущей картине. Французский режиссер признался, что на этот его замысел значительно повлиял «Аватар», снятый Джеймсом Кэмероном. Resumen de la noticia.

В основе будущего фильма — комиксы «Валериан и Лорелин» французского писателя Пьера Кристиана и иллюстратора Жан-Клода Мезье. Картина расскажет о приключениях путешествующего во времени и пространстве межгалактического агента XXVI века и его спутницы Лорелин, которые работают в пространственно-временной службе по защите человечества от преступников. Они живут на космическом корабле, диаметр которого составляет 12 миль, его населяют миллионы различных форм жизни. Noticia.

Режиссер обещает, что он постарается избежать штампов, когда главный злодей появляется в первые десять минут фильма, а зритель заранее знает, чем и как все закончится. Resumen de la noticia.

Figura 15. Estructura del artículo 10CU140815_7654545.TXT de la colección del corpus TEXTRUSS

Para la construcción del corpus TEXTRUSS, después de descargar los artículos, se realizó la clasificación de cada artículo. Los textos originales son llamados textos-fuente mientras que los resúmenes de cada uno de ellos son llamados los resúmenes.

5.1.2. Organización del corpus

El corpus contiene 3 formatos diferentes (Ver Figura 16):

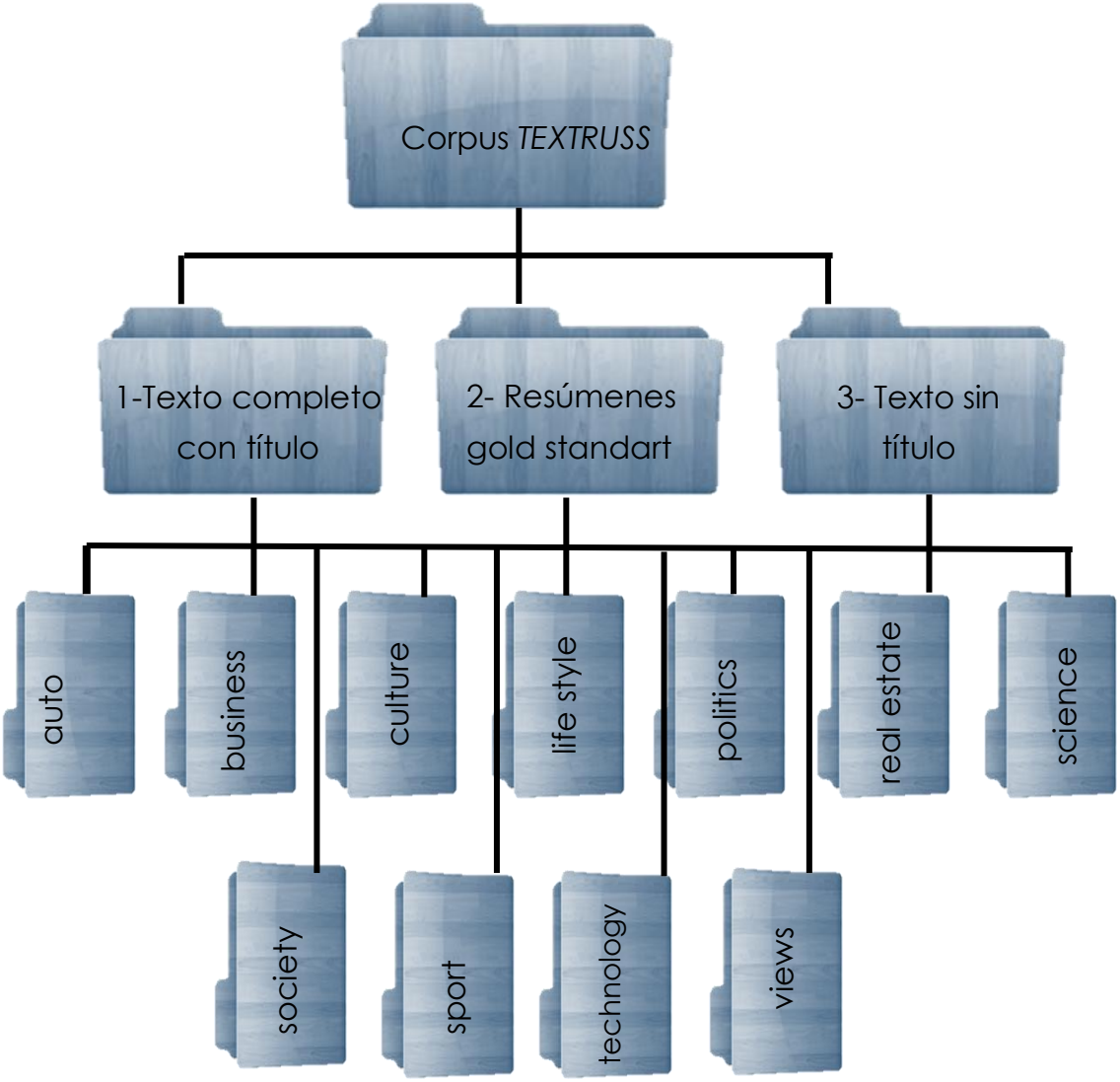


Figura 16. Estructura del corpus TEXTRUSS.

1. Texto completo con título: contiene las noticias descargadas del portal sin ninguna modificación.
2. Resúmenes *gold standard:* contiene solo los resúmenes de cada artículo de noticia.
3. Texto sin título: contiene las noticias a partir de la narración del artículo omitiendo el encabezado (título de la noticia, referencia, autor de la fotografía, fecha de publicación de la noticia y autor de la noticia).

En cada formato el corpus incluye 11 categorías.

5.2 Determinación de parámetros

A continuación se describen los parámetros de las herramientas comerciales, las cuales fueron evaluadas para el corpus *TEXTRUS*.

Para garantizar que cada una de las herramientas comerciales no disminuyera la calidad de los resúmenes, fue necesario calcular el porcentaje adecuado para producir resúmenes con más de 100 palabras. Se implementaron las fórmulas (Ver Figuras 17 y 18).

Para generar resúmenes con 100 palabras en las herramientas: Open Text *Summarizer* [OTS 15], *Text Compactor* [TextCompactor 15], *Microsoft Office Word* [MOW 15] se utilizó la siguiente fórmula tomada de [Matias 13] (Ver Figura 17):

$$\frac{\text{Número de palabras deseadas}}{\text{Número de palabras totales en el documento}} * 100$$

Figura 17. Generación de resúmenes de 100 palabras.

Para generar resúmenes con 100 palabras en la herramienta: *Tools4noobs* [T4NS 15] se aplicó la siguiente formula (Ver Figura 18):

$$\left(\frac{\text{Número de palabras deseadas} * 100}{\text{Número de palabras totales del documento} - 80} \right. * -1$$

Figura 18. Generación de resúmenes con la herramienta Tools4noobs.

Cabe mencionar, que a diferencia de las otras herramientas el umbral de Tools4noobs funciona al contrario.

Para la herramienta *T-Conspectus* se eligió el 20% de umbral ya que es el rango óptimo para conseguir resúmenes con más de 100 palabras.

En el caso de la implementación del corpus *TEXTRUSS* al método [Matias 16], el cual propone un nuevo método para la generación de resúmenes extractivos, mediante un modelo evolutivo. Se investigó el formato de los archivos para ser procesados mediante el algoritmo genético.

5.3 Selección de las herramientas comerciales y métodos del estado del arte

En esta sección, se describirán las herramientas comerciales y los métodos del estado del arte que se eligieron para el desarrollo de esta tesis. Para revisar a detalle la descripción de cada una de las herramientas comerciales seleccionadas, dirigirse al Anexo 1.

Después de investigar cada una de las herramientas comerciales, tanto instalables como en línea y los métodos del estado del arte estuvieran disponibles, además cumplieran con los parámetros establecidos, se eligieron para la realización de los experimentos.

5.3.1 Herramientas comerciales

A continuación se muestran las herramientas comerciales, las cuales fueron evaluadas para el corpus *TEXTRUSS*.

Herramientas en línea:

- *Open Text Summarizer [OTS 15]*
- *Text Compactor [TextCompactor 15]*
- *Tools4noobs [T4NS 15]*
- *T-Conspectus [Conspectus 15]*

Herramientas instalables:

Sistema Operativo *Windows XP* [MOW 15]
- *Microsoft Word 2007*
- *Microsoft Word 2003*

Sistema Operativo *Windows VISTA* [MOW 15]
- *Microsoft Word 2007*
- *Microsoft Word 2003*

Sistema Operativo *Windows 7* [MOW 15]
- *Microsoft Word 2007*
- *Microsoft Word 2003*

Sistema Operativo *Windows 8* [MOW 15]
- *Microsoft Word 2007*
- *Microsoft Word 2003*

En el caso de los métodos del estado del arte se tomaron en cuenta los mejores métodos:
-Comparación de bolsa de palabras, n-gramas y secuencias frecuentes (utilizando algoritmo genético) [Matias 13], [Matias 16].
-*1st Best Method, 2nd Best Method, 3rd Best Method, MFS, 4th Best Method, 5th Best Method, baseline* [Ledeneva 11]
-*n*-gramas, ponderación booleana [García 08]

-SFMs, bolsa de palabras y bigramas [García 09]
-*n*-gramas, bolsa de palabras, y SFMs [Montiel 09]
-Algoritmo *PageRank* [Mihalcea 04]

Estos trabajos se comparan en capítulo 3.

5.4 Evaluación de las herramientas comerciales y los métodos del estado del arte

5.4.1 Herramientas instalables (Microsoft Office Word)

Una vez que se tuvo la extracción de los resúmenes de cada una de las herramientas comerciales y los resumes del método el estado del arte, se procedió a realizar la evaluación con el sistema ROUGE, a continuación se muestran las gráficas.

Primero se muestra la comparación de las evaluaciones de las herramientas comerciales instalables. Después se muestra la evaluación de las herramientas comerciales en línea, en seguida los métodos del estado del arte y por último se presenta el conjunto de los resultados de todas las evaluaciones tanto de las herramientas comerciales, como del método del estado del arte y el *baseline* del corpus *TEXTRUSS*.

Como se aprecia en el gráfico (Ver Figura 19) el sistema operativo *Windows XP* con la paquetería de *Microsoft Word 2007* muestra menor calidad en los resumenes, a comparacion de todas las herramientas comerciales instalables.

En el sistema operativo *Windows 7* con la paquetería de *Microsoft Word 2007,* muestra que los resumenes son de mayor calidad en *todas las herramientas comerciales instalables*.

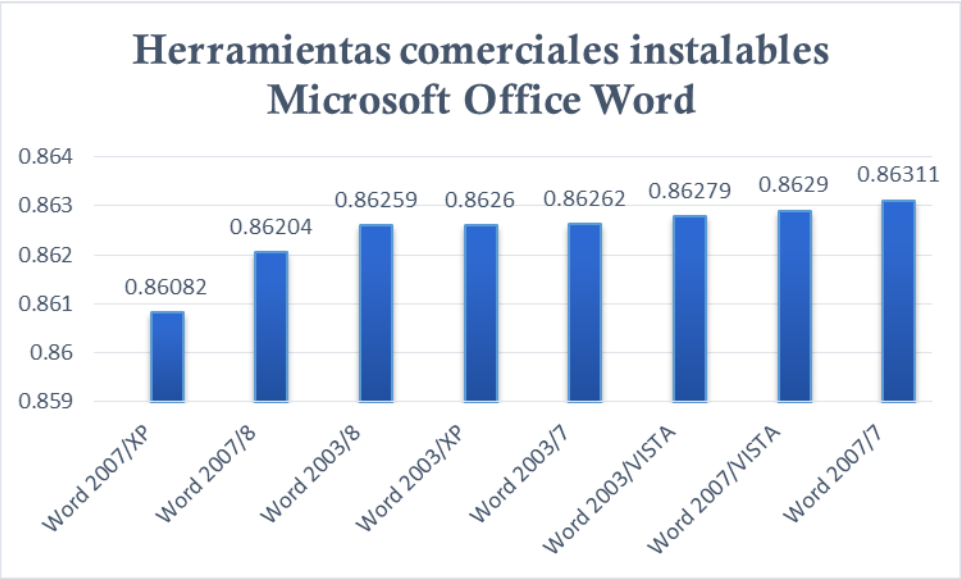


Figura 19. Evaluación de las herramientas instalables Microsoft Word con ROUGE.

5.4.2 Herramientas comerciales en línea

Los resultados de la evaluación de las herramientas en línea fueron los siguientes (Ver Figura 20):

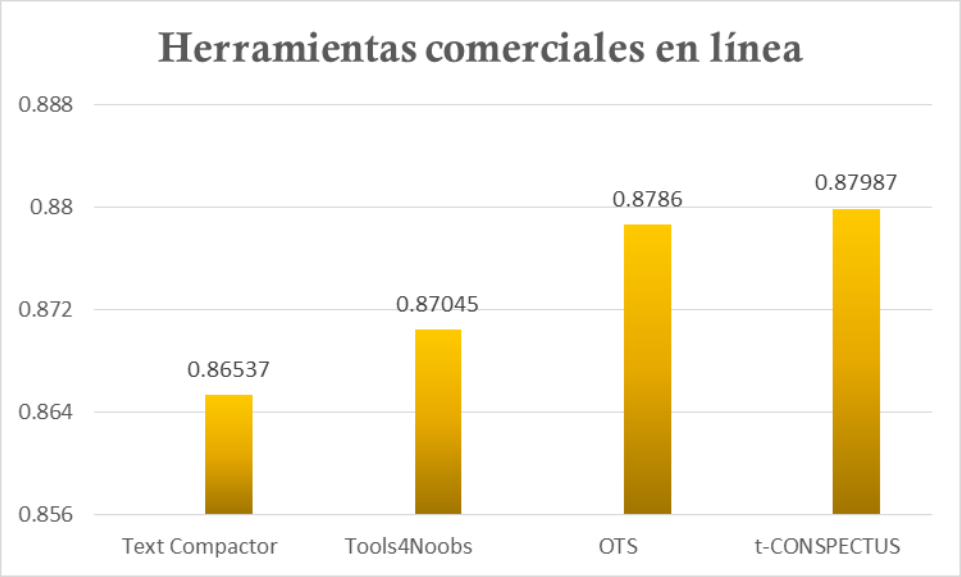


Figura 20. Herramientas en línea evaluadas con ROUGE.

Como se puede ver en la Figura 20 la herramienta en línea que obtuvo menor calidad en la elaboración de resumenes con el corpus TEXTRUSS, es la herramienta *Text Compactor* [TextCompactor 15].

Mientas la herramienta para la elaboración de resumenes con mayor calidad es la herramienta *T-Conspectus* [Conspectus 15].

5.4.3 Herramientas comerciales instalables y en línea

Para ver con mayor claridad la comparación de las herramientas comerciales se muestra la siguiente gráfica (Ver Figura 21):

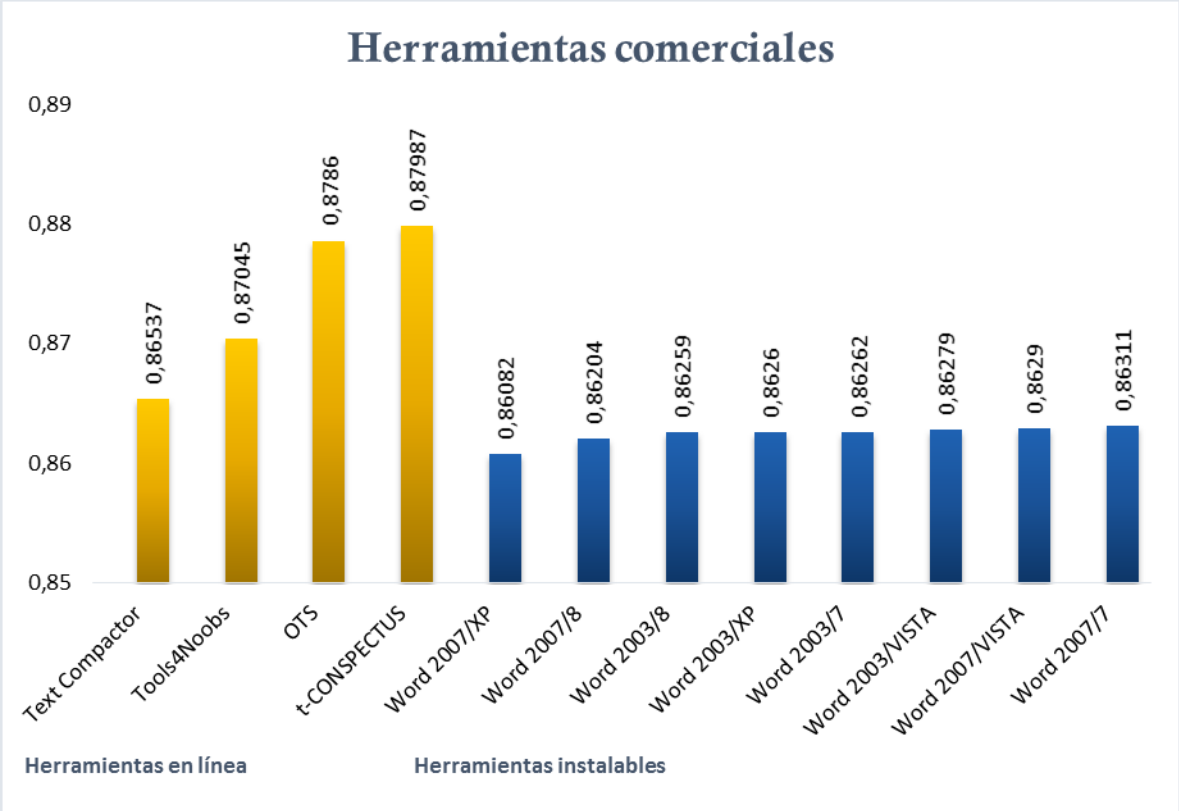


Figura 21. Herramientas comerciales en línea e instalables.

Donde se puede ver que la herramienta comercial que realiza los resúmenes con menor calidad es la herramienta instalable con el sistema operativo *XP* y paquetería de *Microsoft Word 2007*.

Y la herramienta con mayor calidad es la herramienta *T-Conspectus* con un valor de 0.87987.

Para ver a detalle los resultados de las herramientas comerciales, evaluadas con el sistema ROUGE ir al Anexo 3.

5.4.4 Resultados con los métodos del estado del arte

A continuación se describe la evaluación de los resúmenes generados mediante el método [Matias 16] realizados con la herramienta ROUGE.

Recordando que *Translit* forma parte del proceso de la generación de resúmenes del estado del arte, la tarea de *Translit* es pasar los caracteres sirilicos a letras en latín, no es un traductor.

El *baseline* del corpus *TEXTRUSS* es de 0.89993 se menciona con la finalidad de ver la calidad de los resúmenes contenidos en el corpus *TEXTRUSS,* el *baseline* de T*ranslit* no supera el *baseline* de *TEXTRUSS,* por consecuente, los archivos devueltos por el método [Matias 16] se tienen que regresar a caracteres cirílicos, para realizar la evaluación equilibrada entre los resúmenes *gold standart* de *TEXTRUSS* y los resúmenes generados por el método [Matias 16].

En seguida se describen los parámetros por los cuales fueron realizados los diferentes modelos de texto (Ver Tabla 1).

Tabla 1. Parámetros para el método de [Matias 16] aplicados al corpus *TEXTRUSS*.

Modelo de texto	Pre procesamiento	Pendiente	Operador de selección	Función de aptitud	F-measure ROUGE-1
Bolsa de palabras	No	0.3	Ruleta	Belta = 0.5 y Delta = 0.5	0.88969
Bolsa de palabras	No	0.8	Ruleta	Belta = 0.5 y Delta = 0.5	0.87966
Bolsa de palabras	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.89051
Bigramas	No	0.3	Ruleta	Belta = 0.5 y Delta = 0.5	0.88671
Bigramas	No	0.8	Ruleta	Belta = 0.5 y Delta = 0.5	0. 89113
Bigramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.88898
Bigramas	No	0.75	Ruleta	Belta = 0.5 y Delta = 0.5	0.88563
Trigramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.88828
Cuatro gramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.89487
Cinco gramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.89459
Cinco gramas	No	0.75	Ruleta	Belta = 0.5 y Delta = 0.5	0.89065
Seis gramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.89392
Siete gramas	No	0.625	Ruleta	Belta = 0.5 y Delta = 0.5	0.89497
Bolsa de palabras	No	0.90	Ruleta	Belta = 0.4 y Delta = 0.6	0.89128
Bolsa de palabras	No	0.90	Ruleta	Belta = 0.6 y Delta = 0.4	0.89011
Bolsa de palabras	No	0.90	Ruleta	Belta = 0.3 y Delta = 0.7	0.88978
Bolsa de palabras	No	0.90	Ruleta	Belta = 0.5 y Delta = 0.5	0.88963

En seguida se muestran las evaluaciones gráficamente, para ver con claridad los resultados (Ver Figura 22).

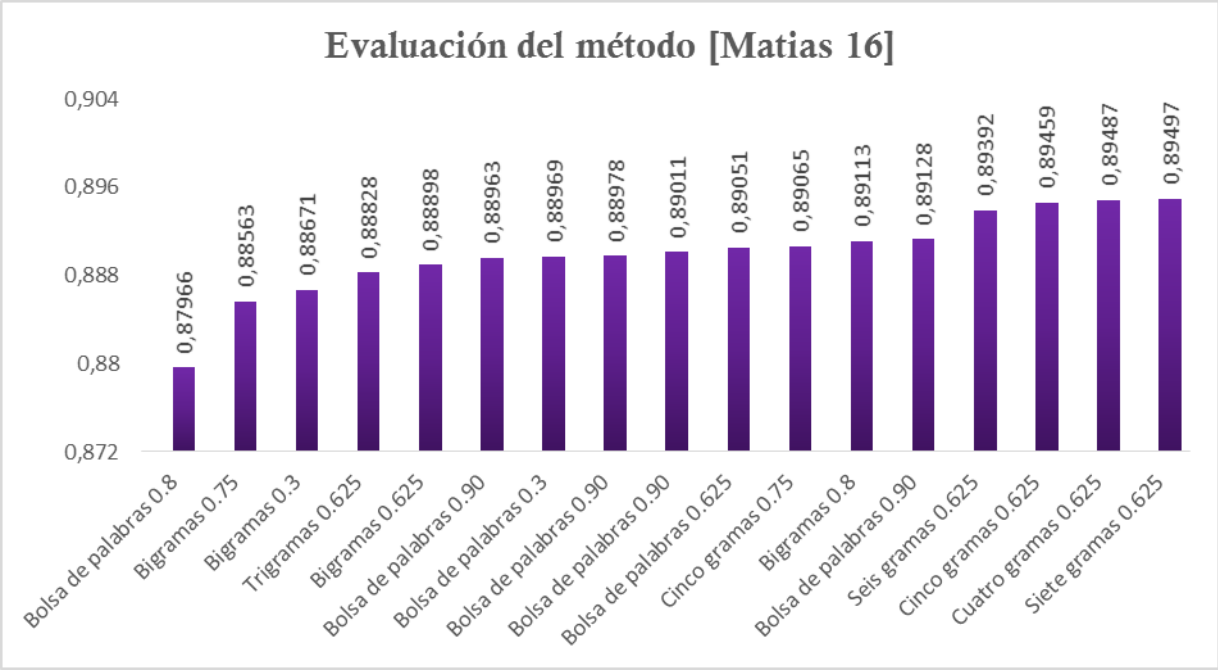


Figura 22. Evaluación de pruebas con el método de [Mastias 16].

Como se puede ver en la gráfica (Ver Fgura 22) el parámetro con menor calidad para la generación automatica de resúmenes es el de bolsa de palabras, mientras que el valor más alto es el de siete gramas con un valor de 0.89497, cabe destacar que este valor se tomara como referencia para la comparación de las herramientas comerciales. Para ver a detalle los parámetros de evaluación dirigirse al Anexo 4.

5.4.5 Comparación de las herramientas comerciales y los métodos del estado del arte

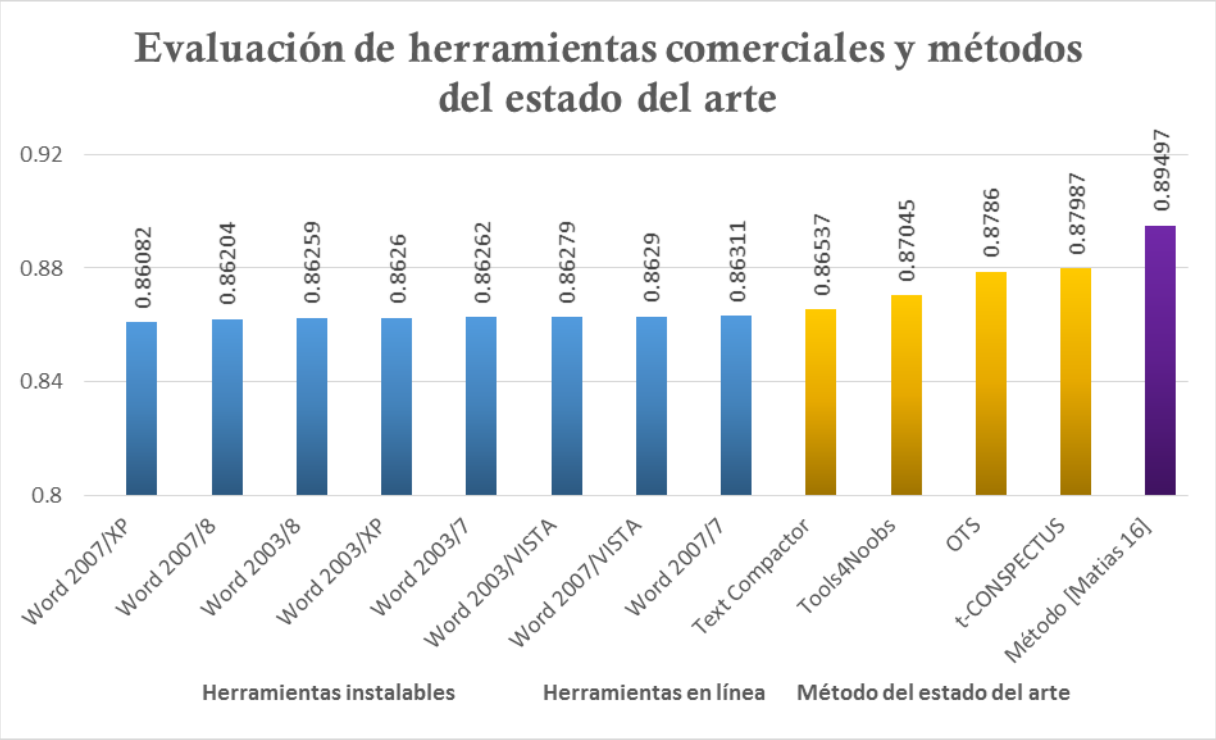


Figura 23. Comparación de las herramientas comerciales evaluadas con ROUGE.

Como se puede observar (Ver Figura 23) en la comparación de las herramientas comerciales y los métodos del estado del arte, el mejor resultado para la elaboración de resúmenes de mayor calidad es el método propuesto por [Matias 16].

A continuación la lista de herramientas comerciales por su calidad (menor-mayor):

Sistema operativo *Windows XP / Microsoft Word 2007*
Sistema operativo *Windows 8 / Microsoft Word 2007*
Sistema operativo *Windows 8 / Microsoft Word 2003*
Sistema operativo *Windows XP / Microsoft Word 2003*
Sistema operativo *Windows 7 / Microsoft Word 2003*

Sistema operativo *Windows VISTA / Microsoft Word 2003*

Sistema operativo *Windows VISTA / Microsoft Word 2007*

Sistema operativo *Windows 7 / Microsoft Word 2007*

Text Compactor [TextCompactor 15]

Tools4noobs [T4NS 15]

Open Text Summarizer [OTS 15]

T-Conspectus [Conspectus 15]

Para ver detallados las medidas de evaluación de estos resultados ir al Anexo 2.

5.4.6 Herramientas comerciales y métodos del estado del arte

En la Figura 24, se puede observar la comparacion de la evaluacion de las herramientas comerciales con el método del estado del arte [Matias 16] utilizado en este trabajo.

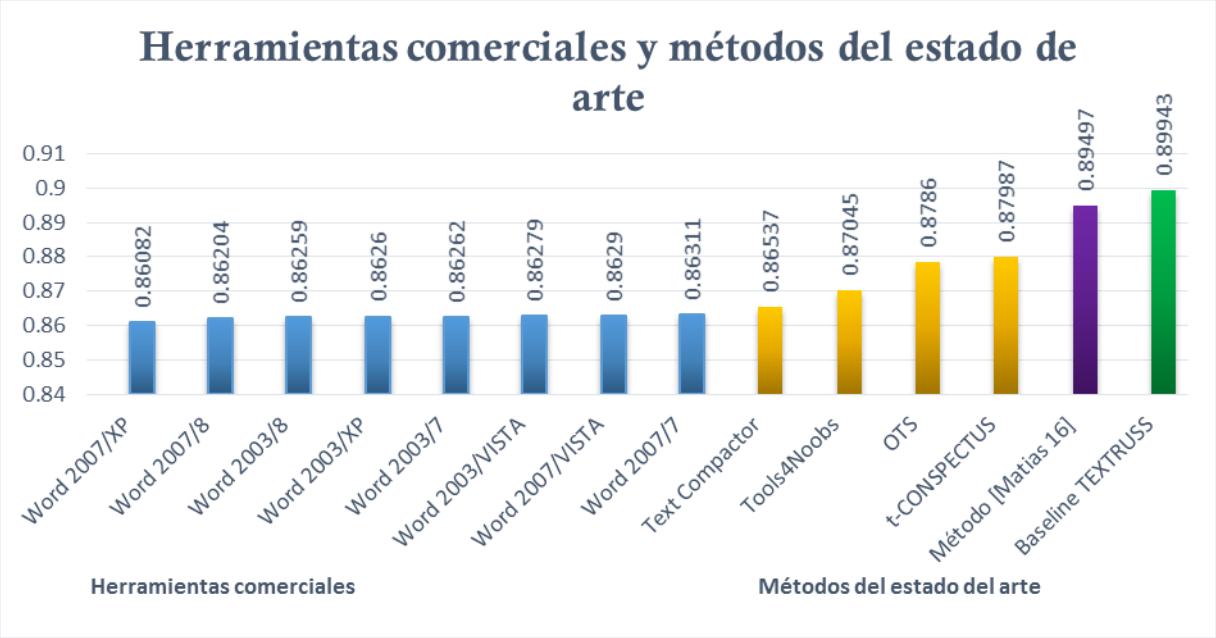


Figura 24. Comparación de las herramientas comerciales y los métodos del estado del arte.

Como se puede apreciar, el método de [Matias 16] muestra mayor calidad en la elaboración de resúmenes que las herramientas comerciales.

Se comprobó que el método propuesto por [Matias 16] aplicado para el idioma ruso tiene una mayor consistencia en comparación de las herramientas comerciales, para la generación automática de resúmenes.

CAPÍTULO 6

Conclusiones y Trabajo Futuro

En este capítulo, se presentan las conclusiones generales de la tesis, así como el trabajo futuro. Se mencionan algunas de las direcciones posibles de investigación a partir de esta tesis.

6.1 Conclusiones

- ✓ Se descubrió que para la elaboración de un corpus se necesita contar con el apoyo de un experto en el idioma.

- ✓ Con la elaboración de este trabajo de tesis se descubrió que los métodos del estado del arte son mejores que las herramientas comerciales para el idioma ruso.

- ✓ De las herramientas comerciales en línea, para realizar resúmenes en idioma ruso, la mejor herramienta es *T-Conspectus*.

- ✓ De las herramientas comerciales instalables la más óptima para realizar resúmenes en el idioma ruso es la que utiliza el sistema Operativo *Windows 7 Microsoft Word 2007*.

- ✓ Se probaron diferentes parámetros del método de [Matias 16] para el idioma ruso.

- ✓ Se comprobó que, con el método [Matias 16] se obtienen los resúmenes extractivos con mayor calidad.

- ✓ En la mayoría de los trabajos del estado del arte la mejor herramienta comercial instalable para realizar resúmenes extractivos fue la herramienta *Copernic*. En este trabajo, no se implementó ya que solo realiza resúmenes en idioma inglés, español, alemán y francés.

- ✓ El *baseline* del corpus *TEXTRUSS* es muy alto por la estructura del propio corpus que se utiliza para redactar las noticias.

6.2 Trabajo futuro

- Probar el corpus TEXTRUSS con otros métodos del estado de arte.
- Comprobar si los resúmenes elaborados con las herramientas comerciales instalables, pueden realizarse mediante otros sistemas operativos (Linux, Ubuntu, Mac OS).
- Elaboración del módulo de transliteración.
- Elaboración el módulo de SFMs enfocados al idioma ruso.
- Realizar las pruebas de los métodos del estado del arte con pre-procesamiento.

Referencias

[Alfonseca 03] Alfonseca Enrique y Pilar Rodríguez, Generating extracts with genetic algorithms, publicación periódica Springer Verlag Berlin s.n Vol. Volumen 2633. Págs. 511-519. -3-540-01274-5, 2003.

[Alguliev 05] Alguliev Rasim M., Ramiz M. Aliguliyev; Effective Summarization Method of Text Documents; Institute of Information Tecnology, Azerbaijan National Academy of Sciences, Baku Azerbaijan, 2005.

[Armeaga 15] Armeaga Geovani García, Tesis de Licenciatura en Ingeniero en Software, "Comparación de medidas de similitud en cadenas textuales, para la detección de plagio en tareas escolares", Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, 2015.

[Berker 11] Berker Mine, Using genetic algorithms with lexical chains for automatic text summarization, Presentado en la Universidad Bogazoci, para obtener el grado de Ingeniero en computación, año 2011.

[Camacho 15] Camacho Marcela Ávila, "Detección de fragmentos de texto como candidato a hipervínculo", Tesis para obtener grado de Maestra en Ciencias de la Computación; Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, Enero 2015.

[Cunha 08] Cunha Fanego y Iria da, "Hacia un modelo lingüístico de resumen automático de artículos médicos en español", Barcelona, Presentada en el Instituto Universitario de Lingüística Aplicada, Tesis para la

obtención del grado de Doctorado en Ciencias del lenguaje y Lingüística Aplicada, 2008.

[Conspectus 15] *T-Conspectus* Herramienta comercial en línea para realizar resúmenes automáticos, fecha de consulta 16 de Diciembre de 2015, http://tconspectus.pythonanywhere.com/summarization.

[Copernic 15] Copernic, Herramienta comercial instalable para realizar resúmenes automáticos, fecha de consulta 12 de Octubre de 2015, http://www.splitbrain.org/services/ots.

[Díaz 05] Díaz, A. y P. Gervás, "Personalisation in news delivery systems: item summarization and multi-tier item selection using relevance feedback". Web Intelligence and Agent Systems, 3(2):135-154, 2005.

[Fayyad 96] Fayyad, U.M.,G. Piatesky Shapiro y P. Smyth, "From data mining to knowledge discovery: an overview", Advances in Knowledge Discovery and Data Mining AAAI/MIT Press., pp 1-34 1996.

[Gantz 14] Gantz Jhon F., Vernon Turner, David Reinsel, Stephen Minton The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, "April 2014, sponsored by EMC. The multimedia content can be viewed at http://www.emc.com/leadership/digital-universe/index.htm. http://idcdocserv.com/1678.

[García 08] García Hernández René Arnulfo, Ledeneva Yulia, Alexander Gelbukh, Erendira Rendon, Rafael Cruz, Text Summarization by Sentences Extraction Using Unsupervised Learning. LNAI 5317, pp133-143, Springer-Verlag, ISSN 0302-9743, 2008.

[García 09] García Hernández René Arnulfo, Ledeneva Yulia, Rafael Cruz Reyes, Romyna Montiel Soto, Comparación de Tres Modelos de

Representación de Texto en la Generación Automática de Resúmenes, Sociedad Española para el Procesamiento de Lenguaje Natural, vol.43, pp. 303-311, ISSN 1135-5948, 2009.

[García 09a] García Hernández René Arnulfo, Yulia Ledeneva, Griselda Matias, Ángel Hernández Dominguez, Jorge Chavez, Alexander Gelbukh, "Comparing Commercial Tools and state-of-the-art methods for generating Text Summaries", IEEE Computer Society Press, pp. 92-96,ISBN 9780769539331, Noviembre, 2009.

[Gazeta.ru 15] Gazeta.ru, Portal de noticias donde se descargaran los artículos que conforman al corpus *TEXTRUSS*, fecha de consulta: lunes 10 de agosto de 2015, http://www.gazeta.ru/

[Gelbukh 06] Gelbukh Alexander y Grigori Sidorov, Ciencias de la Computación. Procesamiento automático del español con enfoque en recursos léxicos grandes, primera edición, impreso en México, Centro de Investigación en Computación, Instituto Politécnico Nacional, 2006.

[Gelbukh 10] Gelbukh Alexander, Artículo de divulgación Procesamiento de Lenguaje Natural y sus Aplicaciones, Sociedad Mexicana de Inteligencia Artificial, Komputer Sapiens ISSN 2007-0691, Año II vol. I. Enero - Junio 2010.

[Márquez 10] Márquez Víctor Gil, Resúmenes automáticos: Enfoque extractivo y evaluación, Escuela Politécnica Superior Universidad Autónoma de Madrid, publicación periódica 23 de mayo de 2010.

[Hernández 16] Hernández Yanet Casimiro, "Extracción de frases clave usando patrones léxicos en artículos científicos", Tesis para obtener grado de Maestra en Ciencias de la Computación; Universidad Autónoma del

Estado de México Unidad Académica Profesional Tianguistenco, Enero 2016.

[Hovy 99a] Hovy E. and C. –Y. Lin Advances in Automatic Text Summarization, chapter Automated text summarization in SUMMARIST, pages 81-94. MIT Press, Cambridge, 1999.

[Hovy 99b] Hovy E. and C. Lin, Automated Text Summarization in SUMMARIST. In 'Advances in Automatic Text Summarization', I. Mani and M. Maybury (editors), 1999.

[Ibáñez 13] Ibáñez Onofre Dulce Yarely, "Evaluación de las herramientas comerciales de generación automática de resúmenes de textos para el idioma Portugués", Tesis de Licenciatura, Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, 2013.

[Ledeneva 11] Ledeneva Yulia Nikolaevna, René García Hernández, Griselda Matias Medoza, Selene Vargas, Abraham García, Comparison of State-of-the-Art Methods and Commercial Tools for Multi-Document Text Summarization, Research in Computer Science. ISSN: 1870-4069, vol.54, pp. 145-159, 2011. (INDIZADO POR LATINEX)

[Ledeneva 08a] Ledeneva Yulia Nikolaevna, Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization, tesis de doctorado Instituto Politécnico Nacional, 2008.

[Ledeneva 08b] Ledeneva Yulia, Alexander Gelbukh, Rene Arnulfo Garcia- Hernandez; Terms Derived from Frequent Sequences for Extractive Text Summarization; Natural Language and Text Processing Laboratory; Center for Computing Research; National Polytechnic Institute; DF 07738;México; 2008.

[Lin 04] Lin Chi-Yew: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proc. of Workshop on Text Summarization of ACL, Spain, 2004.

[Lin 97] Lin, C.Y., Hovy, E.: Automated Text Summarization is SUMMARIST.In: Proc. of ACL Workshop on Intelligent, Scalable Text Summarization, Madrid, Spain, 1997.

[Lloret 08] Lloret E., O. Ferrández, R. Muñoz y M. Palomar. "Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos". Procesamiento del Lenguaje Natural, No. 41, 2008, pp. 183-190, 2008.

[Lloret 12] Lloret E., Palomar M., Text summarization in progress: A literature review. Artificial Intelligence Review, 37(1), 1-41, 2012.

[Lyman 16] Lyman, Peter and Hal R. Varian, How Much Information. Retrieved from http://groups.ischool.berkeley.edu/archive/how-much-info-2003/execsum.html2003, consultado el 9 de Marzo del 2016.

[Mani 01] Mani Inderjeet, I. Jonh Benjamins Publishing Company, Automatic Summarization, Natural Language Processing 3, Company 286 pp., 2001.

[Maña 03] Maña Manuel J. López; Generación automática de resúmenes de Texto para el acceso a la información; Tesis de Doctorado; Escuela Superior de Ingeniería Informática; Universidad de Vigo en Ourense; España; 2003.

[Maqueo 98] Maqueo Ana María y Verónica Méndez, Libro del maestro Español 2 Lengua y Comunicación, Editorial Limusa Noriega, México, Abril 1998.

[Matias 13] Matias Mendoza Griselda Areli, "Generación automática de resúmenes usando algoritmos genéticos", Tesis de Licenciatura; Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, 2013.

[Matias 16] Matias Mendoza Griselda Areli, "Generación automática de resúmenes independientes del lenguaje", Tesis para obtener grado de Maestra en Ciencias de la Computación; Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, Enero 2016.

[Mihalcea 04] Mihalcea Rada and Paul Tarau, TextRank:Bringing Order into Texts, Departamento of Computer Science University of North Texas, 2004.

[Montiel 09] Montiel Romyna Soto, "Generación automática de resúmenes mediante aprendizaje no supervisado", Tesis de Licenciatura; Instituto Tecnológico de Toluca, 2009.

[Montiel 09a] Montiel Romyna Soto, René GarcÍa Hernández, Yulia Ledeneva, Rafael Cruz Reyes, Comparación de Tres Modelos de Representación de Texto en la Generación Automática de Resúmenes, Sociedad Española para el Procesamiento del Lenguaje Natural, vol. 43, pp. 303-311, ISSN 1135-5948, 2009.

[MOW 15] Microsoft Office 2003/2007 descarga de sistemas operativos, fecha de consulta 29 de Noviembre de 2015, de la página principal de Microsoft Office http://www.microsoft.com/es-mx/download

[Nea 02] Nea Leavitt; Data Mining fot the Corporate Masses, IEEE Computer Society Press, U.A 2002.

[Nenkova 11] Nenkova, A. & McKeown, K., Automatic summarization, Foundations and Trends in Information Retrieval, 5(2-3), 103-233, fecha de consulta

12 de Octubre de 2015 http://wwwleavcom.com/pdf/Dataminingstory.pdf

[NewsBlaster 16] NewsBlaster Natural Language, consultado el 03 de Marzo de 2016, Proceessing, http://www1.cs.columbia.edu/nlp/projects.cgi/

[OTS 15] Open Text Summarizer, Herramienta comercial en línea para realizar resúmenes automáticos, fecha de consulta 12 de Octubre de 2015, http://www.splitbrain.org/services/ots.

[Plaza 10] Plaza Laura Morales, Uso de Grafos semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación de Distintos Dominios: Biomedicina, Periodismo y Turismo, Tesis Doctoral, Facultad de informática Universidad Complutense de Madrid, diciembre 2010.

[Saggion 10] Saggion Horacio, E. Lloret, M. Palomar, Using Text Summaries for Predicting Rating Scales, Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Held in conjunction to ECAI, 44-51 pp., 2010.

[Saunmali 11] Saunmali Ladda, Naomie Salim y Mohammed Salem Binwahlan, Genetic algorithm based sentence estraction for text summarization, publicación periódica International Journal of INNOVATIVE computing pág. Vol 1. -2180-4370, 2011.

[Sparck 99a] Sparck Karen Jones, Automatic Summarizing: Factors and Directions, Computer Laboratory, University of Cambridgem, Ed. I. Mani and M. Maybury, 1-12, MIT Press 1999.

[Sparck 99b] Sparck Karen Jones and J. Galliers, Evaluating natural language processing system: An analysis and review, Article in ARTIFICIIAL INTELLIGENCE, Junuary 1999.

[StarWars 16] StarWars película de ciencia y ficción, Consultada el 13 de enero 2016 http://www.StarWars.com

[Torres 05] Torres Georgina Araceli Vargas, Libro: Biblioteca digital, 69 p., Centro Universitario de Investigación Bibliotecológicas, México: UNAM ISBM 970-32-2471-7, 2005.

[Translit 15] Translit.ru, Herramienta en línea para transliterar los artículos del corpus *TEXTRUSS*, fecha de consulta: 5 de Octubre de 2015, http://translit.net/.

[T4NS 15] Tools4noobs Summarizer herramienta comercial en línea para realizar resúmenes automáticos Online summarizer tool (*free summarizing*), fecha de consulta 18 de Octubre de 2015, http://www.tools4noobs.com/summarize/.

[TextCompactor 15] Text Compactor. Herramienta comercial en línea para realizar resúmenes automáticos en línea, fecha de consulta 18 de Octubre de 2015, http://textcompactor.com/.

[Vargas 16] Vargas Flores Selene Itzel, "Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos", Tesis para obtener grado de Maestra en Ciencias de la Computación; Universidad Autónoma del Estado de México Unidad Académica Profesional Tianguistenco, Enero 2016.

[Vázquez 15] 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 06] Villatoro Tello Esaú, "Generación automática de resúmenes de múltiples documentos", Tesis de maestría; Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, 2006.

Anexo 1.

Descripción de las herramientas comerciales

Open Text Summarizer [OTS 15]

Esta herramienta comercial en línea analiza automáticamente los textos y trata de identificar las partes más importantes del texto. Es una herramienta de código abierto para resumir textos. El programa lee un texto y decide qué frases son importantes y cuáles no lo son. *OTS* soporta varios idiomas: inglés, alemán, español, ruso, hebreo y otros (+25). Para apoyar más idiomas o modificar los lenguajes existentes se pueden hacer simplemente editando un archivo XML de reglas. *OTS* es a la vez una biblioteca y una herramienta de línea de comandos que permiten resumir el texto en la consola.

A continuación se describe cada una de las opciones que permite la herramienta Open Text Summarizer (Ver Figura A1):

1. "*input*": Es el área para colocar el texto a resumir.
2. "*or load from URL*": Es el área para colocar el URL de la página web a resumir.
3. "output": Es el área donde el usuario puede elegir el tipo de salida de la información.
   3.1 "output summary": El tipo de información de salida será de resumen general.
   3.2 "output keywords": El tipo de información de salida será de palabras claves.
4. "summarization radio": Aquí se elige el umbral, cabe destacar que mediante esta opción podemos elegir el tamaño de salida de la información.
5. "lenguaje": En esta área se elegirá el idioma que realizará el resumen.
6. "enviar": La función de este botón, es para dar inicio al proceso de la realización del resumen.

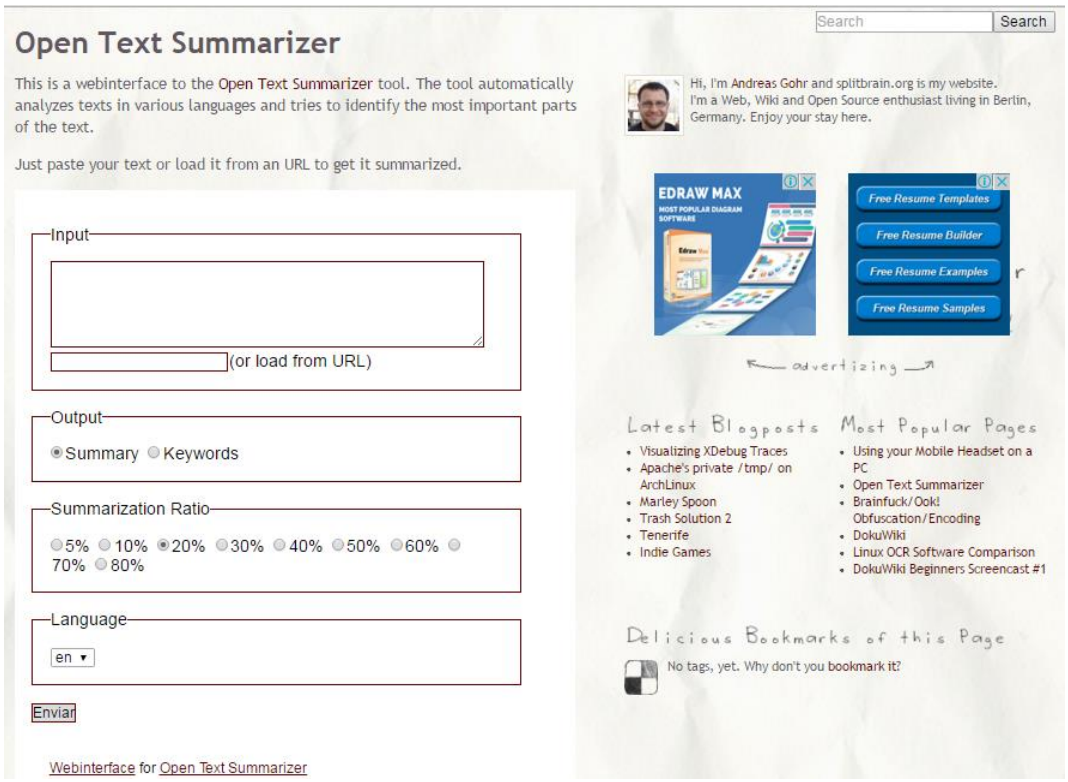


Figura A1. Interfaz de la herramienta comercial en línea Open Text Summarizer [OTS 15].

Text Compactor [TextCompactor 15]

Herramienta de generación de resúmenes gratuita, disponible en línea. Fue creada para ayudar a estudiantes, maestros o profesiones con dificultades de procesar cantidades abrumadoras de información.

Después de que el texto es escrito o pegado en el cuadro de diálogo, la aplicación web calcula la frecuencia de cada palabra en el pasaje. Entonces, una puntuación se calcula para cada frase basándose en el cálculo de la frecuencia asociada con las palabras que contiene. La frase más importante se considera que es la frase con el cálculo de la frecuencia más alta. Los resultados pueden ser manipulados cuando un pasaje tiene sólo

unas pocas frases. La herramienta *Text Compactor* no está lista para el uso de lecturas de ciencia ficción (es decir, las historias sobre personajes imaginarias, lugares, eventos).

La herramienta *Text Compactor* nos da tres pasos para realizar la generación de resúmenes. A continuación se describe cada una de las opciones que permite que la herramienta *Text Compactor* (Ver Figura A2):

a) Paso 1. "Type or paste your tex tinto the box": En este recuadro se escribe o pega el texto a resumir.
b) Paso 2. "Drag the slider, or enter a number in the box, to set the porcentage of text to keep in the summary": Aquí se define el umbral del texto a resumir. El umbral va desde 0 a 100 (el aumento va en unidades).
c) Paso 3. "Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text' into a word processor, or text to speech program or language translation tool": En este último paso, nos da el resumen del texto previamente dado. También nos hace referencia al análisis del resumen resultante, si es que no cumple con los requerimientos del usuario, se pueden volver a realizar los pasos desde la ejecución del paso 2.

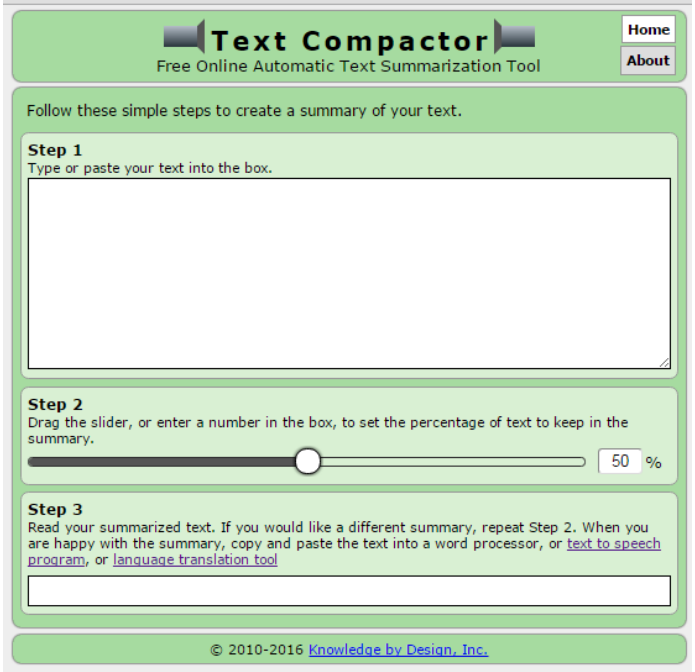


Figura A2. Interfaz de la herramienta comercial en línea Text Compactor [TextCompactor 15].

Tools4noobs [T4NS 15]

Esta herramienta en línea crea automáticamente un resumen de un texto (por lo general de grandes tamaños del texto); ya sea que pegue el texto o pegue el link de la dirección de la página web a resumir y la herramienta da un breve resumen (Ver Figura A3).

La herramienta consta de 3 frases:

a) Extracción de frases del texto dado.

b) Identificación de las palabras clave en el texto y contar relevancia de cada una de ellas.

c) Identificación de las frases con la mayoría de las palabras clave y de salida correspondientes ellas en base a las opciones seleccionadas.

Contiene parámetros de los cuales el usuario puede manejar de acuerdo a sus requerimientos, a continuación se describen:

1. "*URL*": Aquí se pone la dirección electrónica del archivo a resumir.

2. "*OR direct input*": Aquí se escribe o se pega el texto a resumir.
3. "*Hide options*": Pestaña para ocultar o mostrar los parámetros a modificar.
4. "*Threshold*": Aquí se define el umbral, es decir, el valor que se establece limita las frases en función de su relevancia, la cual a su vez determinan el número de palabras relevantes. Las oraciones que selecciona del documento original, las coloca en el mismo órden en el texto resumido.
5. "*Number of lines*": En esta opción se puede definir el número de líneas del cual constará el resumen final dado por la herramienta. Cabe mencionar que si se elige esta opción, la opción de arriba se deshabilita automáticamente.
6. "*Mininum sentence length*": Aquí se puede definir el número de caracteres de una sentencia, por defecto la herramienta marca 50.
7. "*Mininum Word length*": Aquí se puede definir el número mínimo de caracteres de una palabra, por defecto la herramienta marca 4.
8. "*Show sentence relevance*": Por medio de número se calcula la adición de la relevancia de cada palabra clave detectada. La relevancia de cada frase se muestra al final, dentro de un paréntesis.
9. "*Show best words*": Esta opción nos remarca en el texto las palabras claves más relevantes en el texto de resumen.
10. "*Number of best words*": Esta opción nos da el número de palabras claves (esta opción solo se habilita si la opción anterior se habilita).
11. "*Keyword highlighting*": Si se selecciona esta opción se generara automáticamente una salida adicional con el mismo resumen, pero resaltando las palabras claves. Se sugiere seleccionar esta opción si se desea realizar una verificación rápida de la densidad de palabra clave en el texto final.
12. "*Show sentences*": Muestra las frases más destacadas en el texto.
13. "*Summarize it!*": Botón para dar comienzo a la generación automática del resumen.

Figura A3. Interfaz de la herramienta comercial en línea Tools4noobs [T4NS 15].

T-Conspectus [Conspectus 15]

Es una aplicación web para resumir los artículos en inglés, alemán y ruso. Proporciona servicio multilingüe resumen automático de artículos de noticias.

El generador de resúmenes utiliza algunas técnicas de *PLN* para extraer automáticamente las frases más informativas a partir de un texto sin formato insertado en el cuadro de texto, cargado por el usuario o insertado desde una *URL*.

Utiliza un algoritmo para su procesamiento, el cual contempla un proceso de tres fases:

- La primer etapa es el pre-procesamiento la cual realiza cuatro procedimientos principales:

  -Título: Si el a ser resumido contiene título este será utilizado para la asignación de pesos adicionales a las palabras claves (es recomendable introducir textos con título).

  -Divide el texto en párrafos: El generador de resúmenes necesita saber los límites del párrafo para encontrar su primer y última frase y poner puntuaciones basadas en la posición.

  -Divide los párrafos en oraciones: Este proceso se divide en dos sub etapas. La primera la descomposición inicial de cadena, la corrección posterior a la división.

  -Tokenización en cada oración: El módulo se divide en palabras frases, haciendo coincidir una cadena contra el patrón de expresión.
- La segunda etapa es la de puntuación de resúmenes asigna ponderaciones a términos construyendo así dinámicamente un diccionario de palabras clave (ponderación plazo y ponderación frase).
- La tercera etapa es la generación del resumen, selecciona un número "n" de las primeras frases de la lista generada en el paso anterior. El número de sentencias que se seleccionara en el resumen final se calcula en función del usuario. [Conspectus 15].

Esta herramienta proporciona cuatro opciones diferentes:

1) Generación de resumen automático
2) Segmentación de texto
3) Ponderación de términos
4) Comparación de resúmenes

1-. Generación de resumen automático: Aquí se generan resúmenes automáticos a partir de un texto dado ya sea mediante una URL, escribiendo o pegado el texto.

2-. Segmentación de texto: Aquí nos muestra las oraciones más representativas del texto dado, nos dice el idioma del texto resumido, la frase del artículo que más se repitió, el

número de frases que nos mostrará y mediante una tabla nos muestra las frases más representativas del texto.

3-.Ponderación de término: Aquí se puede probar varios métodos de ponderación de términos. Utiliza una técnica basada en la suposición de que las oraciones que contienen palabras que ocurren con frecuencia en un texto tienen mayor peso que el resto. Esto significa que son las frases más importantes las cuales se van a extraer. La importancia de una palabra se calculó utilizando medidas estadísticas (frecuencia de términos y *TF-IDF*), antes de realizar el cálculo de los pesos de términos se realiza una tarea de filtrado (*stopwords* y *lematizacion*), en seguida se aplica el proceso del algoritmo de resumen basado en término. Después de este proceso se muestra el resumen resultante.

4-. Comparación de resúmenes: Aquí compara los resúmenes generados de esta herramienta con otras herramientas comerciales. Además de contener los mismos parámetros de la generación automática de resúmenes, se añade la opción de elegir mediante que algoritmo se desea realizar la comparación ya sea mediante *TextRank, LexRank, Lsa, Edmundson, Luhn* o Aleatorio. Una vez eligiendo el algoritmo se genera el texto resumido.

Una vez explicado el contenido general de la herramienta, se analiza a detalle los parámetros de "generación de resumen automático" ya que estos se aplican en esta tesis (Ver Figura A4):

1. "*Type or copy the text of an article here and click Summarize*": En este recuadro se pega o se escribe el texto a resumir.
2. "*O paste URL*": Aquí el usuario puede pegar la dirección electrónica del archivo a resumir.
3. "*Or upload an article*": Aquí podemos seleccionar el archivo desde nuestra computadora el cual será resumido. Solo se pueden cargar texto sin formato.
4. "*Specify the size of the resulting summary*": Aquí se especifica por medio de valor porcentual el tamaño el texto resultante (va desde el valor 5 hasta el valor 70. El aumento va en una escala de 5 en 5).

5. "*Show keywords and statistics*": Mediante esta opción se mostrarán las palabras claves y estadísticas.
6. "*Keywords in summary*": Mediante esta opción se mostrarán las palabras claves del resumen.
7. "*Summarizer*": Botón para generar el resumen, este solo se activa si hay texto en el recuadro.
8. "*Remove Text*": Elimina el texto o el URL que contenga, este botón solo se activa si los recuadros correspondientes tienen texto.

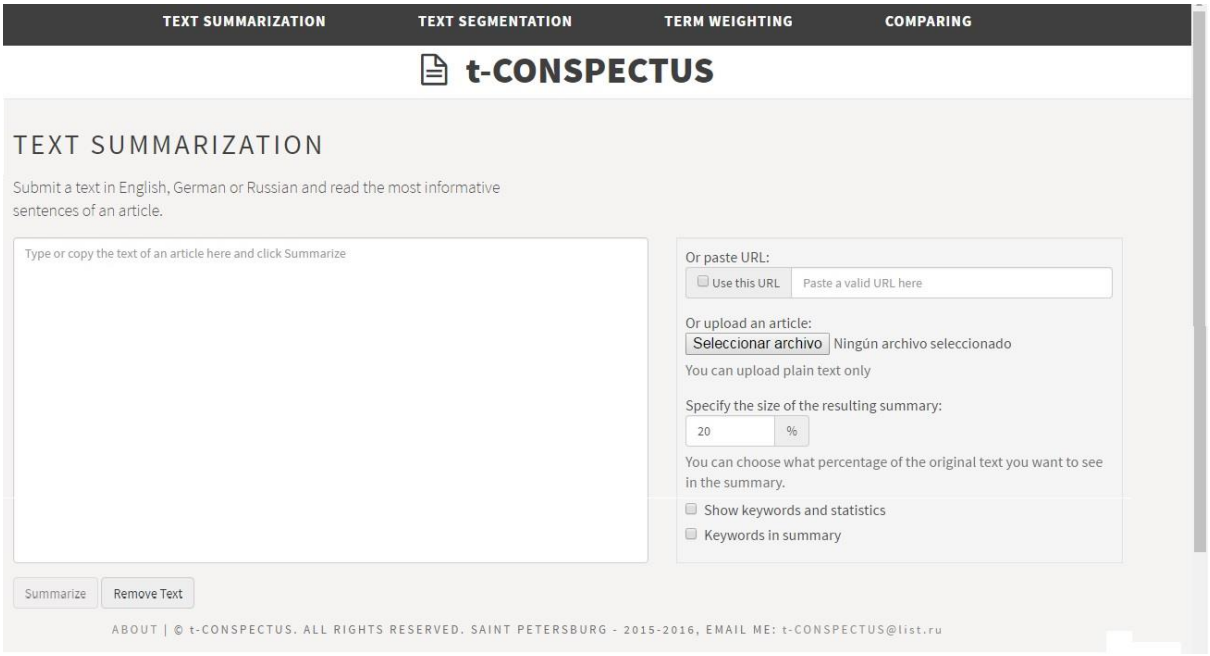


Figura A4. Interfaz de la herramienta comercial en línea *T-Conspectus* [Conspectus 15].

Microsoft Office Word [MOW 15]

En *Microsoft Office Word* únicamente para sus versiones de *Microsoft Office Word* 2003 y *Microsoft Office Word* 2007 contiene la opción para realizar resúmenes automáticos, para ello se tiene que habilitar la opción de "Autorresumen…". A continuación se describirá los pasos para cada versión de *Microsoft Office Word,* primero describiremos los pasos de *Microsoft Office Word* 2003 y en seguida las opciones para *Microsoft Office Word* 2007.

- ***Microsoft Office Word* 2003**

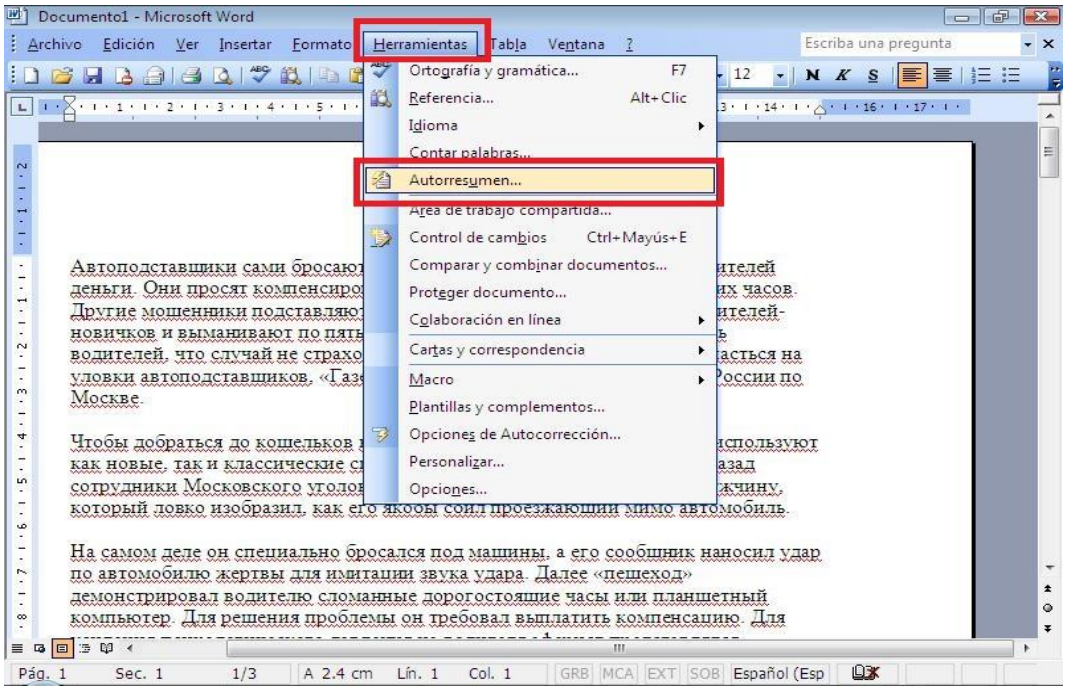
Microsoft Office Word 2003 forma parte del paquete Microsoft Office 2003, formado por un conjunto de programas o aplicaciones con finalidades bien distintas (hojas de cálculo, presentaciones, bases de datos, etc.). Microsoft Office Word 2003 es una versión de procesador de textos más difundido a nivel mundial. Un procesador de textos es un programa que permite crear documentos de texto a los que se le pueden añadir imágenes, gráficos, tablas y un sinfín de objetos que harán más atrayentes los trabajos realizados con él [Villar 05].

Microsoft Office Word 2003 es una herramienta comercial, examina el documento y selecciona las oraciones más relevantes para el tema principal.

En esta tesis se explica los pasos para activar la opción de "Autorresumen…" esta opción nos permitirá generar automáticamente los resúmenes del corpus TEXTRUSS.

Una vez abierta la aplicación de Microsoft Office Word se presenta una ventana con una serie de botones, barras de menús, de desplazamiento, regla etc.

1. El primer paso es colocar el texto a resumir.
2. Damos clic en la pestaña de "Herramientas".
3. Seleccionamos la opción de "Autorresumen…" (Ver Figura A5).



4. Figura A5. Interfaz para activar la opción de autorresumen.

Automáticamente se mostrará un recuadro con los parámetros de Microsoft Word, en los que el usuario podra seleccionar de acuerdo a sus requerimientos. A contiuación se explican:

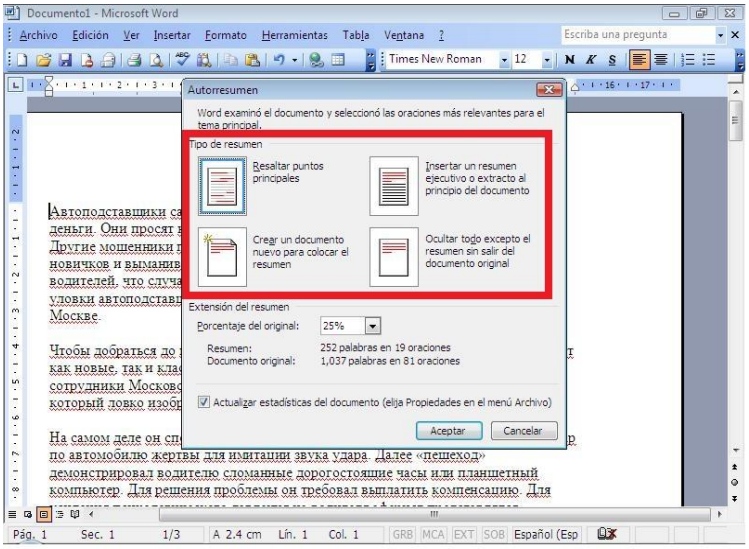


Figura A6. Interfaz para seleccionar parametro de "Tipo de resumen".

Como se puede Ver en la Figura A6, los primeros parametros que nos accede a modificar son el "Tipo de resumen" el cual nos va a permitir visualizar el resumen resultante:

A. Resaltar los puntos principales.
B. Insertar un resumen ejecutivo o extracto al principio del documento.
C. Crear un documento nuevo para colocar el resumen.
D. Ocultar todo excepto el resumen sin salir del documento
d) Para esta tesis se seleccionó la tercera opción "Crear un documento nuevo para colocar el resumen".

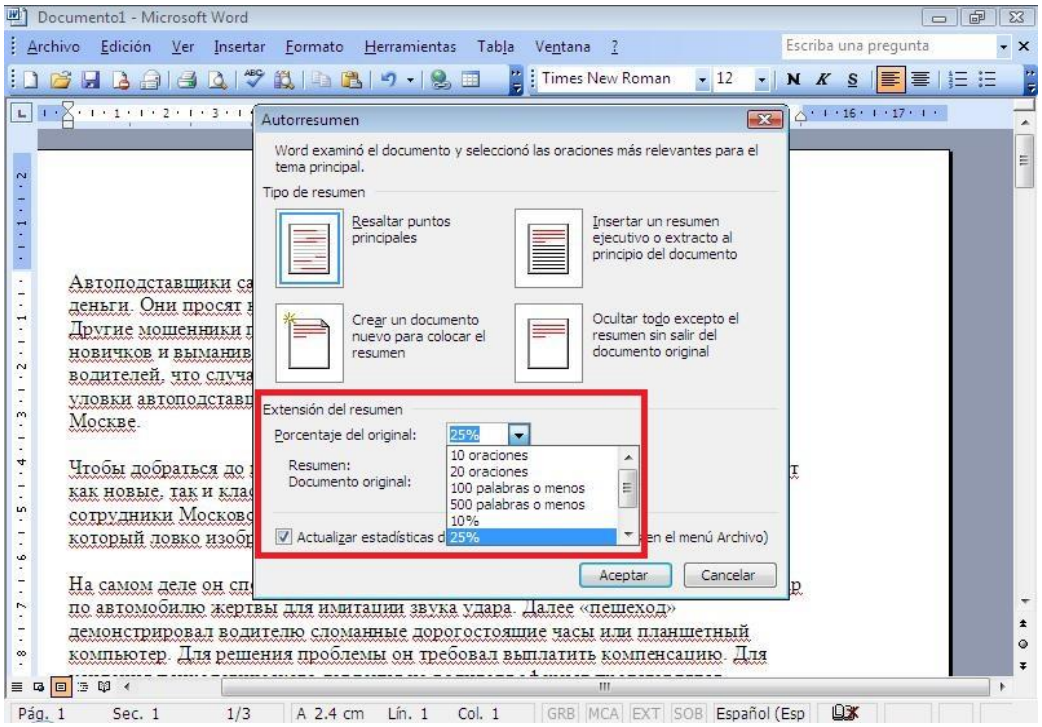


Figura A7. Interfaz para seleccionar parametro de "Extensión de resumen".

El segundo parámetro a modificar es "Extensión del resumen", aquí el usuario seleccionara el requerimiento del tamaño del texto que solicite: por oraciones (10 oraciones o 20 oraciones), número de palabras (100 palabras o menos 500 palabras o menos) o porcentaje (10% 25%). Para esta tesis se eligió la opción "25 %" ya que esta opción genera resúmenes con más de 100 palabras.

e) Por último se selecciona el botón "Aceptar" para generar el resumen automático.

- ***Microsoft Office Word 2007***

Para activar la opción de "Autorresumen…" después de abrir Word 2007:

1. Nos vamos al botón de Office (parte superior izquierda).
2. Seleccionamos el botón "opciones de Word" (Ver Figura A8).

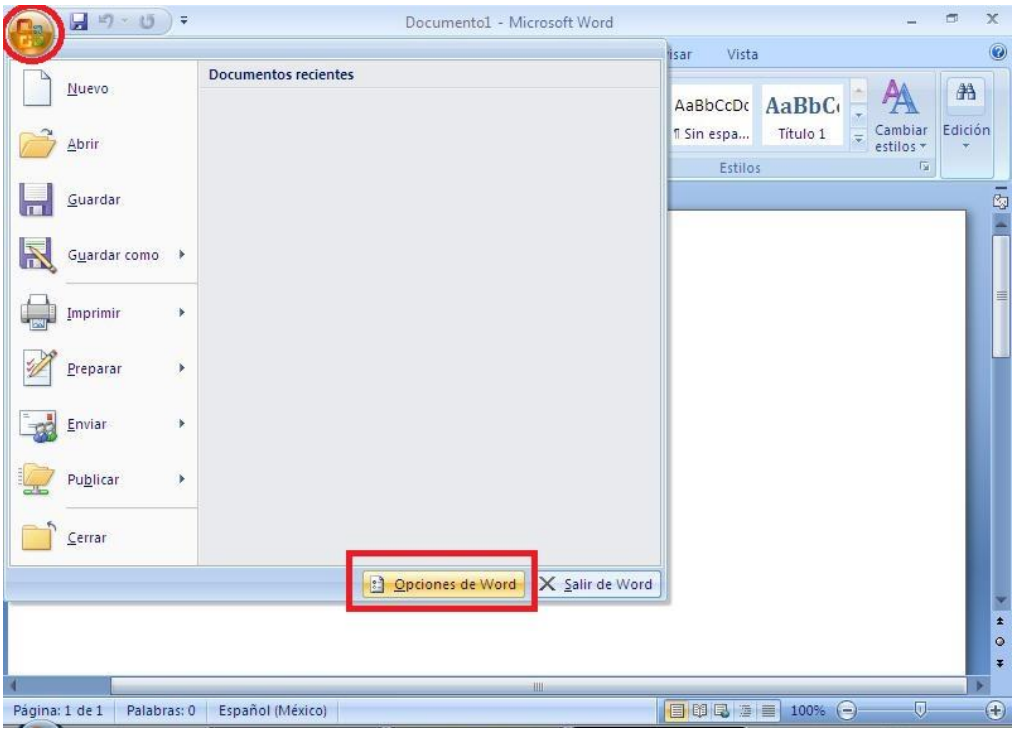


Figura A8. Interfaz para ir a la opción de "autorresumen".

3. Despues seleccionamos la opción"Personalizar" (parte derecha superior-intermedia) (Ver Figura A9).
4. Nos situamos en donde dice "Comandos disponibles en: ".
5. Seleccionamos la opcion de "Todos los comandos".
6. Buscamos y damos clic a la opción "Autorresumen…".
7. Seleccionamos el boton "Agregar > >" .
8. Finalmente damos clik en el boton "Aceptar".
9. En seguida escribimos o pegamos el texto a resumir.

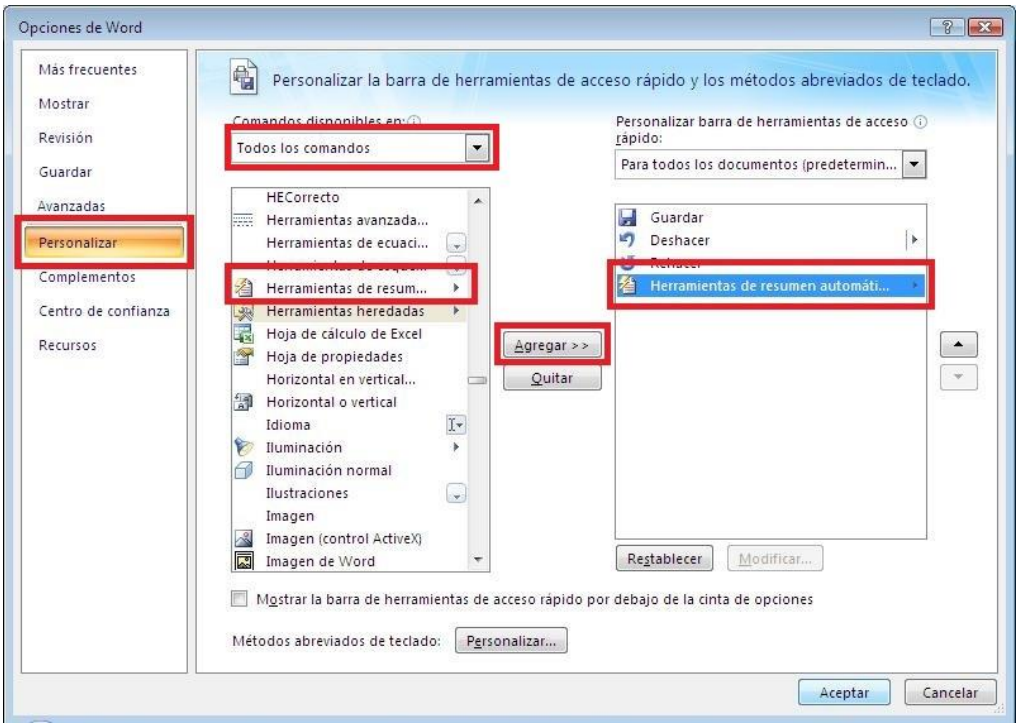


Figura A9. Interfaz para activar opción de "autorresumen".

Automáticamente en la barra de herramientas de acceso rápido nos aparece el icono de "Autorresumen…" (Ver Figura 10).

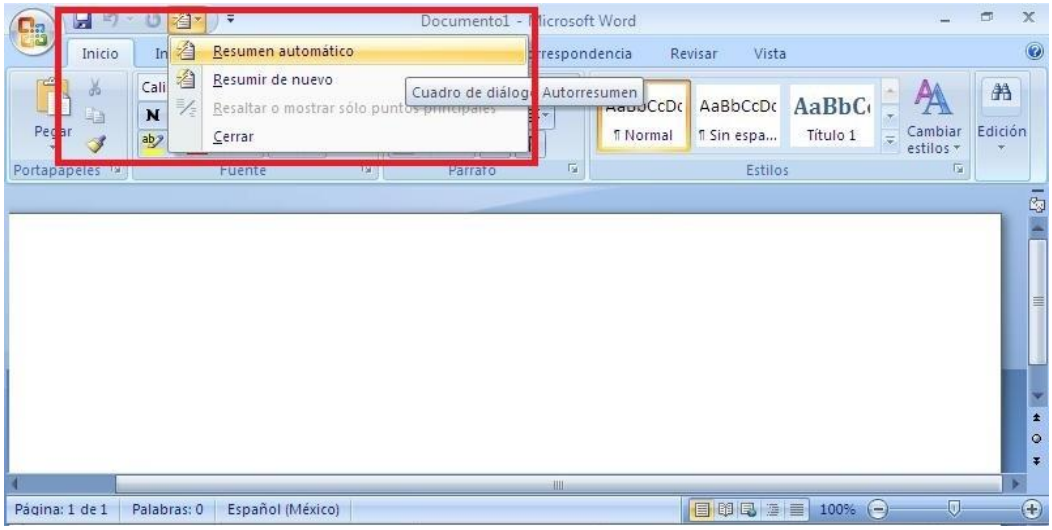


Figura A10. Icono de activación de la opción "autorresumen".

10. Damos clic en la opción de "Resumen automático" y saldra automáticame un cuadro de diálogo con los parámetros a modificar (Ver Figura 11).

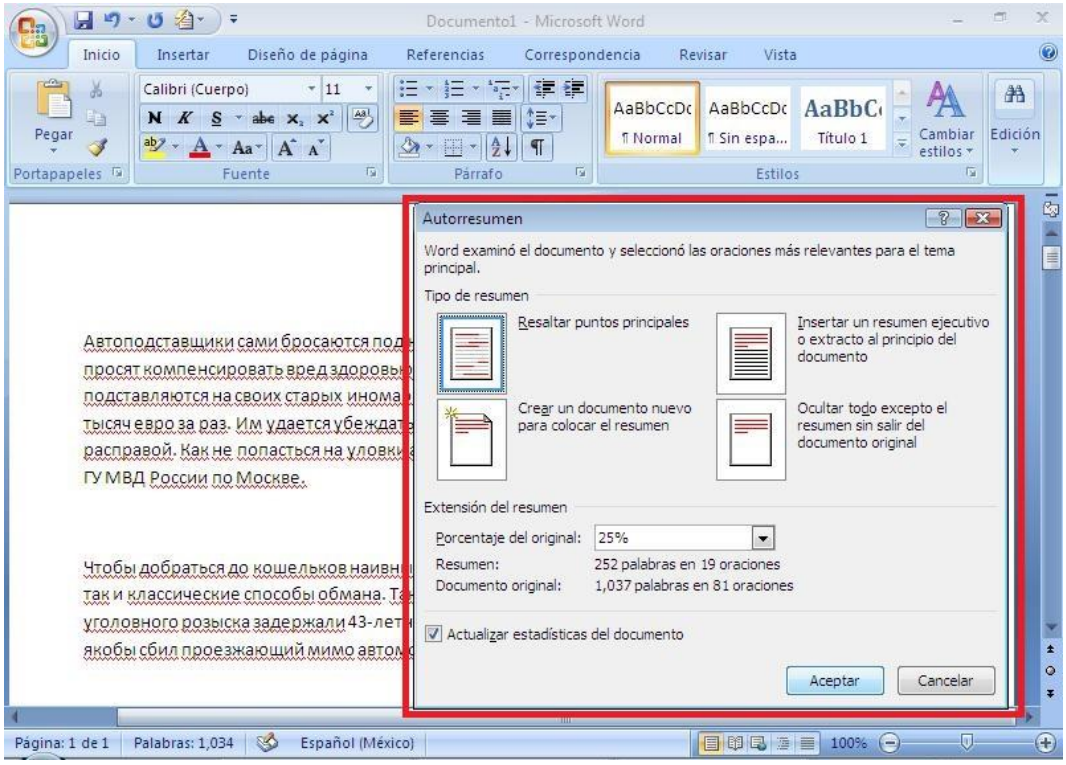


Figura A11. Interfaz para seleccionar los parametros.

11. El primer parámetro a modificar es "Tipo de resumen" el cual nos permitir seleccionar la forma en que Microsoft Office Word nos mostrará el resumen, las opciones de este parámetro son:
    a) Resaltar los puntos principales.
    b) Insertar un resumen ejecutivo o extracto al principio del documento.
    c) Crear un documento nuevo para colocar el resumen.
    d) Ocultar todo excepto el resumen sin salir del documento

Para esta tesis se seleccionó la opción "Crear un documento nuevo para colocar el resumen" el cual nos mostrará en un nuevo documento el resumen final, ya que omite las partes que nos son del resumen del documento original.

12. El segundo parámetro a modificar es "Extensión de resumen", las opciones a seleccionar son: 10 oraciones, 20 oraciones, 100 palabras o menos, 500 palabras o menos, 10% y 25%. Estos parámetros nos permiten seleccionar el tamaño del resumen.

    Para esta tesis se seleccionó la opción "25 %" ya que esta opción genera resúmenes con más de 100 palabras.

13. Por último se selecciona el botón "Aceptar" para generar el resumen automático.

Anexo 2. Medidas de la evaluación

En la tabla A2, se muestran las medidas de las herramientas comerciales y el valor más alto de los métodos del estado del arte, evaluadas con el sistema ROUGE.

Tabla 2. Evaluación de las herramientas comerciales y del método [Matias 2016].

Herramientas comerciales y métodos del estado del arte	Recuerdo ROUGE-1	Precisión ROUGE-1	F-measure ROUGE-1
Microsoft Word 2007 WindXp	0.94051	0.79760	0.86082
Microsoft Word 2007 Wind8	0.94010	0.79929	0.86204
Microsoft Word 2003 Wind8	0.93953	0.80058	0.86259
Microsoft Word 2003 WindXp	0.93974	0.80053	0.86260
Microsoft Word 2003 Wind7	0.93985	0.80043	0.86262
Microsoft Word 2003 WindVista	0.93945	0.80097	0.86279
Microsoft Word 2007 WindVista	0.93947	0.80118	0.86290
Microsoft Word 2007 Wind7	0.93967	0.80137	0.86311
Herramienta Text Compactor	0.92873	0.81397	0.86537
Tool4noobs	0.90739	0.84008	0.87045
Herramienta Open Text Summarizer	0.92021	0.84344	0.87860
T-Conspectus	0.91835	0.84747	0.87987
Método del estado del arte [Matias 16]	0.94409	0.85282	0.89497

Anexo 3. Evaluación de herramientas comerciales

A continuación se muestran los resultados detallados emitidos por la herramienta ROUGE para la evaluación de las herramientas comerciales, que se utilizaron en el desarrollo de esta tesis, primero se muestran las herramientas comerciales instalables y luego las herramientas en línea.

Herramientas comerciales instalables:

- Herramienta comercial instalable en Sistema Operativo *Windows XP* para *Microsoft Word 2007*

1 ROUGE-1 Average_R: 0.94051 (95%-conf.int. 0.93664 - 0.94446)
1 ROUGE-1 Average_P: 0.79760 (95%-conf.int. 0.78914 - 0.80679)
1 ROUGE-1 Average_F: 0.86082 (95%-conf.int. 0.85557 - 0.86633)
---
1 ROUGE-2 Average_R: 0.69754 (95%-conf.int. 0.68856 - 0.70705)
1 ROUGE-2 Average_P: 0.59107 (95%-conf.int. 0.58153 - 0.60158)
1 ROUGE-2 Average_F: 0.63808 (95%-conf.int. 0.62961 - 0.64762)
---
1 ROUGE-SU4 Average_R: 0.83523 (95%-conf.int. 0.82945 - 0.84098)
1 ROUGE-SU4 Average_P: 0.70720 (95%-conf.int. 0.69951 - 0.71648)
1 ROUGE-SU4 Average_F: 0.76374 (95%-conf.int. 0.75801 - 0.77010)

- Herramienta comercial instalable en Sistema Operativo *Windows 8* para *Microsoft Word 2007*

1 ROUGE-1 Average_R: 0.94010 (95%-conf.int. 0.93628 - 0.94407)

1 ROUGE-1 Average_P: 0.79929 (95%-conf.int. 0.79166 - 0.80815)

1 ROUGE-1 Average_F: 0.86204 (95%-conf.int. 0.85788 - 0.86733)

---

1 ROUGE-2 Average_R: 0.69640 (95%-conf.int. 0.68739 - 0.70537)

1 ROUGE-2 Average_P: 0.59187 (95%-conf.int. 0.58275 - 0.60207)

1 ROUGE-2 Average_F: 0.63846 (95%-conf.int. 0.62970 - 0.64799)

---

1 ROUGE-SU4 Average_R: 0.83447 (95%-conf.int. 0.82890 - 0.83991)

1 ROUGE-SU4 Average_P: 0.70850 (95%-conf.int. 0.70123 - 0.71710)

1 ROUGE-SU4 Average_F: 0.76459 (95%-conf.int. 0.75924 - 0.77058)

- Herramienta comercial instalable en Sistema Operativo *Windows 8* para *Microsoft Word 2003*

1 ROUGE-1 Average_R: 0.93953 (95%-conf.int. 0.93548 - 0.94360)

1 ROUGE-1 Average_P: 0.80058 (95%-conf.int. 0.79291 - 0.80897)

1 ROUGE-1 Average_F: 0.86259 (95%-conf.int. 0.85843 - 0.86760)

---

1 ROUGE-2 Average_R: 0.69623 (95%-conf.int. 0.68728 - 0.70545)

1 ROUGE-2 Average_P: 0.59305 (95%-conf.int. 0.58363 - 0.60335)

1 ROUGE-2 Average_F: 0.63910 (95%-conf.int. 0.63045 - 0.64842)

---

1 ROUGE-SU4 Average_R: 0.83394 (95%-conf.int. 0.82836 - 0.83950)

1 ROUGE-SU4 Average_P: 0.70966 (95%-conf.int. 0.70237 - 0.71801)

1 ROUGE-SU4 Average_F: 0.76507 (95%-conf.int. 0.75983 - 0.77114)

- Herramienta comercial instalable en Sistema Operativo *Windows XP* para *Microsoft Word 2003*

1 ROUGE-1 Average_R: 0.93974 (95%-conf.int. 0.93582 - 0.94360)

1 ROUGE-1 Average_P: 0.80053 (95%-conf.int. 0.79315 - 0.80872)

1 ROUGE-1 Average_F: 0.86260 (95%-conf.int. 0.85848 - 0.86764)

---

1 ROUGE-2 Average_R: 0.69660 (95%-conf.int. 0.68718 - 0.70567)

1 ROUGE-2 Average_P: 0.59330 (95%-conf.int. 0.58387 - 0.60368)

1 ROUGE-2 Average_F: 0.63937 (95%-conf.int. 0.63055 - 0.64874)

---

1 ROUGE-SU4 Average_R: 0.83426 (95%-conf.int. 0.82873 - 0.83975)

1 ROUGE-SU4 Average_P: 0.70974 (95%-conf.int. 0.70265 - 0.71829)

1 ROUGE-SU4 Average_F: 0.76521 (95%-conf.int. 0.75973 - 0.77117)

- Herramienta comercial instalable en Sistema Operativo *Windows 7* para *Microsoft Word 2003*

1 ROUGE-1 Average_R: 0.93985 (95%-conf.int. 0.93597 - 0.94382)

1 ROUGE-1 Average_P: 0.80043 (95%-conf.int. 0.79291 - 0.80907)

1 ROUGE-1 Average_F: 0.86262 (95%-conf.int. 0.85841 - 0.86765)

---

1 ROUGE-2 Average_R: 0.69615 (95%-conf.int. 0.68700 - 0.70528)

1 ROUGE-2 Average_P: 0.59260 (95%-conf.int. 0.58347 - 0.60279)

1 ROUGE-2 Average_F: 0.63880 (95%-conf.int. 0.63020 - 0.64828)

---

1 ROUGE-SU4 Average_R: 0.83410 (95%-conf.int. 0.82890 - 0.83967)

1 ROUGE-SU4 Average_P: 0.70937 (95%-conf.int. 0.70227 - 0.71829)

1 ROUGE-SU4 Average_F: 0.76495 (95%-conf.int. 0.75968 - 0.77108)

- Herramienta comercial instalable en Sistema Operativo *Windows VISTA* para *Microsoft Word 2003*

1 ROUGE-1 Average_R: 0.93945 (95%-conf.int. 0.93562 - 0.94344)

1 ROUGE-1 Average_P: 0.80097 (95%-conf.int. 0.79355 - 0.80944)

1 ROUGE-1 Average_F: 0.86279 (95%-conf.int. 0.85862 - 0.86791)

---

1 ROUGE-2 Average_R: 0.69593 (95%-conf.int. 0.68682 - 0.70507)

1 ROUGE-2 Average_P: 0.59309 (95%-conf.int. 0.58414 - 0.60335)

1 ROUGE-2 Average_F: 0.63900 (95%-conf.int. 0.63040 - 0.64840)

---

1 ROUGE-SU4 Average_R: 0.83371 (95%-conf.int. 0.82816 - 0.83927)

1 ROUGE-SU4 Average_P: 0.70985 (95%-conf.int. 0.70280 - 0.71847)

1 ROUGE-SU4 Average_F: 0.76510 (95%-conf.int. 0.75977 - 0.77111)

- Herramienta comercial instalable en Sistema Operativo *Windows VISTA* para *Microsoft Word 2007*

1 ROUGE-1 Average_R: 0.93947 (95%-conf.int. 0.93556 - 0.94345)

1 ROUGE-1 Average_P: 0.80118 (95%-conf.int. 0.79364 - 0.80981)

1 ROUGE-1 Average_F: 0.86290 (95%-conf.int. 0.85867 - 0.86799)

---

1 ROUGE-2 Average_R: 0.69584 (95%-conf.int. 0.68660 - 0.70502)

1 ROUGE-2 Average_P: 0.59309 (95%-conf.int. 0.58401 - 0.60340)

1 ROUGE-2 Average_F: 0.63895 (95%-conf.int. 0.63034 - 0.64840)

---

1 ROUGE-SU4 Average_R: 0.83372 (95%-conf.int. 0.82817 - 0.83928)

1 ROUGE-SU4 Average_P: 0.71000 (95%-conf.int. 0.70286 - 0.71868)

1 ROUGE-SU4 Average_F: 0.76517 (95%-conf.int. 0.75996 - 0.77117)

- Herramienta comercial instalable en Sistema Operativo *Windows 7* para *Microsoft Word 2007*

1 ROUGE-1 Average_R: 0.93967 (95%-conf.int. 0.93584 - 0.94366)
1 ROUGE-1 Average_P: 0.80137 (95%-conf.int. 0.79375 - 0.80992)
1 ROUGE-1 Average_F: 0.86311 (95%-conf.int. 0.85894 - 0.86818)
---
1 ROUGE-2 Average_R: 0.69578 (95%-conf.int. 0.68687 - 0.70484)
1 ROUGE-2 Average_P: 0.59312 (95%-conf.int. 0.58394 - 0.60353)
1 ROUGE-2 Average_F: 0.63895 (95%-conf.int. 0.63051 - 0.64839)
---
1 ROUGE-SU4 Average_R: 0.83387 (95%-conf.int. 0.82830 - 0.83947)
1 ROUGE-SU4 Average_P: 0.71017 (95%-conf.int. 0.70288 - 0.71878)
1 ROUGE-SU4 Average_F: 0.76534 (95%-conf.int. 0.76013 - 0.77130)

Herramientas comerciales en línea:

- Herramienta comercial en línea Text Compactor

1 ROUGE-1 Average_R: 0.92873 (95%-conf.int. 0.92416 - 0.93338)
1 ROUGE-1 Average_P: 0.81397 (95%-conf.int. 0.80674 - 0.82151)
1 ROUGE-1 Average_F: 0.86537 (95%-conf.int. 0.86118 - 0.86927)
---
1 ROUGE-2 Average_R: 0.66494 (95%-conf.int. 0.65646 - 0.67343)
1 ROUGE-2 Average_P: 0.58195 (95%-conf.int. 0.57379 - 0.59031)
1 ROUGE-2 Average_F: 0.61913 (95%-conf.int. 0.61171 - 0.62706)
---
1 ROUGE-SU4 Average_R: 0.81489 (95%-conf.int. 0.80950 - 0.82078)
1 ROUGE-SU4 Average_P: 0.71313 (95%-conf.int. 0.70584 - 0.72025)
1 ROUGE-SU4 Average_F: 0.75867 (95%-conf.int. 0.75389 - 0.76346)

- Herramienta comercial en línea Tools4Noobs

1 ROUGE-1 Average_R: 0.90739 (95%-conf.int. 0.90125 - 0.91281)

1 ROUGE-1 Average_P: 0.84008 (95%-conf.int. 0.83308 - 0.84731)

1 ROUGE-1 Average_F: 0.87045 (95%-conf.int. 0.86649 - 0.87420)

---

1 ROUGE-2 Average_R: 0.63688 (95%-conf.int. 0.62844 - 0.64521)

1 ROUGE-2 Average_P: 0.58957 (95%-conf.int. 0.58158 - 0.59844)

1 ROUGE-2 Average_F: 0.61090 (95%-conf.int. 0.60338 - 0.61846)

---

1 ROUGE-SU4 Average_R: 0.79031 (95%-conf.int. 0.78374 - 0.79669)

1 ROUGE-SU4 Average_P: 0.73118 (95%-conf.int. 0.72430 - 0.73844)

1 ROUGE-SU4 Average_F: 0.75783 (95%-conf.int. 0.75309 - 0.76273)

- Herramienta comercial en línea OTS

1 ROUGE-1 Average_R: 0.92024 (95%-conf.int. 0.91576 - 0.92463)

1 ROUGE-1 Average_P: 0.84344 (95%-conf.int. 0.83644 - 0.85042)

1 ROUGE-1 Average_F: 0.87860 (95%-conf.int. 0.87439 - 0.88230)

---

1 ROUGE-2 Average_R: 0.67620 (95%-conf.int. 0.66778 - 0.68415)

1 ROUGE-2 Average_P: 0.62017 (95%-conf.int. 0.61076 - 0.62963)

1 ROUGE-2 Average_F: 0.64583 (95%-conf.int. 0.63748 - 0.65399)

---

1 ROUGE-SU4 Average_R: 0.81359 (95%-conf.int. 0.80787 - 0.81919)

1 ROUGE-SU4 Average_P: 0.74526 (95%-conf.int. 0.73813 - 0.75237)

1 ROUGE-SU4 Average_F: 0.77652 (95%-conf.int. 0.77126 - 0.78158)

•Herramienta comercial en línea *T-Conspectus*

1 ROUGE-1 Average_R: 0.91835 (95%-conf.int. 0.91302 - 0.92319)

1 ROUGE-1 Average_P: 0.84747 (95%-conf.int. 0.84053 - 0.85392)

1 ROUGE-1 Average_F: 0.87987 (95%-conf.int. 0.87605 - 0.88377)

---

1 ROUGE-2 Average_R: 0.68254 (95%-conf.int. 0.67293 - 0.69155)

1 ROUGE-2 Average_P: 0.62974 (95%-conf.int. 0.62068 - 0.63867)

1 ROUGE-2 Average_F: 0.65390 (95%-conf.int. 0.64508 - 0.66259)

---

1 ROUGE-SU4 Average_R: 0.81496 (95%-conf.int. 0.80812 - 0.82136)

1 ROUGE-SU4 Average_P: 0.75144 (95%-conf.int. 0.74431 - 0.75824)

1 ROUGE-SU4 Average_F: 0.78047 (95%-conf.int. 0.77463 - 0.78592)

Anexo 4. Evaluación de métodos del estado del arte

A continuación se muestran los parámetros para la evaluación de cada uno de los resúmenes de los métodos del estado del arte, así como sus resultados emitidos por la herramienta ROUGE:

Experimento 1:

Parámetros:
Modelo de texto: Bigramas
Pre-procesamiento: No
Pendiente: 0.3
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:
1 ROUGE-1 Average_R: 0.94558 (95%-conf.int. 0.94178 - 0.94945)
1 ROUGE-1 Average_P: 0.83682 (95%-conf.int. 0.83060 - 0.84285)
1 ROUGE-1 Average_F: 0.88671 (95%-conf.int. 0.88299 - 0.89008)
---
1 ROUGE-2 Average_R: 0.74019 (95%-conf.int. 0.73144 - 0.74962)
1 ROUGE-2 Average_P: 0.65485 (95%-conf.int. 0.64617 - 0.66333)
1 ROUGE-2 Average_F: 0.69398 (95%-conf.int. 0.68562 - 0.70239)
---
1 ROUGE-SU4 Average_R: 0.85396 (95%-conf.int. 0.84850 - 0.85978)
1 ROUGE-SU4 Average_P: 0.75489 (95%-conf.int. 0.74830 - 0.76137)
1 ROUGE-SU4 Average_F: 0.80030 (95%-conf.int. 0.79545 - 0.80551)

Experimento 2:

Parámetros:

Modelo de texto: Bigramas
Pre-procesamiento: No
Pendiente: 0.8
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94489 (95%-conf.int. 0.94120 - 0.94859)
1 ROUGE-1 Average_P: 0.84511 (95%-conf.int. 0.83949 - 0.85118)
1 ROUGE-1 Average_F: 0.89113 (95%-conf.int. 0.88767 - 0.89460)
---
1 ROUGE-2 Average_R: 0.74830 (95%-conf.int. 0.73978 - 0.75703)
1 ROUGE-2 Average_P: 0.66887 (95%-conf.int. 0.66018 - 0.67754)
1 ROUGE-2 Average_F: 0.70549 (95%-conf.int. 0.69706 - 0.71383)
---
1 ROUGE-SU4 Average_R: 0.85695 (95%-conf.int. 0.85161 - 0.86224)
1 ROUGE-SU4 Average_P: 0.76563 (95%-conf.int. 0.75929 - 0.77204)
1 ROUGE-SU4 Average_F: 0.80772 (95%-conf.int. 0.80275 - 0.81279)

Experimento 3:

Parámetros:

Modelo de texto: Bigramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94230 (95%-conf.int. 0.93355 - 0.94813)
1 ROUGE-1 Average_P: 0.84349 (95%-conf.int. 0.83542 - 0.85057)
1 ROUGE-1 Average_F: 0.88898 (95%-conf.int. 0.88156 - 0.89417)
---
1 ROUGE-2 Average_R: 0.74365 (95%-conf.int. 0.73255 - 0.75338)
1 ROUGE-2 Average_P: 0.66523 (95%-conf.int. 0.65437 - 0.67491)
1 ROUGE-2 Average_F: 0.70137 (95%-conf.int. 0.69089 - 0.71046)
---
1 ROUGE-SU4 Average_R: 0.85232 (95%-conf.int. 0.84290 - 0.85953)
1 ROUGE-SU4 Average_P: 0.76206 (95%-conf.int. 0.75283 - 0.76961)
1 ROUGE-SU4 Average_F: 0.80362 (95%-conf.int. 0.79493 - 0.81017)

Experimento 4:

Parámetros:

Modelo de texto: Bigramas
Pre-procesamiento: No
Pendiente: 0.75
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.93882 (95%-conf.int. 0.92915 - 0.94545)
1 ROUGE-1 Average_P: 0.84038 (95%-conf.int. 0.83022 - 0.84862)
1 ROUGE-1 Average_F: 0.88563 (95%-conf.int. 0.87676 - 0.89175)
---
1 ROUGE-2 Average_R: 0.73871 (95%-conf.int. 0.72739 - 0.74969)
1 ROUGE-2 Average_P: 0.66091 (95%-conf.int. 0.64966 - 0.67096)
1 ROUGE-2 Average_F: 0.69667 (95%-conf.int. 0.68536 - 0.70683)
---
1 ROUGE-SU4 Average_R: 0.84934 (95%-conf.int. 0.83965 - 0.85687)
1 ROUGE-SU4 Average_P: 0.75946 (95%-conf.int. 0.74993 - 0.76784)
1 ROUGE-SU4 Average_F: 0.80074 (95%-conf.int. 0.79149 - 0.80783)

Experimento 5:

Parámetros:

Modelo de texto: Trigramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados

1 ROUGE-1 Average_R: 0.93789 (95%-conf.int. 0.92799 - 0.94566)
1 ROUGE-1 Average_P: 0.84576 (95%-conf.int. 0.83734 - 0.85359)
1 ROUGE-1 Average_F: 0.88828 (95%-conf.int. 0.88006 - 0.89471)
---
1 ROUGE-2 Average_R: 0.74411 (95%-conf.int. 0.73210 - 0.75548)
1 ROUGE-2 Average_P: 0.66974 (95%-conf.int. 0.65899 - 0.67993)
1 ROUGE-2 Average_F: 0.70413 (95%-conf.int. 0.69290 - 0.71443)
---
1 ROUGE-SU4 Average_R: 0.85037 (95%-conf.int. 0.83992 - 0.85975)
1 ROUGE-SU4 Average_P: 0.76556 (95%-conf.int. 0.75603 - 0.77424)
1 ROUGE-SU4 Average_F: 0.80473 (95%-conf.int. 0.79563 - 0.81286)

Experimento 6:

Parámetros:

Modelo de texto: Cuatro gramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados

1 ROUGE-1 Average_R: 0.94336 (95%-conf.int. 0.93949 - 0.94708)
1 ROUGE-1 Average_P: 0.85317 (95%-conf.int. 0.84757 - 0.85865)
1 ROUGE-1 Average_F: 0.89487 (95%-conf.int. 0.89173 - 0.89803)
---
1 ROUGE-2 Average_R: 0.75216 (95%-conf.int. 0.74364 - 0.76043)
1 ROUGE-2 Average_P: 0.67966 (95%-conf.int. 0.67156 - 0.68691)
1 ROUGE-2 Average_F: 0.71318 (95%-conf.int. 0.70536 - 0.72033)
---
1 ROUGE-SU4 Average_R: 0.85725 (95%-conf.int. 0.85175 - 0.86259)
1 ROUGE-SU4 Average_P: 0.77445 (95%-conf.int. 0.76850 - 0.78018)
1 ROUGE-SU4 Average_F: 0.81271 (95%-conf.int. 0.80792 - 0.81726)

Experimento 7:

Parámetros:

Modelo de texto: Cinco gramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94398 (95%-conf.int. 0.93999 - 0.94757)
1 ROUGE-1 Average_P: 0.85226 (95%-conf.int. 0.84662 - 0.85818)
1 ROUGE-1 Average_F: 0.89459 (95%-conf.int. 0.89140 - 0.89775)
---
1 ROUGE-2 Average_R: 0.75316 (95%-conf.int. 0.74447 - 0.76117)
1 ROUGE-2 Average_P: 0.67944 (95%-conf.int. 0.67133 - 0.68687)
1 ROUGE-2 Average_F: 0.71347 (95%-conf.int. 0.70547 - 0.72056)
---
1 ROUGE-SU4 Average_R: 0.85819 (95%-conf.int. 0.85272 - 0.86337)
1 ROUGE-SU4 Average_P: 0.77397 (95%-conf.int. 0.76799 - 0.77988)
1 ROUGE-SU4 Average_F: 0.81282 (95%-conf.int. 0.80807 - 0.81726)

Experimento 8:

Parámetros:

Modelo de texto: Cinco gramas
Pre-procesamiento: No
Pendiente: 0.75
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.93950 (95%-conf.int. 0.93004 - 0.94591)
1 ROUGE-1 Average_P: 0.84878 (95%-conf.int. 0.83934 - 0.85689)
1 ROUGE-1 Average_F: 0.89065 (95%-conf.int. 0.88206 - 0.89669)
---
1 ROUGE-2 Average_R: 0.74683 (95%-conf.int. 0.73618 - 0.75624)
1 ROUGE-2 Average_P: 0.67428 (95%-conf.int. 0.66488 - 0.68348)
1 ROUGE-2 Average_F: 0.70777 (95%-conf.int. 0.69820 - 0.71665)
---
1 ROUGE-SU4 Average_R: 0.85317 (95%-conf.int. 0.84394 - 0.86032)
1 ROUGE-SU4 Average_P: 0.76997 (95%-conf.int. 0.76130 - 0.77787)
1 ROUGE-SU4 Average_F: 0.80835 (95%-conf.int. 0.79995 - 0.81491)

Experimento 9:

Parámetros:

Modelo de texto: Seis gramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94415 (95%-conf.int. 0.94031 - 0.94786)
1 ROUGE-1 Average_P: 0.85089 (95%-conf.int. 0.84509 - 0.85672)
1 ROUGE-1 Average_F: 0.89392 (95%-conf.int. 0.89089 - 0.89706)
---
1 ROUGE-2 Average_R: 0.75306 (95%-conf.int. 0.74485 - 0.76104)
1 ROUGE-2 Average_P: 0.67810 (95%-conf.int. 0.67029 - 0.68543)
1 ROUGE-2 Average_F: 0.71269 (95%-conf.int. 0.70484 - 0.71977)
---
1 ROUGE-SU4 Average_R: 0.85833 (95%-conf.int. 0.85288 - 0.86364)
1 ROUGE-SU4 Average_P: 0.77269 (95%-conf.int. 0.76669 - 0.77859)
1 ROUGE-SU4 Average_F: 0.81218 (95%-conf.int. 0. 80749 - 0.81668)

Experimento 10:

Parámetros:

Modelo de texto: Siete gramas
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94409 (95%-conf.int. 0.94023 - 0.94773)
1 ROUGE-1 Average_P: 0.85285 (95%-conf.int. 0.84709 - 0.85868)
1 ROUGE-1 Average_F: 0.89497 (95%-conf.int. 0.89176 - 0.89826)
---
1 ROUGE-2 Average_R: 0.75229 (95%-conf.int. 0.74439 - 0.76053)
1 ROUGE-2 Average_P: 0.67914 (95%-conf.int. 0.67112 - 0.68694)
1 ROUGE-2 Average_F: 0.71292 (95%-conf.int. 0.70522 - 0.72026)
---
1 ROUGE-SU4 Average_R: 0.85785 (95%-conf.int. 0.85246 - 0.86300)
1 ROUGE-SU4 Average_P: 0.77415 (95%-conf.int. 0.76776 - 0.77998)
1 ROUGE-SU4 Average_F: 0.81277 (95%-conf.int. 0.80809 - 0.81728)

Experimento 11:

Parámetros:

Modelo de texto: Bolsa de palabras
Pre-procesamiento: No
Pendiente: 0.3
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94378 (95%-conf.int. 0.93978 - 0.94776)
1 ROUGE-1 Average_P: 0.84359 (95%-conf.int. 0.83760 - 0.85023)
1 ROUGE-1 Average_F: 0.88969 (95%-conf.int. 0.88610 - 0.89336)
---
1 ROUGE-2 Average_R: 0.73980 (95%-conf.int. 0.72996 - 0.74932)
1 ROUGE-2 Average_P: 0.66119 (95%-conf.int. 0.65161 - 0.67048)
1 ROUGE-2 Average_F: 0.69737 (95%-conf.int. 0.68835 - 0.70635)
---
1 ROUGE-SU4 Average_R: 0.85260 (95%-conf.int. 0.84657 - 0.85861)
1 ROUGE-SU4 Average_P: 0.76137 (95%-conf.int. 0.75447 - 0.76824)
1 ROUGE-SU4 Average_F: 0.80332 (95%-conf.int. 0.79757 - 0.80874)

Experimento 12:

Parámetros:

Modelo de texto: Bolsa de palabras
Pre-procesamiento: No
Pendiente: 0.8
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.93666 (95%-conf.int. 0.92865 - 0.94289)
1 ROUGE-1 Average_P: 0.83235 (95%-conf.int. 0.82302 - 0.84093)
1 ROUGE-1 Average_F: 0.87966 (95%-conf.int. 0.87159 - 0.88614)
---
1 ROUGE-2 Average_R: 0.72211 (95%-conf.int. 0.70879 - 0.73417)
1 ROUGE-2 Average_P: 0.64236 (95%-conf.int. 0.62858 - 0.65551)
1 ROUGE-2 Average_F: 0.67865 (95%-conf.int. 0.66570 - 0.69104)
---
1 ROUGE-SU4 Average_R: 0.84247 (95%-conf.int. 0.83237 - 0.85052)
1 ROUGE-SU4 Average_P: 0.74784 (95%-conf.int. 0.73701 - 0.75769)
1 ROUGE-SU4 Average_F: 0.79079 (95%-conf.int. 0.78073 - 0.79893)

Experimento 13:

Parámetros:

Modelo de texto: Bolsa de palabras
Pre-procesamiento: No
Pendiente: 0.625
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94356 (95%-conf.int. 0.93947 - 0.94765)
1 ROUGE-1 Average_P: 0.84538 (95%-conf.int. 0.83904 - 0.85146)
1 ROUGE-1 Average_F: 0.89051 (95%-conf.int. 0.88674 - 0.89412)
---
1 ROUGE-2 Average_R: 0.74108 (95%-conf.int. 0.73128 - 0.75086)
1 ROUGE-2 Average_P: 0.66376 (95%-conf.int. 0.65369 - 0.67330)
1 ROUGE-2 Average_F: 0.69931 (95%-conf.int. 0.69021 - 0.70881)
---
1 ROUGE-SU4 Average_R: 0.85315 (95%-conf.int. 0.84700 - 0.85890)
1 ROUGE-SU4 Average_P: 0.76359 (95%-conf.int. 0.75641 - 0.77023)
1 ROUGE-SU4 Average_F: 0.80473 (95%-conf.int. 0.79934 - 0.80985)

Experimento 14:

Parámetros:

Modelo de texto: Bolsa de palabras

Pre-procesamiento: No

Pendiente: 0.90

Función de aptitud: 0.4 β + 0.6 δ

Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94594 (95%-conf.int. 0.94218 - 0.94975)

1 ROUGE-1 Average_P: 0.84478 (95%-conf.int. 0.83858 - 0.85123)

1 ROUGE-1 Average_F: 0.89128 (95%-conf.int. 0.88777 - 0.89492)

---

1 ROUGE-2 Average_R: 0.74794 (95%-conf.int. 0.73824 - 0.75760)

1 ROUGE-2 Average_P: 0.66767 (95%-conf.int. 0.65819 - 0.67694)

1 ROUGE-2 Average_F: 0.70458 (95%-conf.int. 0.69581 - 0.71352)

---

1 ROUGE-SU4 Average_R: 0.85740 (95%-conf.int. 0.85164 - 0.86325)

1 ROUGE-SU4 Average_P: 0.76490 (95%-conf.int. 0.75780 - 0.77178)

1 ROUGE-SU4 Average_F: 0.80739 (95%-conf.int. 0.80192 - 0.81265)

Experimento 15:

Parámetros:

Modelo de texto: Bolsa de palabras

Pre-procesamiento: No

Pendiente: 0.90

Función de aptitud: 0.6 β + 0.4 δ

Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94466 (95%-conf.int. 0.94072 - 0.94830)

1 ROUGE-1 Average_P: 0.84364 (95%-conf.int. 0.83723 - 0.84963)

1 ROUGE-1 Average_F: 0.89011 (95%-conf.int. 0.88655 - 0.89359)

---

1 ROUGE-2 Average_R: 0.74573 (95%-conf.int. 0.73692 - 0.75450)

1 ROUGE-2 Average_P: 0.66578 (95%-conf.int. 0.65726 - 0.67459)

1 ROUGE-2 Average_F: 0.70255 (95%-conf.int. 0.69470 - 0.71026)

---

1 ROUGE-SU4 Average_R: 0.85575 (95%-conf.int. 0.85033 - 0.86140)

1 ROUGE-SU4 Average_P: 0.76348 (95%-conf.int. 0.75655 - 0.77001)

1 ROUGE-SU4 Average_F: 0.80589 (95%-conf.int. 0.80056 - 0.81073)

Experimento 16:

Parámetros:

Modelo de texto: Bolsa de palabras
Pre-procesamiento: No
Pendiente: 0.90
Función de aptitud: 0.3 β + 0.7 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94459 (95%-conf.int. 0.94061 - 0.94824)
1 ROUGE-1 Average_P: 0.84313 (95%-conf.int. 0.83666 - 0.84961)
1 ROUGE-1 Average_F: 0.88978 (95%-conf.int. 0.88604 - 0.89338)
---
1 ROUGE-2 Average_R: 0.74168 (95%-conf.int. 0.73236 - 0.75080)
1 ROUGE-2 Average_P: 0.66211 (95%-conf.int. 0.65210 - 0.67135)
1 ROUGE-2 Average_F: 0.69870 (95%-conf.int. 0.68917 - 0.70773)
---
1 ROUGE-SU4 Average_R: 0.85473 (95%-conf.int. 0.84888 - 0.85993)
1 ROUGE-SU4 Average_P: 0.76227 (95%-conf.int. 0.75499 - 0.76916)
1 ROUGE-SU4 Average_F: 0.80475 (95%-conf.int. 0.79906 - 0.81015)

Experimento 17:

Parámetros:

Modelo de texto: Bolsa de palabras
Pre-procesamiento: No
Pendiente: 0.90
Función de aptitud: 0.5 β + 0.5 δ
Operador de selección: Ruleta

Resultados:

1 ROUGE-1 Average_R: 0.94628 (95%-conf.int. 0.94277 - 0.94996)
1 ROUGE-1 Average_P: 0.84146 (95%-conf.int. 0.83527 - 0.84773)
1 ROUGE-1 Average_F: 0.88963 (95%-conf.int. 0.88616 - 0.89314)
---
1 ROUGE-2 Average_R: 0.74823 (95%-conf.int. 0.73948 - 0.75711)
1 ROUGE-2 Average_P: 0.66526 (95%-conf.int. 0.65649 - 0.67433)
1 ROUGE-2 Average_F: 0.70339 (95%-conf.int. 0.69505 - 0.71187)
---
1 ROUGE-SU4 Average_R: 0.85786 (95%-conf.int. 0.85221 - 0.86334)
1 ROUGE-SU4 Average_P: 0.76204 (95%-conf.int. 0.75552 - 0.76865)
1 ROUGE-SU4 Average_F: 0.80604 (95%-conf.int. 0.80098 - 0.81111)