D AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Determinación del desempeño de resúmenes generados automáticamente para el idioma español"

Tesis
Para Obtener el Título de
Ingeniero en Software

Que Presenta

Nancy Nagay González González

Directora:
Dra. Yulia Nikolaevna Ledeneva

TIANGUISTENCO, MÉX. Noviembre 2016

Resumen

En la actualidad el crecimiento rápido de internet ha provocado gran cantidad de información que está disponible en formato electrónico que crece de manera exponencial. Esto da lugar a millones de documentos cuya magnitud dificulta en gran medida su manejo. Esto lleva a la búsqueda de nuevos programas que suplan las tareas cada vez más específicas. Por ejemplo, cuando se quiere saber más de un tema es necesario revisar más de un documento ya sea en internet o en la computadora. Después se necesita identificar un documento con mayor relevancia de información para nuestros fines. Lo que facilitaría la tarea de búsqueda es si pudiéramos tomar solamente las partes más relevantes (documentos, renglones, oraciones, frases o palabras) y leer solo lo importante. Precisamente estas partes formarían un resumen de un tema buscado.

Un resumen se define como un texto muy corto que comunica la información más importante del documento original (Ledeneva 2008). Esta tesis se trata de la generación automática de resúmenes, que es una tarea de gran utilidad para hacer las tareas más rápidas con ayuda de una herramienta.

En el presente trabajo se evaluan las diferentes herramientas comerciales tanto en línea como las que son instalables para saber su desempeño en el idioma español. Los experimentos se llevan a cabo sobre el corpus TER (corpus en español). Posteriormente, el desempeño de las herramientas comercioales se compara con el método del estado de arte de (Matias 2016), ya que se había probado como uno de los mejores métodos para los idiomas inglés y portugués (Matias 2013, Ibañez 2013).

Contenido

Página

Lista de figuras

Página

Lista de tablas

CAPÍTULO 1.

Introducción

1.1 Relevancia de la cantidad de información

Hoy en día es muy importante el acceso a la información textual por lo que ha aumentado la cantidad de esta información en internet. Se requiere el desarrollo y evaluación de las herramientas comerciales y de los métodos del estado de arte para saber su desempeño y poder mejorar la calidad.

Una de las tareas más importantes dentro del área del Procesamiento de Lenguaje Natural (PLN) es la generación automática de resúmenes. Actualmente se requiere generar los resúmenes de noticias, artículos

científicos, libros, de la información busqueda en internet (Ledeneva 08a, Ledeneva 08b).

(Gelbukh, 2010) define el procesamiento de lenguaje natural de la siguiente manera:

> *"El procesamiento de lenguaje natural (PLN, denominado también NLP por sus siglas en el idioma inglés) se entiende la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o sonidos del lenguaje".*

1.2 Definición de un resumen

(Ledeneva 2008) define el concepto de resumen, cito textualmente:

> *"El resumen del documento es un texto (muy) corto que comunica la información más importante del documento original".*

Por otro lado (Garcés 2008) define el término de resumen como:

> *"Conocemos el término resumen como la representación abreviada y precisa de una o varias fuentes. El resumen de una fuente debe de representar las ideas principales de ésta, ya que el contenido global de la fuente gira alrededor de estas ideas. Para que un resumen sea correcto, debe tener una serie de propiedades, entre ellas tenemos la objetividad, la coherencia y la facilidad de compresión. Sin embargo, no*

siempre es suficiente evaluar la utilidad de un resumen en relación a esas propiedades, ya que está altamente dependiente del usuario que lo está leyendo".

1.3 Resumen automático

Un resumen es un texto generado nuevamente, producido por alguna representación interna que resulta después del texto de entrada. Esta idea surge con el motivo de mejorar la calidad final del texto del resumen y, para ello se lleva a cabo un proceso posterior acortando y revisando el material (Mani et ál., 1999; Knight y Marcu, 2000).

El resumen automático es la técnica que se utiliza para generar resúmenes de una o varias fuentes mediante un programa informático (software).

1.4 Tipos de resúmenes

En el área de procesamiento de lenguaje natural las formas para generar resúmenes de manera automática se clasifican principalmente en dos: resúmenes extractivos y abstractivos.

En la tesis, se toma en cuenta los resúmenes de tipo extractivo porque están formados por la combinación de fragmentos o enunciados que son extraídos del documento fuente.

1.4.1 Resúmenes extractivos

Es el enfoque que trata las fuentes como un conjunto de frases, en este tipo de resumen no se limitan a extraer fragmentos relevantes del texto, sino que se hace un análisis más profundo para poder comprenderlo y poder generar un nuevo resumen.

Para llevar a cabo el proceso de extracción, el primer paso es identificar las frases más relevantes de las fuentes. Para proceder en esta fase se pueden utilizar criterios de selección basados en métodos estadísticos durante el tratamiento de la fuente.

En segundo lugar, a partir de los fragmentos extraídos previamente, hay que generar un resumen procurando perder la menor cantidad de información posible y evitando la redundancia.

Por último, pueden aplicarse tratamientos posteriores al resumen para conseguir un texto coherente y bien relacionado, uniendo los diferentes fragmentos o completando frases que pueden haber quedado incompletas.

1.4.2 Resúmenes abstractivos

Los resúmenes abstractivos no se limitan a extraer fragmentos relevantes de un texto, si no que analizan el texto con mayor profundidad para poder comprenderlo y poder generar un nuevo resumen a partir de la información analizada de la fuente.

El resumen por abstracción es por lo tanto similar al método que emplearía un ser humano para hacer un resumen propio de un texto, y la dificultad de reproducir ese método por una computadora es lo que hace que los métodos de extracción sean los que han conseguido avances más significativos.

1.5 Herramientas comerciales y métodos del estado de arte

Actualmente existen trabajos que han determinado la calidad de los resúmenes para idioma inglés y portugués. Las herramientas comerciales que ayudan para la creación de resúmenes en línea son: Open Text Summarizer (OTS), Summarizing, Text Compactor. Las herramientas instalables son: Microsoft Word 2003 (Sistema operativo XP), Microsoft Word 2007 (Sistema operativo XP), Microsoft Word 2003 (Sistema operativo 7), Microsoft Word 2007 (Sistema operativo 7), Microsoft Word 2003 (Sistema operativo 8.1), Microsoft Word 2007 (Sistema operativo 8.1), Copernic Summarizer (Sistema operativo 7) (CS/7), entre otros (Ver en el Anexo 2).

Dentro de los métodos del estado de arte que han propuesto generar los resúmenes automáticamente en los idiomas inglés e portugués, se encuentran los trabajos: (Matias 2016), (Ledeneva 2008a), (Ledeneva 2008b), (Ledeneva 2011), (Ledeneva 2014), (Mihalcea 2004), entre otros.

1.6 Planteamiento del problema

Actualmente el rápido crecimiento de internet ha provocado la gran cantidad de información disponible en formato electrónico que crece de manera exponencial, dando lugar a millones de documentos cuya magnitud dificulta en gran medida su manejo, por ello se lleva a cabo la búsqueda de nuevos programas que suplan las tareas cada vez más específicas.

Cuando se quiere saber más de un tema es necesario leer más de un documento ya sea en internet o en nuestra computadora. Se necesita identificar un documento con mayor relevancia de información para nuestros fines, que facilitaría la tarea de búsqueda si pudiéramos tomar documentos y leer solo lo importante o mejor aún "renglones", "frases" o "palabras" relevantes. Debido a esto la generación automática de resúmenes es de gran utilidad para hacer las tareas más rápidas con ayuda de una herramienta.

La pregunta de investigación de esta tesis es:

¿Qué tan eficiente son las herramientas comerciales y cuál es la mejor herramienta comercial para la generación de resúmenes automáticos para un solo documento en el idioma español?

1.7 Objetivos

1.7.1 Objetivo general

- Determinar la calidad de los resúmenes generados automáticamente del corpus en el idioma español.

1.7.2 Objetivos específicos

- Presentar la descripción y el estado de arte de las herramientas de la generación automática de los resúmenes que funcionan para el idioma español.

- Generar los resúmenes utilizando herramientas comerciales y el método del estado de arte en el idioma español, para el corpus en el idioma español que lleva por el nombre TER (Matias 2016).

- Evaluar los resúmenes generados con las herramientas comerciales y el método del estado de arte.

- Analizar los resúmenes generados con las herramientas comerciales y el método del estado de arte.

- Comparar los resultados generados con las herramientas comerciales y el método del estado de arte.

- Encontrar la mejor herramienta comercial para la evaluación de resúmenes en el corpus TER.

- Aplicar el evaluador ROUGE para evaluar la calidad de las herramientas comerciales.

1.8 Hipótesis

Al aplicarse el método de evaluación ROUGE (Lin 2004), se podrá determinar la calidad de las herramientas comerciales en el idioma español.

1.9 Delimitación del problema

- Los resúmenes del corpus TER tendrán que ser de 100 palabras para evaluarlas en ROUGE. Ya que los resúmenes generados por el humano son de 100 palabras.

- Se utilizarán herramientas tanto en línea como instalables para poder evaluarlas.

1.10 Estructura de la tesis

En el presente capítulo, se da una introducción y antecedentes sobre la generación automática de resúmenes. Se presenta el problema, los objetivos generales y específicos, la hipótesis, la delimitación del problema.

El resto del documento está organizado de la siguiente manera:

En el capítulo 2, se definen los conceptos utilizados en esta tesis. Se presentan los conceptos necesarios para entender la tesis. Finalmente, se mencionan la medida de evaluación más utilizada para la generación de resúmenes.

En el capítulo 3, se presenta el estado del arte para la generación de resúmenes automáticos.

En el capítulo 4, se describe la metodología de trabajo propuesta en esta tesis. Se detalla cada una de las etapas que se siguen.

En el capítulo 5, se describe el corpus que se utilizó para generar los resúmenes, también se menciona la herramienta utilizada para la evaluación. Principalmente en este capítulo se muestran las gráficas con los resultados de los experimentos para el idioma español.

Finalmente en el capítulo 6, se describen las conclusiones obtenidas de la evaluación y comparación de las herramientas comerciales con el método del estado de arte. Además se presenta el trabajo futuro.

CAPÍTULO 2.

Marco Teórico

En este capítulo, se definen los conceptos utilizados en esta tesis. Se presentan los conceptos necesarios para entender la tesis. Finalmente, se mencionan la medida de evaluación más utilizada para la generación de resúmenes.

2.1 Procesamiento de lenguaje natural

(Gelbukh 2014) define el procesamiento de lenguaje natural de la siguiente manera:

> *"El procesamiento del lenguaje natural es una rama de la ciencia que pertenece a la intersección de la lingüística aplicada y las ciencias de la computación, que estudia los métodos necesarios para que la computadora pueda ejecutar varias tareas relacionadas con el lenguaje humano, como el español, y requiere cierto grado de «entendimiento» de su contenido. Por otro lado esta ciencia desarrolla las herramientas que ayudan al lingüista en su trabajo cotidiano e incluso pueden llevar a descubrimientos lingüísticos nuevos".*

2.2 Generación automática de resúmenes

Los resúmenes se pueden clasificar por su propósito de acuerdo a su entrada y por su estrategia de condensación como se muestra en la figura [Alfonseca 03], ver Figura 1.

Los resúmenes se clasifican:

Por su propósito:

Genéricos: son para uso general, no tienen ningún propósito en específico.

Orientados al usuario: son los que se tienen que adaptar a las necesidades de un usuario, el usuario puede hacer resúmenes de temas específicos.

De acuerdo a su entrada:

Un solo documento: se tiene como entrada un solo documento.

Multiples documentos: se tiene como entrada más de un documento.

Multimedia: se puede tener como entrada imágenes, grabaciones de audio.

Según su estrategia de condensación:

Abstractivo: en este resumen no se limitan a extraer fragmentos relevantes del texto si no que se hace un análisis más profundo para poder comprenderlo y poder generar un nuevo resumen.

Extractivo: en este tipo de resumen las oraciones que forman parte del contenido se extrae del texto fuente, se hace un análisis superficial.

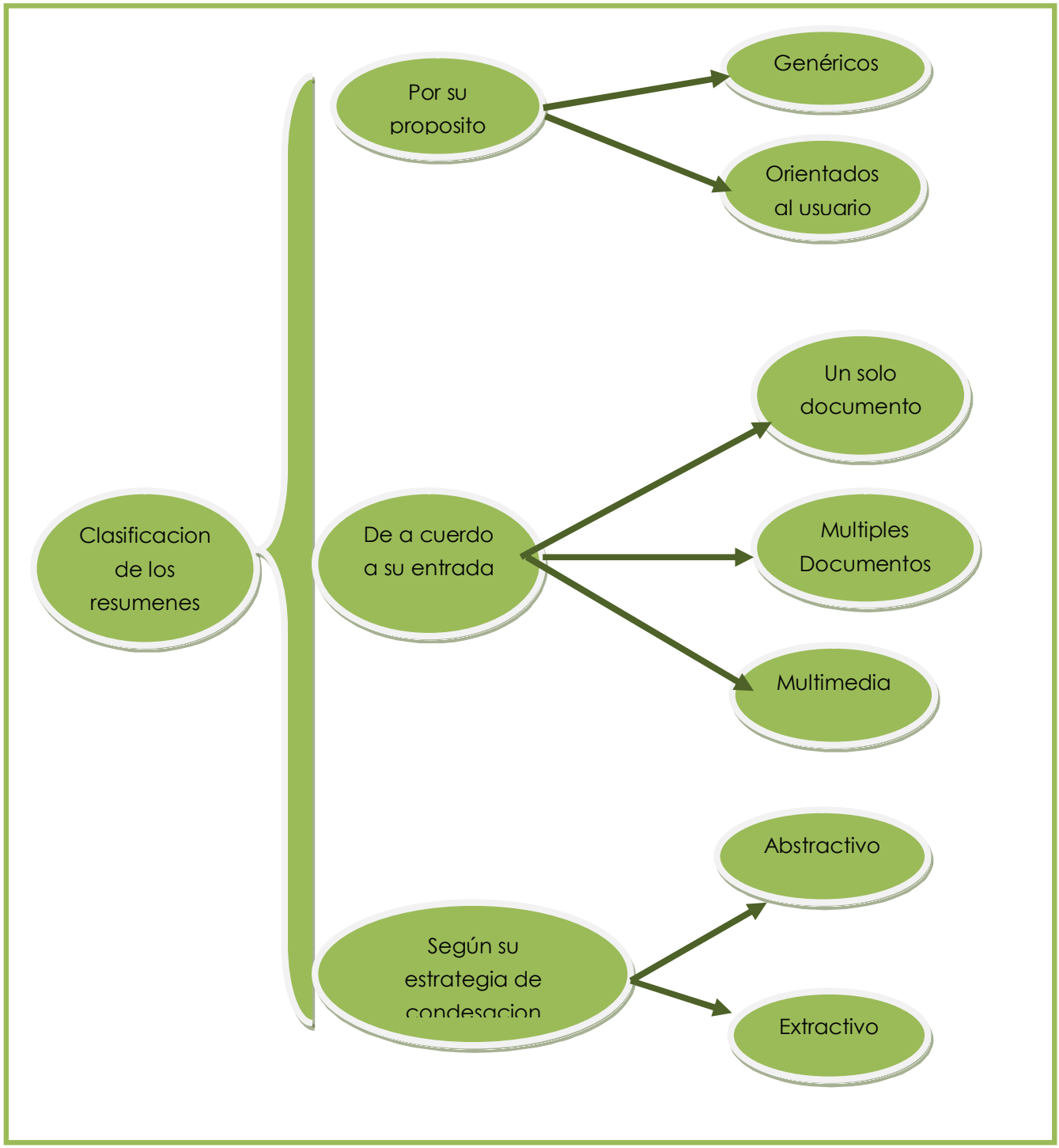

Figura 1. Clasificación de resúmenes (Alfonseca 03).

2.3 Evaluación de resúmenes para la generación automática de resúmenes

Todos los trabajos del estado de arte manejan la herramienta ROUGE (Lin, 2004). Para poder comparar los resultados de este trabajo se utilizará la herramienta ROUGE.

ROUGE proporciona tres medidas:

- ✓ Precisión
- ✓ Recuerdo
- ✓ F-medida

Estas medidas se calculan de la siguiente manera:

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

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

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

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

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

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

2.4 Heurística en la generación automática de resúmenes

2.4.1 Baseline

Baseline es una heurística utilizada en las tareas de la generación automática de resúmenes. Se considera que en el corpus de noticias la información más importante se encuentra al inicio. De esta manera, se extraen las primeras 100 palabras para calcular el baseline para poder considerarlo en las comparaciones de las herramientas comerciales y los métodos del estado de arte.

Para el idioma inglés, el baseline para las primeras 100 palabras fue calculado en el trabajo de (Ledeneva 2008b).

CAPÍTULO 3

Estado del Arte

En este capítulo, se presenta el estado del arte para la generación de resúmenes automáticos. Se describen los trabajos más importantes. Se hizo la revisión de los trabajos para la generación automática de resúmenes que fueron probados para los idiomas inglés y portugués, y básicamente se basó en la importancia de estos trabajos dentro del estado de arte para presentalos en este capítulo.

3.1 Generación automática de resúmenes independientes del lenguaje (Matias 2016)

En el trabajo de la generación automática de resúmenes independientes del lenguaje (Matias 2016), se propone el algoritmo genético para las colecciones en tres idiomas: inglés, portugués e español. La aportación más importante de este trabajo es la compilación del corpus de los Textos en Español para Resúmenes (TER). Esta colección de documentos contiene 240 noticias en el idioma español. El corpus TER es de noticias periodísticas adquiridos del periódico mexicano Crónica, sobre 12 diferentes categorías, academia, bienestar, ciudad, cultura, deportes, espectáculos, estados, mundo, nacional, negocios, opinión y sociedad. Para cada documento de la colección se crearon dos resúmenes por dos humanos expertos.

Este trabajo solamente compara con los métodos del estado de arte y no compara con las herramientas comerciales disponibles. Entonces surge la pregunta: Cuál es la calidad de los resúmenes generados por las herramientas comerciales para el idioma español?

En la figura 2, se muestran los resultados obtenidos con la colección de documentos TER para el lenguaje español en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones.

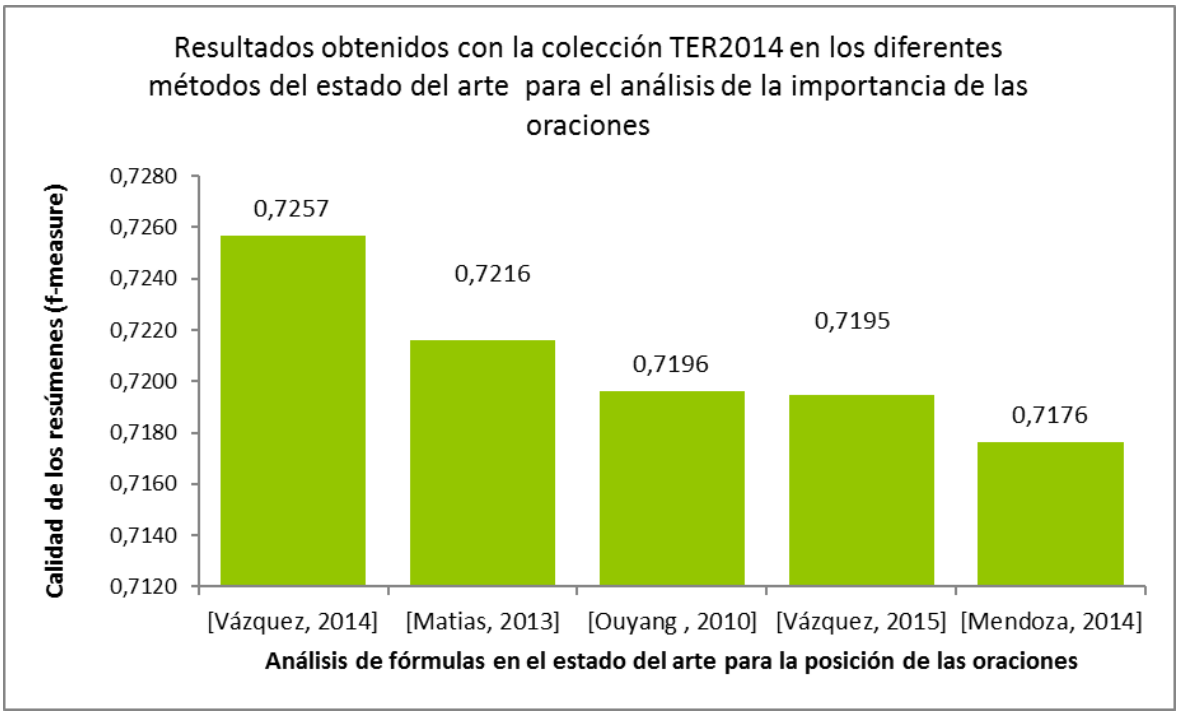

Figura 2. Resultados obtenidos con la colección en el lenguaje español en los diferentes métodos del estado del arte para el análisis de la importancia de las oraciones (Matias 2016).

Como se puede observar en la gráfica anterior la fórmula propuesta en el trabajo de (Vázquez, 2015) es la que obtiene mejores resultados. Entonces para poder comparar en este trabajo se utilizará el resultado de F-measure 0.7257.

3.2 Trabajo (Matias 2013)

Se propone un algoritmo genético probando con la collección para el idioma ingles que se llama DUC 2002 (Document Understanding Conference) que es una colección de documentos creada por National Institute of Standars and Technology (NIST) para la generación de resúmenes. El método

propuesto supera las herramientas comerciales y los métodos del estado de arte.

Copernic Summarizer resultó como mejor herramienta comercial. Dentro de las herramientas de Microsoft Word se obtuvo mejores resultados en Microsoft Word 2003 en Windows Vista y Shvoong es la mejor herramienta en línea. (Matías 2013).

3.3 Generación automática de resúmenes usando algoritmos genéticos (Ibañez 2013)

En este trabajo se realizó la generación automática de resúmenes para el idioma portugués. Se experimentó con el corpus que se llama TeMario que contiene artículos periódicos.

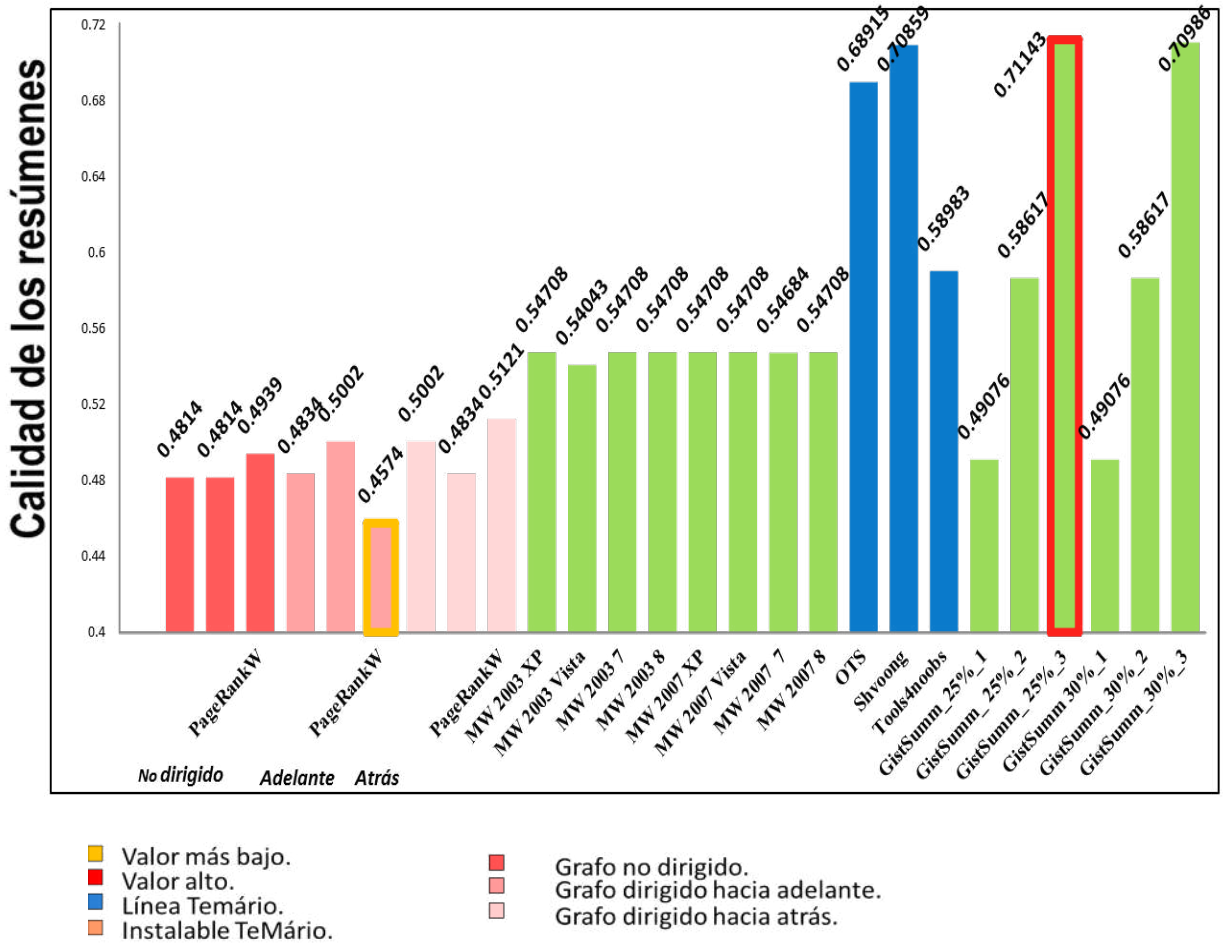

Figura 3. Herramientas comerciales instalables y en línea del trabajo (Ibañez 2013) con los métodos del estado del arte.

3.4 Terminos derivados de Secuencias Frecuentes Maximales (Ledeneva 2008b)

Este trabajo propone utilizar las Secuencias Frecuentes Maximales (SFM) para generación automática de resúmenes. Secuencias Frecuentes Maximales son

secuencias de palabras que se repiten cierto número de veces y que además no están contenidas en otras secuencias frecuentes (García 2006).

La colección de documentos está en el idioma inglés que se llama DUC-2002. Consiste de 567 noticias y tiene dos resúmenes generados para cada texto original. Su longitud es de 100 palabras.

Los mejores resultados se obtienen con el modelo de SFM comparando con el modelo de ngramas y palabras, a continuación en la tabla 1 se muestran los resultados.

Tabla 1. Resúmenes generados con la longitud de 100 palabras (Ledeneva 2008).

Términos	Con palabras claves	Sin palabras claves
Palabras	0.39421	0.41371
ngramas	0.40810	0.42173
SFM	0.43066	0.44085

CAPÍTULO 4

Metodología Propuesta

En este capítulo, se describe la metodología de trabajo propuesta en esta tesis. Se detalla cada una de las etapas que se siguen.

4.1. Metodología de trabajo

En la figura 4, se presenta la metodología de trabajo propuesta a utilizar en esta tesis. La metodología consta de 6 pasos que se describen a detalle en las secciones posteriores.

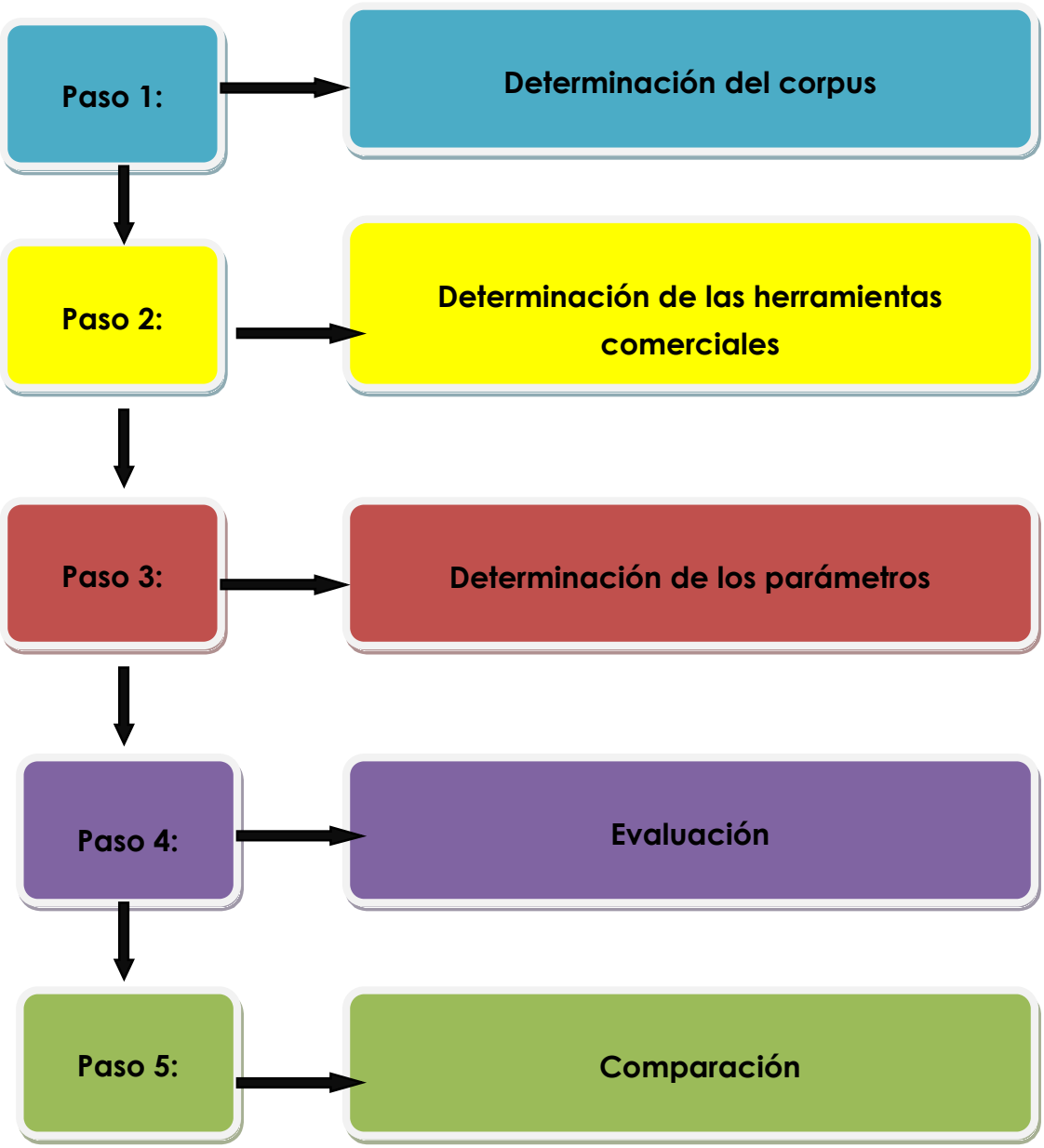

Figura 4. Metodología propuesta.

4.1.1. Determinación del corpus

En este paso se hace la revisión de los corpus disponibles en el idioma español. En esta tesis, el corpus que se va a probar se llama el corpus TER en el idioma español mexicano (Matias 2016). El corpus de describe en la sección 5.1.

4.1.2. Determinación de las herramientas comerciales

En este paso, se hace la revisión del estado de arte de las herramientas comerciales que se encuentra disponibles en la web. Estas herramientas deben de cumplir las siguientes características:

1. La primera característica que deben de cumplir es que sean para la generación automática de resúmenes para un solo documento.
2. La segunda característica que deben de cumplir es que hagan los resúmenes en el idioma español.
3. Debido a que el corpus TER tiene resúmenes de gold estándar de 100 palabras, entonces la herramienta comercial debe de ofrecer la opción de generar los resúmenes con la longitud de 100 palabras.

4.1.3. Determinación de los parámetros

En este paso, se determinan los parámetros que ocupan las herramientas comerciales determinadas en el paso anterior, de la sección 4.1.2.

Los pasos generales que se requieren determinar son los siguientes:

1. El idioma para generar los resúmenes.
2. La longitud del resumen generado por la cantidad del resumen o el porcentaje de las palabras a extraer del documento original.
3. El tipo de resumen a generar: un solo documento o varios documentos.

4.1.4. Evaluación

En este paso, se evalúa el desempeño de los resúmenes generados con las medidas utilizadas en el estado de arte: Precisión, Recuerdo y F-measure (en español, F-medida).

4.1.5. Comparación

Finalmente, los resultados se presentan y se comparan a través de F-medida. Se analizan y se redactan las conclusiones.

CAPÍTULO 5

Experimentación

En este capítulo, se aplica la metodología propuesta y se describe cada de los pasos descritos en el capítulo anterior. Se presentan y se comparan los resultados. Se describe el corpus que se utilizó para generar los resúmenes, también se menciona la herramienta utilizada para la evaluación. Principalmente en este capítulo se muestran las gráficas con los resultados de los experimentos para el idioma español.

5.1. Corpus TER

En el trabajo (Matias 2016) se propone y se utiliza el corpus español mexicano que tiene como nombre Texto en Español para Resúmenes (TER).

El corpus TER para resúmenes fue creado a partir del corpus de noticias obtenido del periódico CRÓNICA sobre 12 diferentes categorías y noticias de los años 2012, 2013 y 2014. Para cada documento de la colección se crearon dos resúmenes por dos humanos expertos (Ver Anexo A).

Las noticias seleccionadas tienen la longitud más de 100 palabras, pero son de diferentes longitudes. Como describe (Matias 2016), las noticias seleccionadas fueron noticias de abril del 2015.

Los nombres que tienen los archivos del corpus TER se formaron de una forma especial que considera las siguientes reglas:

1. Se consideró un numero consecutivo (1-20)
2. Se tomaron dos letras para la categoría
3. Se consideró la fecha de la noticia
4. La clave de la noticia

El corpus selecciona de manera aleatoria 20 noticias de cada una de las categorías proporcionada por el periódico, de la siguiente manera:

1. Academia

2. Bienestar
3. Ciudad
4. Cultura
5. Deportes
6. Espectáculos
7. Estados
8. Mundo
9. Nacional
10. Negocios
11. Opinión
12. Sociedad

Se utilizaron las siguientes abreviaciones para formar parte del nombre según los nombres de las categorías como se indica a continuación.

- Academia: AC
- Bienestar: BI
- Ciudad: CI
- Cultura: CU
- Deportes: DE
- Espectáculos: ES
- Estados: ED
- Mundo: MU
- Nacional: NA
- Negocios: NE
- Opinión: OP
- Sociedad: SO

5.2. Determinación de las herramientas comerciales

En este paso, se hizo la revisión del estado de arte de las herramientas comerciales que se encuentra disponibles en la web. Se seleccionaron las siguientes herramientas comerciales:

- ✓ Copernic Summarizer [Copernic 15]
- ✓ Open Text Summarizer [OTS 15]
- ✓ Summarizing [Summarizing 15]
- ✓ Text Compactor [TextCompactor 15]
- ✓ Word 03 Win XP [MOW 2003]
- ✓ Word 07 Win XP [MOW 2007]
- ✓ Word 03 Win 7 [MOW 2003]
- ✓ Word 07 Win 7 [MOW 2007]
- ✓ Word 03 Win 8.1 [MOW 2003]
- ✓ Word 07 Win 8.1 [MOW 2007]

5.3. Determinación de los parámetros

En este paso, se determinaron los parámetros que ocuparon las herramientas comerciales determinadas en el paso anterior, de la sección 5.2, como se describe a continuación:

Para las herramientas comerciales Copernic Summarizer y Summarizing se utilizó el parámetro de la longitud del resumen a generar: 100 palabras.

Las herramientas comerciales donde se tiene que proporcionar como parámetro el porcentaje de las palabras que tendría que tener el resumen a generar, son las siguientes:

- ✓ Open Text Summarizer
- ✓ Text Compactor
- ✓ Microsoft Word 03 Windows XP
- ✓ Microsoft Word 07 Windows XP
- ✓ Microsoft Word 03 Windows 7
- ✓ Microsoft Word 07 Windows 7
- ✓ Microsoft Word 03 Windows 8.1
- ✓ Microsoft Word 07 Windows 8.1.

En las herramientas que utilizaron porcentaje se aplicó la siguiente formula:

$$\% = \left(\frac{\text{Num de palabras}}{\text{Num de palabras totales en el documento}}\right) \times 100$$

5.4. Evaluación

Primero, se presentan los resultados de la evaluación de las herramientas comerciales y posteriormente de describe el método del estado de arte.

5.4.1 Evaluación con las herramientas comerciales

En la Figura 5, se presentan los resultados de la evaluación de las herramientas cometciales de Microsoft Word.

- ✓ Word 03 Win XP
- ✓ Word 07 Win XP
- ✓ Word 03 Win 7
- ✓ Word 07 Win 7
- ✓ Word 03 Win 8.1
- ✓ Word 07 Win 8.1

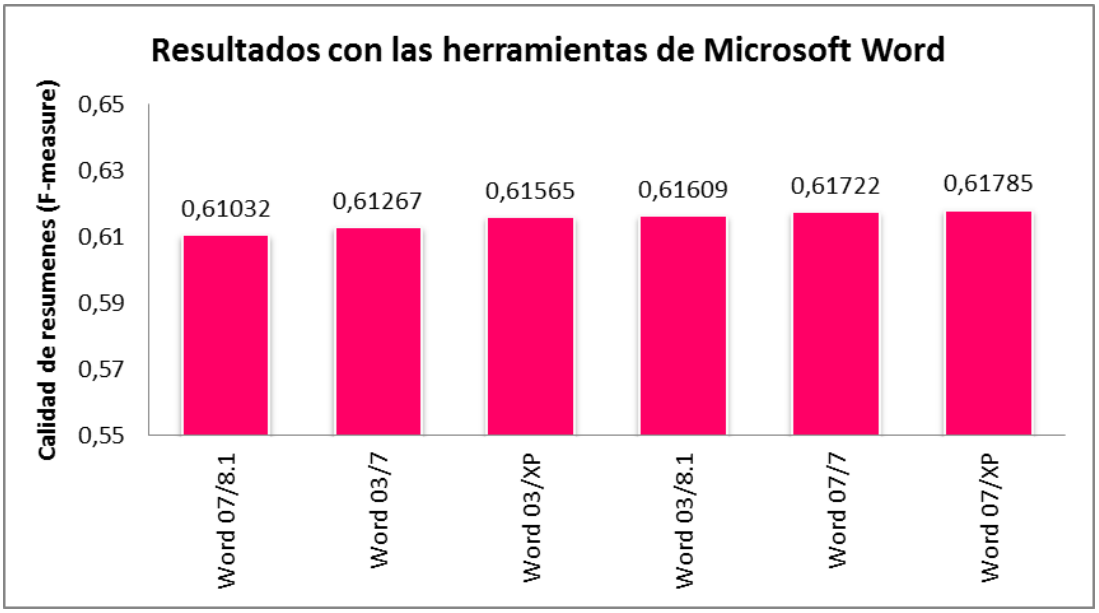

Figura 5. Resultados obtenidos con la colección en el lenguaje español para la herramienta comercial de Microsoft Word.

En la Figura 6, se presentan los resultados de la evaluación de las herramientas comerciales de Microsoft Word.

- ✓ Copernic Summarizer
- ✓ Open Text Summarizer
- ✓ Summarizing
- ✓ Text Compactor

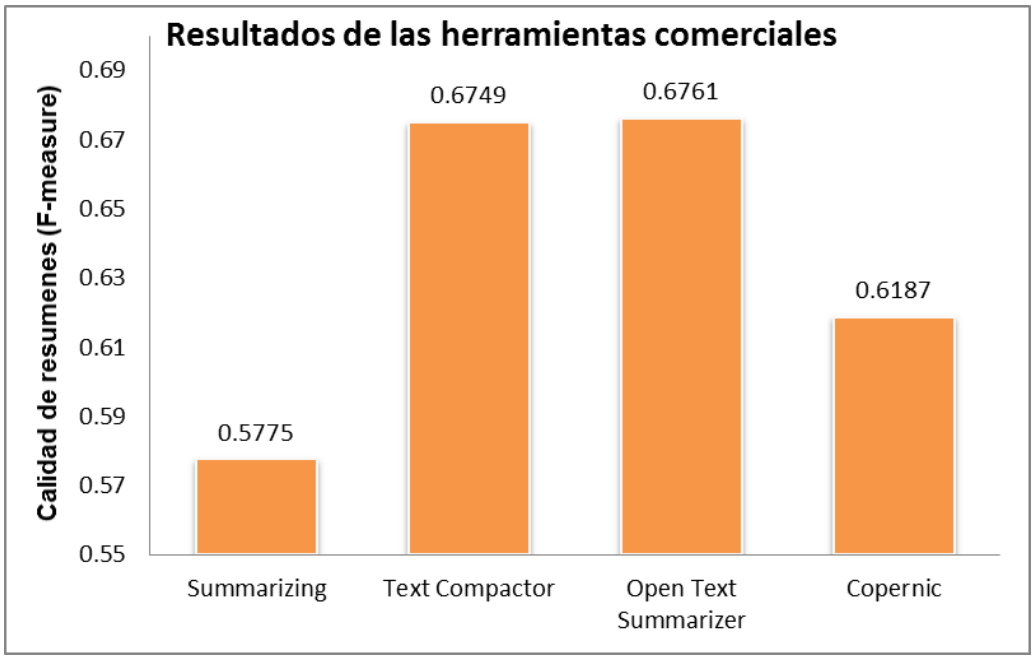

Figura 6. Resultados obtenidos con la colección en el lenguaje español para las herramientas comerciales de Copernic Summarizer, Open Text Summarizer, Summarizing, Text Compactor.

5.4.2 Evaluación con el método del estado del arte

Los mejores resultados obtenidos con la colección en el lenguaje español para el método de estado de arte de (Matias 2016) con diferentes parámetros, son F-measure 0.7257 (Ver Anexo 4).

Los parámetros utilizados en el mejor método propuesto del trabajo (Matias 2016) son como sigue en la tabla 2 a continuación:

Tabla 2. Parámetros para el lenguaje español (TER) para el mejor método del trabajo (Matias 2016).

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

En este paso, se evaluó la calidad de los resúmenes generados a través de las medidas que se utilizan en el estado de arte: Precisión, Recuerdo y F-measure (en español, F-medida). Esto se realizó utilizando la herramienta de evaluación estándar que se llama ROUGE (Lin 2004).

5.5 Comparación con el método del estado del arte y las herramientas comerciales

En la Figura 7, se muestran los resultados de los experimentos con el lenguaje español, con las herramientas comerciales y los métodos del estado del arte. Los experimentos se realizaron utilizando la colección de documentos TER y se evaluaron con la herramienta ROUGE.

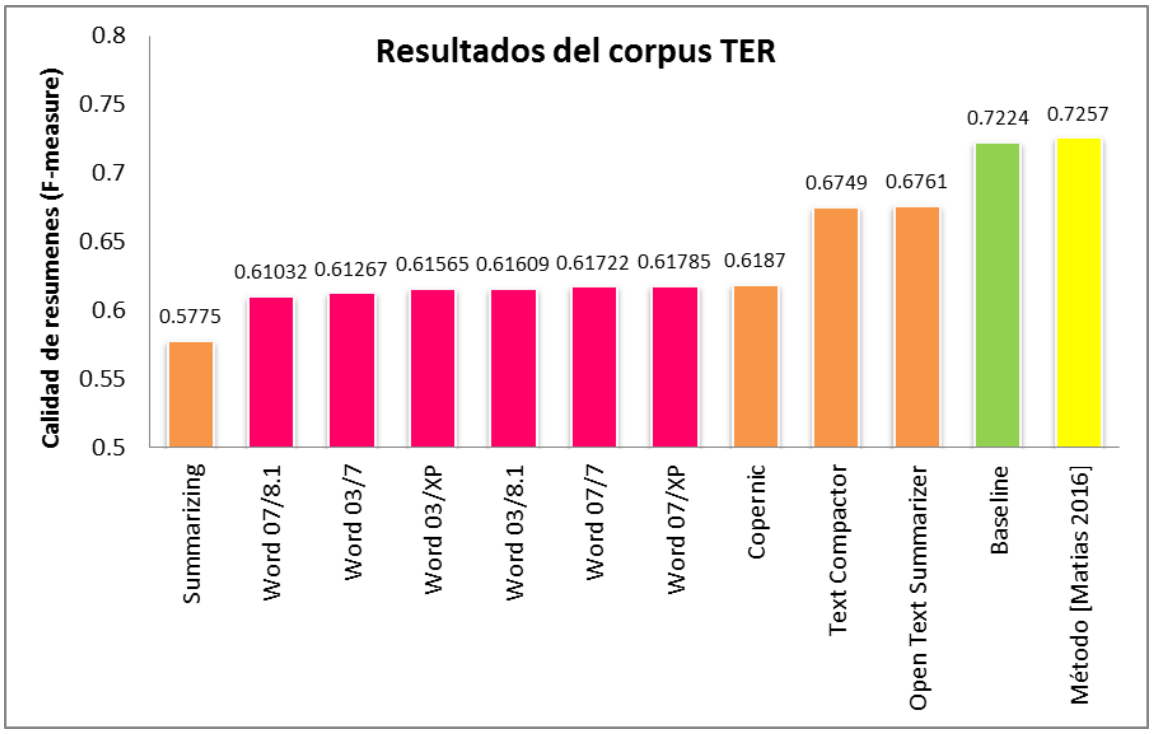

Figura 7. Resultados obtenidos con la colección en el lenguaje español con los parámetros pre-procesamiento y con el modelo de texto.

Como se puede observar el método de (Matias 2016) supera los resultados obtenidos con las herramientas comerciales instalables y en línea.

El baseline de esta colección es muy alto y solamente el método de estado de arte lo supera.

Para los resultados de las medidas Precisión y Recuerdo se pueden consultar en el Anexo 3.

CAPÍTULO 6

Conclusiones y Trabajo Futuro

En este capítulo, se presentan las conclusiones generales del trabajo de tesis. Se describen las conclusiones obtenidas de la evaluación y comparación de las herramientas comerciales con el método del estado de arte. Además se presenta el trabajo futuro.

6.1. Conclusiones

A continuación se presentan las conclusiones de este trabajo:

- ✓ Se comparó con el método propuesto de (Matias 2016) que es independiente de lenguaje, que ya se había probado como uno de los mejores métodos para los idiomas inglés y portugués (Matias 2013, Ibañez 2013).

- ✓ Se compararon con 10 herramientas comerciales que es la aportación de esta tesis.

- ✓ Otra aportación de esta tesis que se probaron las herramientas comerciales para el idioma español.

- ✓ La herramienta con la cual se obtuvo el mejor resultado fue con Open Text Summarizer. Que a partir de esto se puede concluir que esta herramienta comercial generó los resúmenes más parecidos a como los realizaría un humano.

- ✓ Cabe mencionar, que el método de (Matias 2016) superó a la calidad de los resúmenes realizados por las herramientas comerciales (Ver Figura 2).

- ✓ El método de (Matias 2016) superó a la configuración de *baseline* (Ver Figura 2).

- ✓ En el caso de las herramientas de Microsoft Word se probó en sus versiones 2003, 2007 y XP para saber si se obtenían los mismos resultados. De acuerdo a los experimentos realizados, se obtuvo lo siguiente:
  Miscrosoft Office Word 2003 en los sistemas operativos: XP, 7, 8.1 y Miscrosoft Office Word 2007 en los sistemas operativos: XP, 7, 8.1 se obtuvieron los resultados con muy poca diferencia.

- ✓ De las herramientas de Microsoft Word, la mejor herramienta es Microsoft Word 2007/XP.

6.2. Trabajo futuro

A continuación se presentan algunas ideas para el trabajo futuro:

- ✓ Desarrollar un método para generar resúmenes en otros idiomas, basándose en los resultados obtenidos.

- ✓ Probar el enfoque de Secuencias Frecuentes Maximales generando nuevo método para el mismo corpus (García 08, 09, 09a, 13).

- ✓ Probar el enfoque de n-gramas sintácticas generando nuevo método para el mismo corpus.

- ✓ Probar el enfoque de patrones léxicos propuesto en (Hernández 2016).

- ✓ Probar otro corpus en el idioma español para las mismas herramientas comerciales.

Referencias

[Alfonseca 03] Alfonseca, E., & Rodríguez, P. (2003). Generating extracts with genetic algorithms. Springer-Verlag, 2633, 511-519.

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

[DUC 02] Document understanding conference 2002. http://wwwnlpir. nist.gov/projects/duc

(Garcés 2008) Bernardo Garcés Chapero. Curso Resumen automático IA. Presentación de la página web.

http://www.cs.upc.edu/~bejar/ia/material/trabajos/Resumen_Automatico1.pdf

[García 06] García-Hernández, R. A., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2006). A new algorithm for fast discovery of maximal sequential patterns in a document collection. In Computational Linguistics and Intelligent Text Processing (pp. 514-523). Springer Berlin Heidelberg.

[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.

[García 13] García Hernández René Arnulfo, Ledeneva Yulia, Matias Mendoza Griselda, Grigori Sidorov, "Evaluación de las herramientas comerciales y métodos del estado del arte para la generación de resúmenes extractivos individuales", Research in Computing Science 70, p.p.265-274, ISBN 978331903673-1, Año 2013.

[Gelbukh 10] Alexander Gelbukh, 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.

[Gelbukh 14] Centro de Investigación en Computación, Instituto Politécnico Nacional, México. International Journal of Computational Linguistics and Applications, 5(1), 8.

[Hernández 16] Yanet Hernández 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.

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

[Ledeneva 08b] Yulia Ledeneva, 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.

[Ledeneva 11] Yulia Nikolaevna Ledeneva, 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)

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

[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.

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

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

[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.

[Summarizing 15] Summarizing.biz herramienta commercial en línea para realizar resúmenes automáticos en línea, fecha de consulta 25 de diciembre, http://www.summarizing.biz/

[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/.

[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.

Anexo 1. Ejemplo de la estructura del corpus TER en el idioma español

Texto original

En la figura A1, se muestran las carpetas en donde se encuentra el Corpus TER. El corpus está dividido por tres subcarpetas:

- ✓ Separado por oraciones
- ✓ Texto
- ✓ Texto-titulo

Para hacer el resumen con las herramientas comerciales utilizaremos la carpeta de texto.

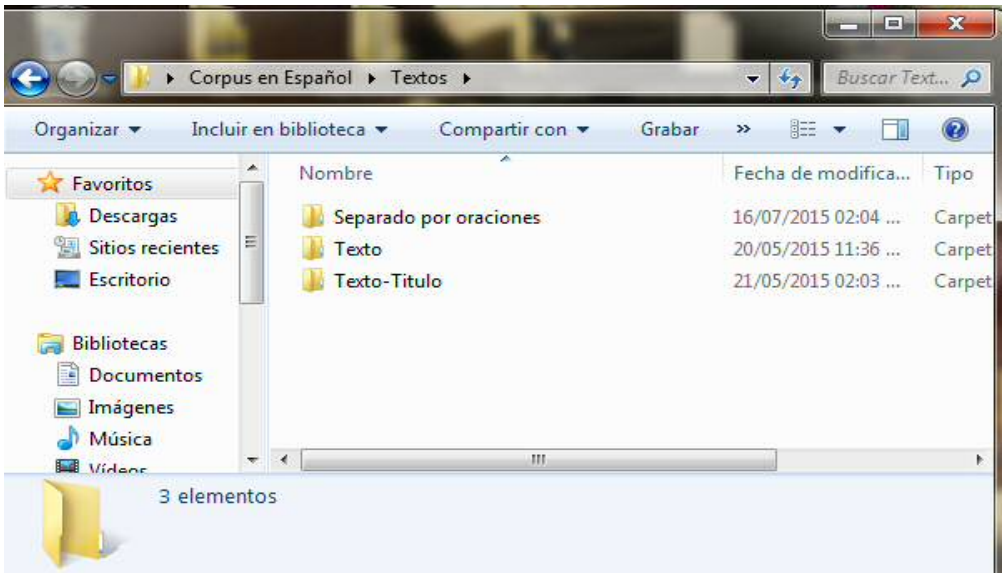

Figura A8. Carpeta donde se encuentra el texto original.

En la figura A2, se muestran las 12 carpetas con 240 archivos en total que tiene el Corpus TER para realizar los resumenes en las diferentes herramientas comerciales, tambien se presenta un ejemplo de una noticia.

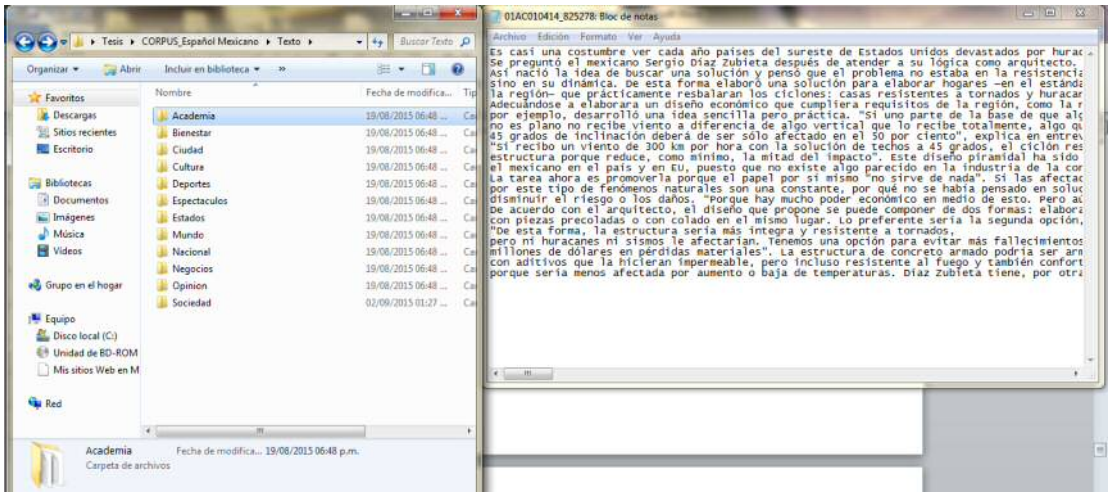

Figura A9. Carpeta de textos originales y un ejemplo del texto original.

Resumen generado por el humano

En la figura A3, se muestra el resumen que generó el humano para la evaluación de cada noticia.

C:\Users\NANCY\Desktop\Corpus en Español\Resumen del experto\Resumenes_sin titulo\Academia

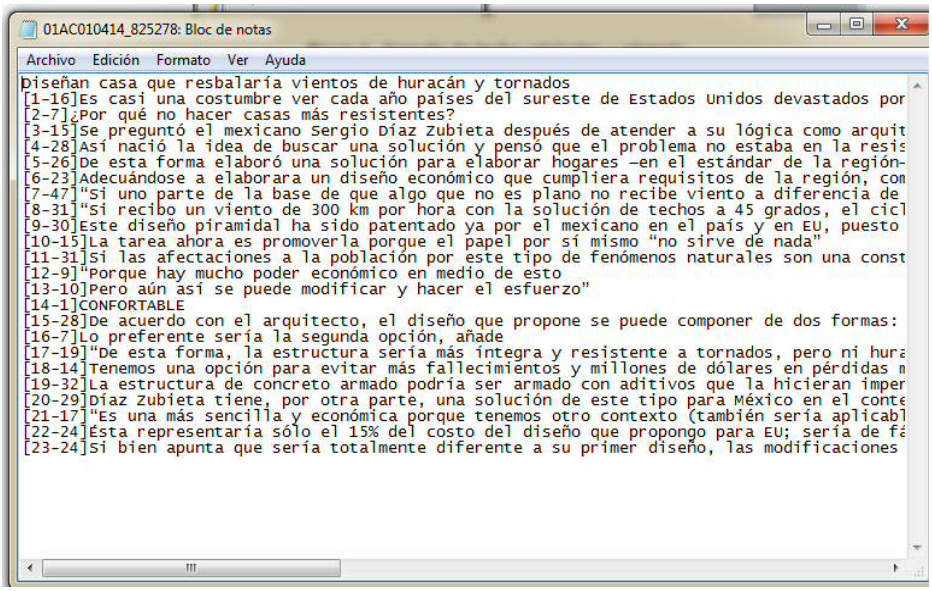

Figura A10. Carpeta de textos generados por el humano.

Resumen generado por la herramienta Copernic Summarizer

En la figura A4, se presenta el resumen de 100 palabras generado con la herramienta Copernic.

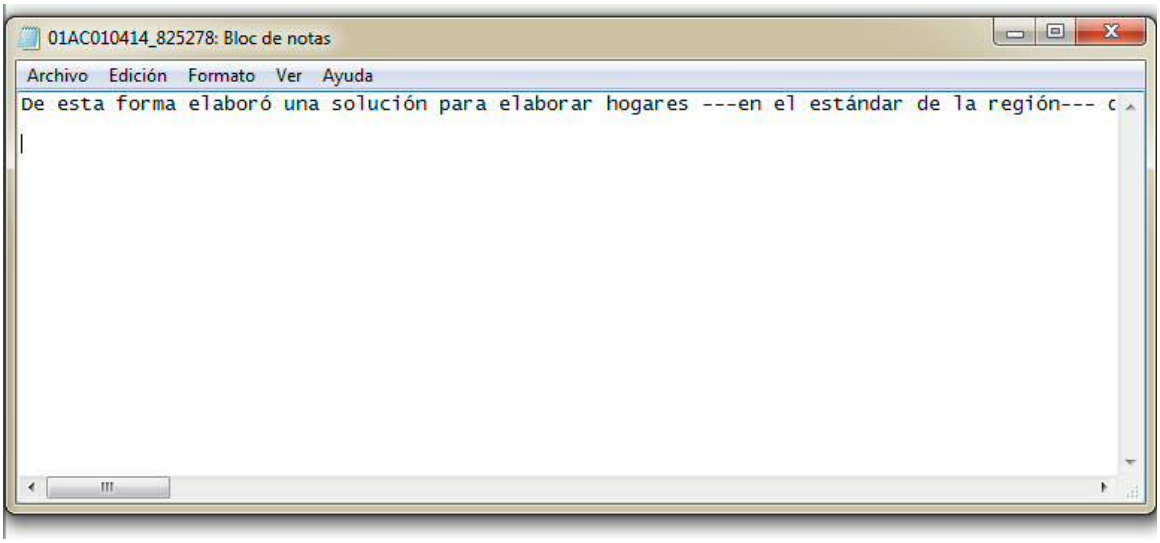

Figura A11. Resumen generado por Copernic Summarizer.

Anexo 2. Descripción de las herramientas comerciales para la generación de resúmenes automáticos

Para la generación de resúmenes automáticos se utilizaron diferentes herramientas comerciales como son:

1. Copernic Summarizer
2. Text Compactor
3. Open Text Summarizer
4. Summarizing
5. Microsoft Office Word 2003 (Sistemas operativos Windows XP, Windows 7 Home Premium y Windows 8.1)
6. Microsoft Office Word 2007 (Sistemas operativos Windows XP, Windows 7 Home Premium y Windows 8.1)

Copernic Summarizer

A continuación se muestra la forma para realizar un resumen con la herramienta Copernic Summarizer, en el sistema operativo Windows 7 Home Premium.

1. Se pega el texto a resumir.
2. Posteriormente, se selecciona la opción de 100 palabras para resumir como lo muestra la figura y automáticamente el texto se queda en 100 palabras.

Algoritmo utilizado para realizar resúmenes con la herramienta Copernic Summarizer es el siguiente:

Copernic Summarizer lee un texto irrelevante y de contenido. Copernic Summarizer se centra en los elementos de texto esenciales, lo que se traduce incluso en los sumarios más relevantes. Una vez que los resúmenes han sido generados, se puede imprimir, guardar y enviar la información.

La herramienta de Copernic se descarga desde su página principal: http://www.copernic.com/en/products/summarizer/

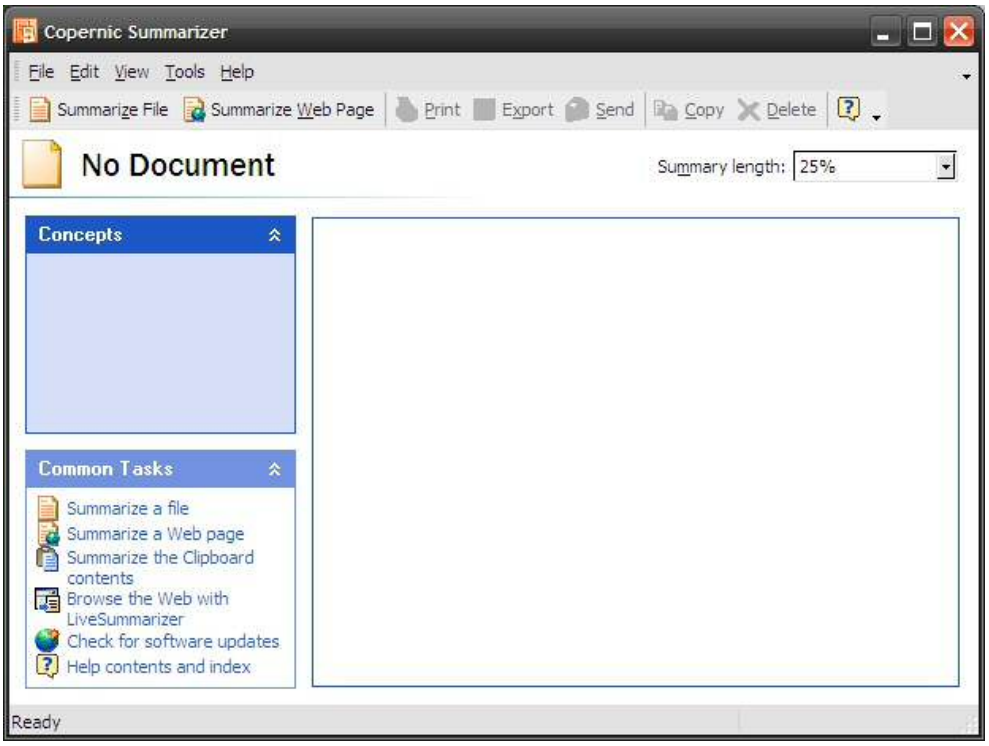

Figura A12. Herramienta en línea Text Compactor.

Text Compactor

A continuación se muestra la forma de realizar un resumen con la herramienta Text Compactor, en el sistema operativo Windows 7 Home Premium (Ver figura A5).

1. Se escribe o se pega el texto en el recuadro.
2. Posteriormente, se elige el porcentaje de texto que quieres que mantenga el resumen como lo muestra la figura.
3. Al asignar el porcentaje automáticamente realiza el resumen que aparece en el recuadro inferior.

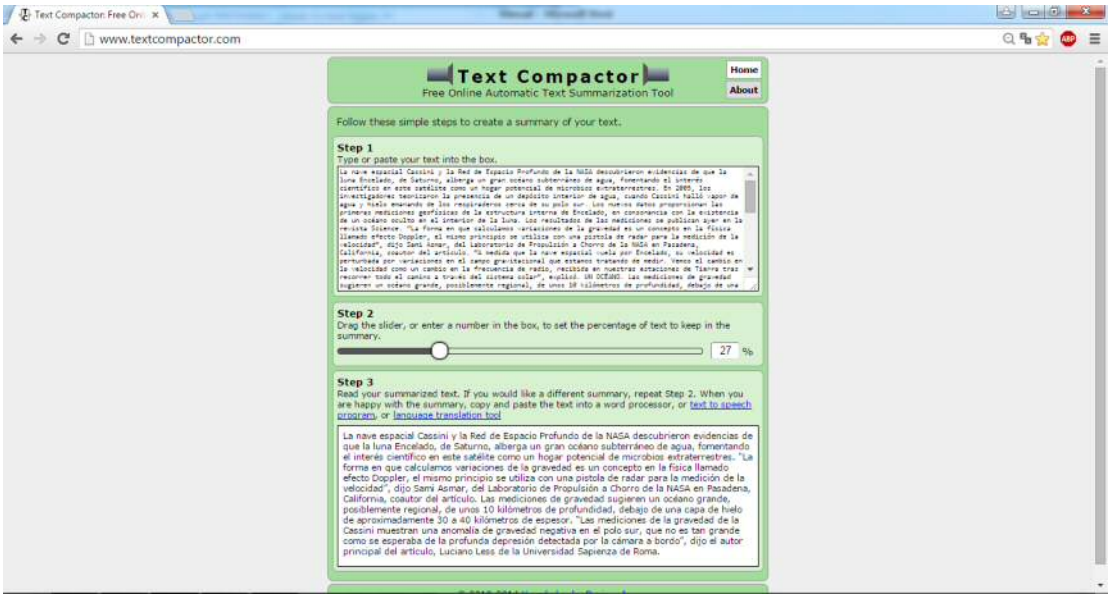

Figura A13. Herramienta en línea Text Compactor.

Algoritmo utilizado por el Text Compactor:

Determina primero cuales son las palabras clave del documento por medio de las veces que aparece repetidamente y después analiza cuales de las frases están más presentes. Esto tiene la finalidad de que el resumen solo contenga párrafos relacionados con el concepto general del texto.

Donde podemos encontrar la herramienta en línea es:
http://www.textcompactor.com/

Open Text Summarizer (OTS)

A continuación se muestra la forma de realizar un resumen con la herramienta OTS, en el sistema operativo Windows 7 Home Premium (Ver Figura A6).

1. Se pega el texto a resumir en el área de trabajo de OTS.
2. Posteriormente, se colocan las opciones para resumir en este caso el tipo de porcentaje que se requiere como lo muestra la figura A6.
3. Después se da clic en el botón donde dice: "Enviar".

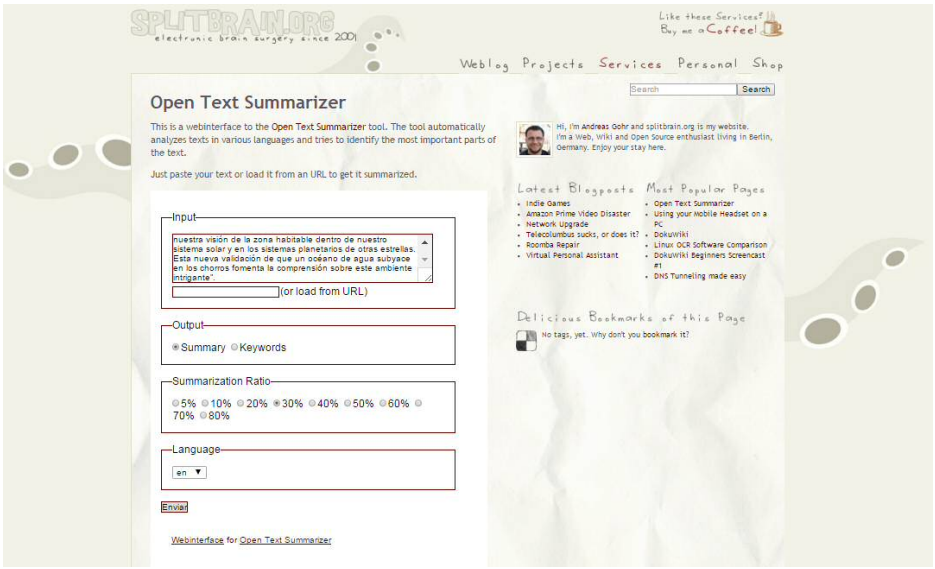

Figura A14. Herramienta en línea Open Text Summarizer.

Algoritmo utilizado por la herramienta OTS es el siguiente:

OTS lee un texto y decide que oraciones son más importantes y cuáles no. El proceso lo hace de la siguiente manera:

a) OTS elimina palabras comunes, como artículos como "el" o "a" o conjunciones como "y" y "pero".
b) Las frases que tienen el mayor porcentaje de las palabras más frecuentes son las que se utilizan en la salida.

OTS es una herramienta para resumir textos de código abierto.

Esta herramienta se basa en el contenido del documento para generar el resumen automático.

Donde se encuentra la herramienta en línea es en el siguiente link:
http://www.splitbrain.org/services/ots

Summarizing

A continuación se presenta la forma para realizar un resumen con la herramienta en línea Summarizing (Ver Figura A7 y A8).

1. Se pega el texto en el recuadro donde dice "Your Text" de la herramienta Summarizing.
2. Una vez que este el texto en el recuadro de Summary length daremos el número de palabras que requerimos que sea el resumen (en este caso

lo dejaremos con 100 words, porque es el número que queremos que tenga nuestro resumen).

3. Después resolveremos la operación que nos aparece debajo.
4. Daremos clic en Summarize para que nos muestre el resumen.

La herramienta Summarizing no proporciona más información sobre el algoritmo utilizado.

Figura A15. Pantalla principal de la Herramienta en línea Summarizing.

Figura A16. Resumen generado con la herramienta Summarizing.

Microsoft Office Word 2003

A continuación se muestra la forma para realizar un resumen en Microsoft Office Word 2003 (el procedimiento utilizado es el mismo usando el sistema operativo XP, 7 Home Premium, y Windows 8.1), ver figuras A9 y A10.

1.- Se pega el texto a resumir en el área de trabajo de Microsoft Office Word 2003.

2.- Se da clic en la opción de herramientas y después buscaremos la opción de Autorresumen Automático.

3.- Cuando aparece la ventana, vamos a seleccionar "Crear un documento nuevo para colocar el resumen" y colocar el porcentaje necesario para cada texto. En este caso de esta tesis todos los resúmenes deben de tener por lo menos100 palabras.

4.- Para finalizar daremos clic en "Aceptar".

Algoritmo utilizado para realizar resúmenes con la herramienta Microsoft Office Word 2003:

1.- "Autorresumen". Esta función consiste en realizar resúmenes automáticos ya sea en un nuevo documento o en el mismo documento.

2.- Con un porcentaje variado que el usuario debe elegir.

3.- Realiza el resumen de las oraciones principales y lo visualiza de la forma en que el usuario haya elegido.

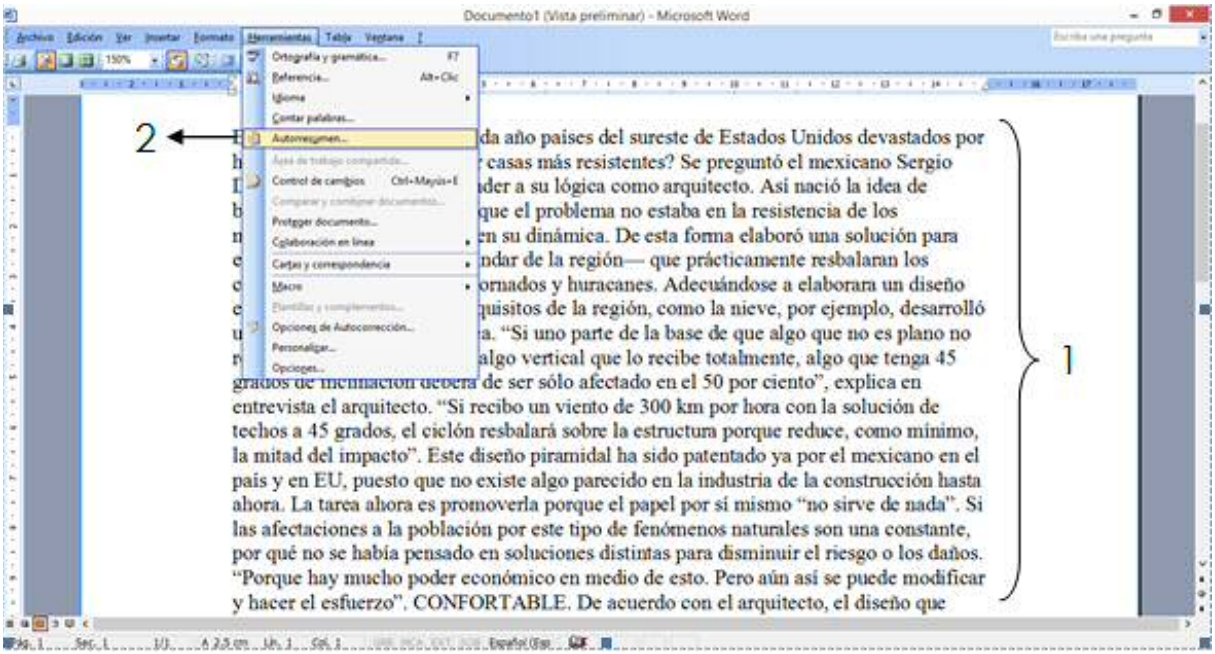

Figura A17. Herramienta instalable Microsoft Word 2003

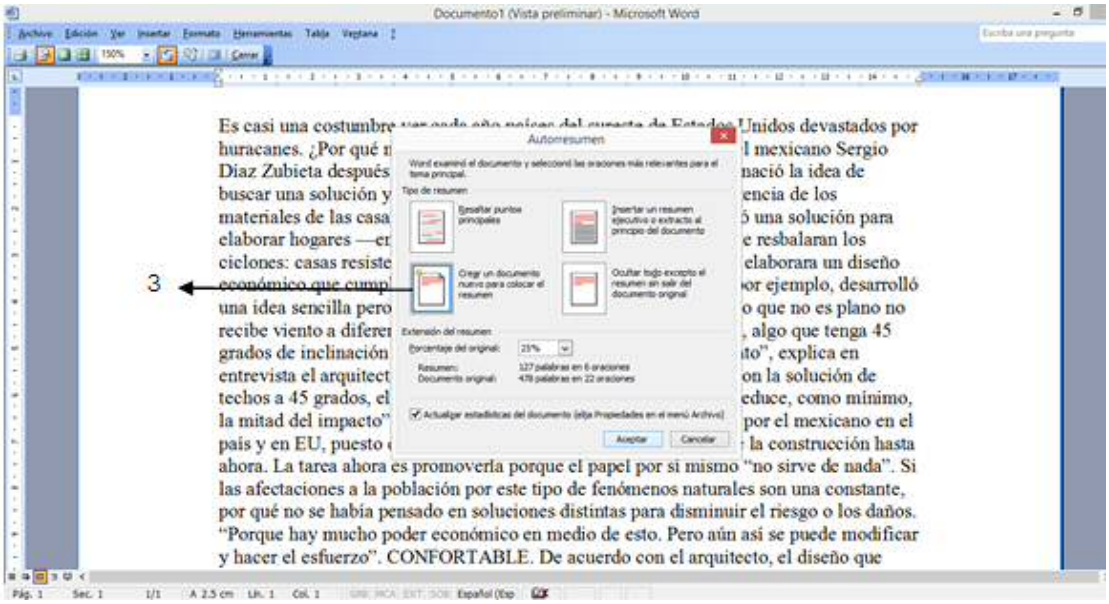

Figura A18. Herramienta instalable Microsoft Word 2003.

Microsoft Office Word 2007

A continuación se muestran los pasos que se deben de seguir para realizar un resumen en Microsoft Office Word 2007, cabe señalar, que el procedimiento no cambia; no importa que se utilicen diferente sistema operativo como: Windows XP, Windows 7 Home Premium y Windows 8.1.

1. Se pega el texto a resumir en el área de trabajo de Microsoft Office Word 2007.
2. Después daremos clic en resumen automático.

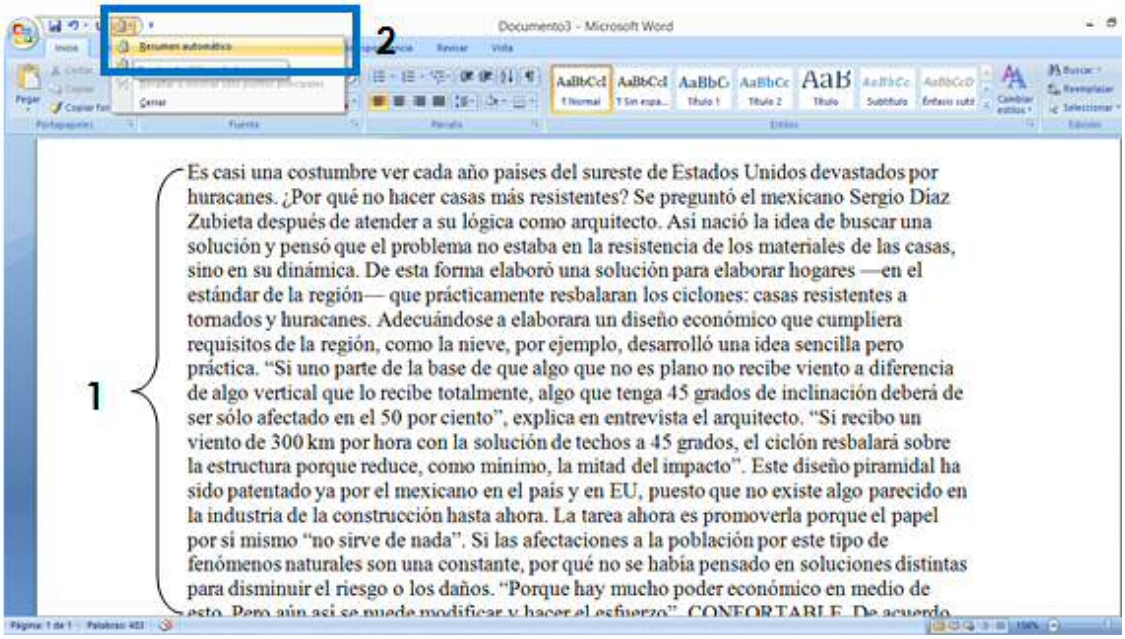

Figura A1913. Herramienta instalable Microsoft Word 2007.

3. Aparecerá la siguiente pantalla en donde seleccionaremos "crear documento nuevo para colocar el resumen" y le daremos el porcentaje según corresponda en cada texto a resumir.

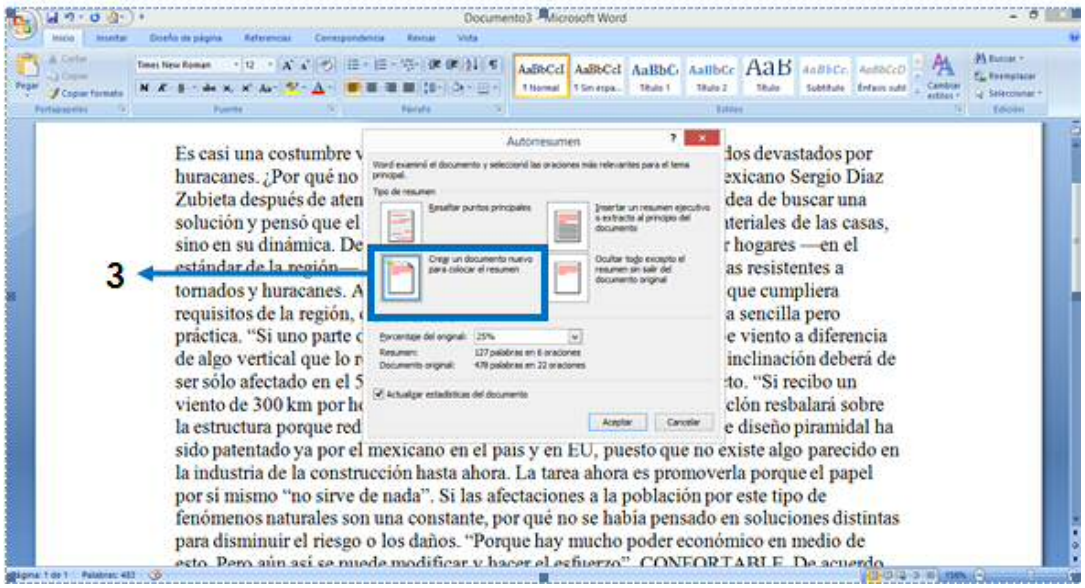

Figura A20. Herramienta instalable Microsoft Word 2007.

4. Por ultimo, daremos clic en "Aceptar".

Algoritmo utilizado para realizar resúmenes con las herramientas Microsoft Office Word 2007:

a) La función de la opción de "Autorresumen" es hacer un resumen automático del contenido de un documento determinando las oraciones principales.
b) Se asigna una puntuación a cada oración.
c) De acuerdo al tipo de resumen que haya elegido el usuario para visualizar el porcentaje de oraciones con puntuación más alta se mostrará en el resumen.

Para Microsoft Office Word 2007, se utilizarán las mismas opciones que en la versión de Microsoft Office 2003. Esta herramienta se basa en el contenido del documento para generar el resumen automático.

Anexo 3. Evaluación de herramientas comerciales con ROUGE para el corpus TER

Resultados de la evaluación de Copernic Summarizer

```
1 ROUGE-1 Average_R: 0.61909 (95%-conf.int. 0.59401 - 0.64338)
1 ROUGE-1 Average_P: 0.61859 (95%-conf.int. 0.59341 - 0.64268)
1 ROUGE-1 Average_F: 0.61872 (95%-conf.int. 0.59367 - 0.64314)
---------------------------------------------
1 ROUGE-2 Average_R: 0.47136 (95%-conf.int. 0.43592 - 0.50552)
1 ROUGE-2 Average_P: 0.47106 (95%-conf.int. 0.43630 - 0.50529)
1 ROUGE-2 Average_F: 0.47114 (95%-conf.int. 0.43604 - 0.50553)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.49010 (95%-conf.int. 0.45698 - 0.52175)
1 ROUGE-SU4 Average_P: 0.48977 (95%-conf.int. 0.45703 - 0.52162)
1 ROUGE-SU4 Average_F: 0.48985 (95%-conf.int. 0.45702 - 0.52182)
```

Resultados de la evaluación de Summarizing

```
1 ROUGE-1 Average_R: 0.57623 (95%-conf.int. 0.55259 - 0.59934)
1 ROUGE-1 Average_P: 0.57917 (95%-conf.int. 0.55519 - 0.60200)
1 ROUGE-1 Average_F: 0.57759 (95%-conf.int. 0.55418 - 0.60069)
---------------------------------------------
1 ROUGE-2 Average_R: 0.40915 (95%-conf.int. 0.37603 - 0.44249)
1 ROUGE-2 Average_P: 0.41073 (95%-conf.int. 0.37769 - 0.44474)
1 ROUGE-2 Average_F: 0.40988 (95%-conf.int. 0.37660 - 0.44353)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.42817 (95%-conf.int. 0.39695 - 0.45969)
1 ROUGE-SU4 Average_P: 0.43001 (95%-conf.int. 0.39869 - 0.46166)
1 ROUGE-SU4 Average_F: 0.42902 (95%-conf.int. 0.39771 - 0.46066)
```

Resultados de la evaluación de Open Text Summarizer

```
1 ROUGE-1 Average_R: 0.67631 (95%-conf.int. 0.65344 - 0.69911)
1 ROUGE-1 Average_P: 0.67608 (95%-conf.int. 0.65332 - 0.69856)
1 ROUGE-1 Average_F: 0.67610 (95%-conf.int. 0.65339 - 0.69876)
---------------------------------------------
1 ROUGE-2 Average_R: 0.55634 (95%-conf.int. 0.52391 - 0.58870)
1 ROUGE-2 Average_P: 0.55619 (95%-conf.int. 0.52398 - 0.58799)
1 ROUGE-2 Average_F: 0.55620 (95%-conf.int. 0.52386 - 0.58842)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.56999 (95%-conf.int. 0.53946 - 0.59955)
1 ROUGE-SU4 Average_P: 0.56981 (95%-conf.int. 0.53912 - 0.59939)
1 ROUGE-SU4 Average_F: 0.56982 (95%-conf.int. 0.53925 - 0.59958)
```

Resultados de la evaluación de Text Compactor

```
1 ROUGE-1 Average_R: 0.67509 (95%-conf.int. 0.65176 - 0.69750)
1 ROUGE-1 Average_P: 0.67496 (95%-conf.int. 0.65145 - 0.69730)
1 ROUGE-1 Average_F: 0.67493 (95%-conf.int. 0.65130 - 0.69747)
---------------------------------------------
1 ROUGE-2 Average_R: 0.55389 (95%-conf.int. 0.52067 - 0.58530)
1 ROUGE-2 Average_P: 0.55373 (95%-conf.int. 0.52052 - 0.58535)
1 ROUGE-2 Average_F: 0.55374 (95%-conf.int. 0.52072 - 0.58530)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.56800 (95%-conf.int. 0.53609 - 0.59760)
1 ROUGE-SU4 Average_P: 0.56783 (95%-conf.int. 0.53644 - 0.59768)
1 ROUGE-SU4 Average_F: 0.56784 (95%-conf.int. 0.53621 - 0.59729)
```

Resultados de la evaluación de Microsoft Word 2003 Windows 8.1

```
1 ROUGE-1 Average_R: 0.61744 (95%-conf.int. 0.59620 - 0.64137)
1 ROUGE-1 Average_P: 0.61498 (95%-conf.int. 0.59356 - 0.63860)
1 ROUGE-1 Average_F: 0.61609 (95%-conf.int. 0.59485 - 0.63978)
---------------------------------------------
1 ROUGE-2 Average_R: 0.46585 (95%-conf.int. 0.43500 - 0.50043)
1 ROUGE-2 Average_P: 0.46428 (95%-conf.int. 0.43308 - 0.49804)
1 ROUGE-2 Average_F: 0.46499 (95%-conf.int. 0.43403 - 0.49917)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.48286 (95%-conf.int. 0.45377 - 0.51614)
1 ROUGE-SU4 Average_P: 0.48121 (95%-conf.int. 0.45199 - 0.51318)
1 ROUGE-SU4 Average_F: 0.48195 (95%-conf.int. 0.45274 - 0.51445)
```

Resultados de la evaluación de Microsoft Word 2003 Windows XP

```
1 ROUGE-1 Average_R: 0.61703 (95%-conf.int. 0.59547 - 0.64069)
1 ROUGE-1 Average_P: 0.61453 (95%-conf.int. 0.59272 - 0.63748)
1 ROUGE-1 Average_F: 0.61565 (95%-conf.int. 0.59403 - 0.63864)
---------------------------------------------
1 ROUGE-2 Average_R: 0.46761 (95%-conf.int. 0.43754 - 0.50229)
1 ROUGE-2 Average_P: 0.46588 (95%-conf.int. 0.43615 - 0.50046)
1 ROUGE-2 Average_F: 0.46666 (95%-conf.int. 0.43717 - 0.50131)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.48447 (95%-conf.int. 0.45540 - 0.51687)
1 ROUGE-SU4 Average_P: 0.48267 (95%-conf.int. 0.45377 - 0.51497)
1 ROUGE-SU4 Average_F: 0.48348 (95%-conf.int. 0.45460 - 0.51584)
```

Resultados de la evaluación de Microsoft Word 2003 Windows 7

```
1 ROUGE-1 Average_R: 0.61406 (95%-conf.int. 0.59301 - 0.63669)
1 ROUGE-1 Average_P: 0.61154 (95%-conf.int. 0.58992 - 0.63403)
1 ROUGE-1 Average_F: 0.61267 (95%-conf.int. 0.59174 - 0.63495)
---------------------------------------------
1 ROUGE-2 Average_R: 0.46257 (95%-conf.int. 0.43074 - 0.49462)
1 ROUGE-2 Average_P: 0.46094 (95%-conf.int. 0.42930 - 0.49260)
1 ROUGE-2 Average_F: 0.46168 (95%-conf.int. 0.42993 - 0.49354)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.47980 (95%-conf.int. 0.44958 - 0.50988)
1 ROUGE-SU4 Average_P: 0.47807 (95%-conf.int. 0.44816 - 0.50808)
1 ROUGE-SU4 Average_F: 0.47885 (95%-conf.int. 0.44874 - 0.50863)
```

Resultados de la evaluación de Microsoft Word 2007 Windows 7

```
1 ROUGE-1 Average_R: 0.61885 (95%-conf.int. 0.59754 - 0.64319)
1 ROUGE-1 Average_P: 0.61716 (95%-conf.int. 0.59625 - 0.64036)
1 ROUGE-1 Average_F: 0.61785 (95%-conf.int. 0.59670 - 0.64152)
---------------------------------------------
1 ROUGE-2 Average_R: 0.46978 (95%-conf.int. 0.43807 - 0.50426)
1 ROUGE-2 Average_P: 0.46881 (95%-conf.int. 0.43717 - 0.50271)
1 ROUGE-2 Average_F: 0.46919 (95%-conf.int. 0.43751 - 0.50327)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.48621 (95%-conf.int. 0.45677 - 0.51901)
1 ROUGE-SU4 Average_P: 0.48510 (95%-conf.int. 0.45569 - 0.51762)
1 ROUGE-SU4 Average_F: 0.48554 (95%-conf.int. 0.45607 - 0.51823)
```

Resultados de la evaluación de Microsoft Word 2007 Windows 8.1

```
1 ROUGE-1 Average_R: 0.61165 (95%-conf.int. 0.58907 - 0.63535)
1 ROUGE-1 Average_P: 0.60925 (95%-conf.int. 0.58639 - 0.63279)
1 ROUGE-1 Average_F: 0.61032 (95%-conf.int. 0.58754 - 0.63397)
---------------------------------------------
1 ROUGE-2 Average_R: 0.45898 (95%-conf.int. 0.42585 - 0.49226)
1 ROUGE-2 Average_P: 0.45750 (95%-conf.int. 0.42447 - 0.49074)
1 ROUGE-2 Average_F: 0.45816 (95%-conf.int. 0.42485 - 0.49133)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.47638 (95%-conf.int. 0.44536 - 0.50843)
1 ROUGE-SU4 Average_P: 0.47481 (95%-conf.int. 0.44325 - 0.50656)
1 ROUGE-SU4 Average_F: 0.47551 (95%-conf.int. 0.44396 - 0.50741)
```

Resultados de la evaluación de Microsoft Word 2007 Windows XP

```
1 ROUGE-1 Average_R: 0.61861 (95%-conf.int. 0.59673 - 0.64232)
1 ROUGE-1 Average_P: 0.61610 (95%-conf.int. 0.59463 - 0.63942)
1 ROUGE-1 Average_F: 0.61722 (95%-conf.int. 0.59543 - 0.64075)
---------------------------------------------
1 ROUGE-2 Average_R: 0.46918 (95%-conf.int. 0.43666 - 0.50392)
1 ROUGE-2 Average_P: 0.46747 (95%-conf.int. 0.43462 - 0.50183)
1 ROUGE-2 Average_F: 0.46825 (95%-conf.int. 0.43557 - 0.50274)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.48591 (95%-conf.int. 0.45486 - 0.51849)
1 ROUGE-SU4 Average_P: 0.48412 (95%-conf.int. 0.45317 - 0.51674)
1 ROUGE-SU4 Average_F: 0.48493 (95%-conf.int. 0.45389 - 0.51767)
```

Anexo 4. Evaluación de método del estado de arte con ROUGE para el corpus TER

En la tabla A3, se muestra la lista de parámetros modificados y el mejor valor obtenido de cada uno de ellos para el lenguaje español.

Tabla A3. Parámetros para el lenguaje español (TER) (Matias 2016).

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

Resultados de la evaluación del método (Matias 2016)

```
1 ROUGE-1 Average_R: 0.72625 (95%-conf.int. 0.70427 - 0.74725)
1 ROUGE-1 Average_P: 0.72522 (95%-conf.int. 0.70366 - 0.74600)
1 ROUGE-1 Average_F: 0.72565 (95%-conf.int. 0.70394 - 0.74626)
---------------------------------------------
1 ROUGE-2 Average_R: 0.62570 (95%-conf.int. 0.59384 - 0.65557)
1 ROUGE-2 Average_P: 0.62460 (95%-conf.int. 0.59259 - 0.65514)
1 ROUGE-2 Average_F: 0.62509 (95%-conf.int. 0.59309 - 0.65530)
---------------------------------------------
1 ROUGE-SU4 Average_R: 0.63485 (95%-conf.int. 0.60446 - 0.66417)
1 ROUGE-SU4 Average_P: 0.63376 (95%-conf.int. 0.60405 - 0.66290)
1 ROUGE-SU4 Average_F: 0.63424 (95%-conf.int. 0.60391 - 0.66334)
```

Anexo 5. Ejemplo del texto original y su resumen del corpus TER

Ejemplo del texto original

(Archivo 06AC030414_825787.txt)

Kaspárov promoverá la educación del ajedrez en Iberoamérica desde México

[1-19]Gary Kaspárov lanzó en México la creación de su Fundación de Ajedrez, la cual será sede para toda Iberoamérica

[2-37]La institución buscará acercar a niños y jóvenes al ajedrez a través de diversos programas de enseñanza para ser empleados en la educación básica, además de organizar torneos, competencias, festivales, talleres, cursos y conferencias, entre otras actividades

[3-35]La Fundación Kaspárov de Ajedrez para Iberoamérica —que tendrá presencia activa en 20 países de la región— será la cuarta iniciada por el ajedrecista, después de constituir otras en Estados Unidos, Asia, Europa y África

[4-36]"México se une Nueva York, Singapur, Bruselas y Johannesburgo, para concluir una red global que promueva la educación del ajedrez en el mundo", dijo ayer en conferencia, previo a la presentación oficial en el Museo Soumaya

[5-43]De acuerdo con el campeón de ajedrez 1985-2000, en la fundación han tenido muchas experiencias en diferentes países que continúan recopilando, pero han comprobado que desarrollar la educación en ajedrez en escuelas primarias mejora dramáticamente su desempeño, principalmente en aquellas de bajos recursos

[6-26]También destacó que además en cada continente es perceptible el incremento del interés en los padres por despertar el gusto por el ajedrez en sus hijos

[7-26]El interés cambia gradualmente la idea que por muchos años predominó en la sociedad, acotó, sobre que el ajedrez era una forma agresiva de deporte intelectual

[8-34]En México, apuntó Kaspárov, previamente había hablado con el secretario de Educación, Emilio Chuayffet, la posibilidad de integrar este tipo de programas a las escuelas del país, lo que ha tenido una respuesta favorable

[9-26]Pero parte del proyecto global de la fundación incluye además un plan integral sobre educación del ajedrez en tres temas clave: educación, tecnología y redes sociales

[10-9]No habría que sorprenderse sobre el último punto, aclara

[11-26]Como deporte, el ajedrez no podría competir con el futbol u otros más populares dentro de la “pantalla grande”, es decir televisores u otros medios convencionales

[12-19]Pero en la “pequeña pantalla” (donde los anteriores ya no son tan atractivos de seguir) tendría una enorme ventaja

[13-16]Esto se refiere a dispositivos móviles donde se puede seguir una partida y aprender sobre ajedrez

[14-36]“Veo el futuro del ajedrez como una enorme red social: no como una estructura vertical que viene de arriba hacia

abajo, sino como una cooperación a nivel regional que también se base en el lenguaje compartido”

[15-16]Eso busca con su fundación, pero sería más efectivo si contara con apoyo de la FIDE

[16-2]JUGADA POLÍTICA

[17-31]El “movimiento” de Kaspárov con su fundación y su red en el orbe tiene además otro objetivo, uno político, en un tablero en el que lleva jugando desde hace mucho tiempo

[18-52]Y es que el personaje que compitió por primera vez contra una computadora y la derrotó —para posteriormente empatar y ser vencido polémicamente por la Deep Blue de IBM— busca el apoyo en el mundo con diversas federaciones nacionales para postularse en la dirección de la Federación Internacional de Ajedrez (FIDE), puntualiza

[19-23]“El enorme reto del ajedrez es la falta de patrocinios y eso es porque la actual administración no ha sido capaz de obtenerla

[20-9]Incluso ha fallado en encontrar licitaciones para los campeonatos”

[21-42]Tras cuestionar la calidad moral y profesional del actual presidente de la federación, Kirsán Iliumzhínov, refirió que en todo el mundo ha visto el mismo problema: el juego no está listo para avanzar a completa velocidad por falta de financiamiento y promoción

[22-14]“El ajedrez tiene un enorme potencial, es un producto que habla por sí mismo”

[23-28]El maestro del ajedrez refirió que la comunidad debe aceptar que han perdido 20 años con la actual administración del FIDE y que no ha contribuido en nada

[24-26]Agregó que es un tema que debe de discutirse públicamente y debatirse en sociedad y "no dejarlo en manos de funcionarios con una mente tan limitada"
[25-12]Kaspárov lleva tiempo planeando su jugada y se allega sigilosamente ganando simpatías
[26-32]Ajedrecista y activista político, el personaje habla sobre la combinación de su habilidad en ambas ante la pregunta de una periodista en el evento en un hotel de la ciudad de México
[27-32]Como jugador de ajedrez, explica, tiene mucha presión lo que ha sido útil para sus estimaciones políticas, le ha permitido ver una macroimagen, la película completa sobre la complejidad de un problema
[28-11]"Ver el efecto de una acción en ambos lados del tablero"
[29-26]Estamos en un mundo de "microgestión" y vemos muchos políticos separar problemas y manejarlos uno a uno, pero cuando solucionamos uno puede tener impacto los demás
[30-29]"Esa es la lección más importantes en el juego del ajedrez, todos los movimientos están conectados y si no haces tus movimientos pensando en cómo afecta un todo perderás"

Ejemplo del resumen generado

(Archivo 06AC030414_825787.txt)

Gary Kaspárov lanzó en México la creación de su Fundación de Ajedrez, la cual será sede para toda Iberoamérica. La institución buscará acercar a niños y jóvenes al ajedrez a través de diversos programas de enseñanza para ser empleados en la educación básica, además de organizar torneos, competencias, festivales, talleres, cursos y conferencias, entre otras actividades. La Fundación Kaspárov de Ajedrez para Iberoamérica —que tendrá presencia activa en 20 países de la región— será la cuarta iniciada por el ajedrecista, después de constituir otras en Estados Unidos, Asia, Europa y África. "México se une Nueva York, Singapur, Bruselas y Johannesburgo, para concluir una red global que promueva la educación del ajedrez en el mundo", dijo ayer en conferencia, previo a la presentación oficial en el Museo Soumaya.