UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

DESEMPEÑO DE LOS MÉTODOS DEL ESTADO DEL ARTE PARA LA GENERACIÓN AUTOMÁTICA DE RESÚMENES EXTRACTIVOS PARA EL CORPUS TEXTRUSS

TESIS

PARA OBTENER EL TÍTULO DE

INGENIERA EN SOFTWARE

Q U E P R E S E N T A

PALOMA TERESA HERNÁNDEZ MAYA

ASESORA:

DRA. YULIA NIKOLAEVNA LEDENEVA

TIANGUISTENCO, MÉX. MAYO 2018

Declaración de originalidad del trabajo escrito

Mediante esta carta hago constar que el trabajo de tesis presentado en este documento es original porque cito debidamente los contenidos utilizados como soporte a la investigación presentada, por lo que exonero a la Universidad Autónoma del Estado de México de cualquier problema de derechos de propiedad intelectual.

_____

Paloma Teresa Hernández Maya

El comité revisor designado por el Departamento Académico de la Unidad Académica Profesional Tianguistenco de la Universidad Autónoma del Estado de México, aprobó la tesis: **"DESEMPEÑO DE LOS MÉTODOS DEL ESTADO DEL ARTE PARA LA GENERACIÓN AUTOMÁTICA DE RESÚMENES EXTRACTIVOS PARA EL CORPUS TEXTRUSS"** y autorizó la impresión de la misma de la C. **PALOMA TERESA HERNÁNDEZ MAYA** el día **03 de mayo de 2018**.

ATENTAMENTE
PATRIA, CIENCIA Y TRABAJO
"2018, CXC Aniversario de la Universidad Autónoma del Estado de México"

Revisor
M. en Ing. Gerardo Arturo Ávila Vilchis

Revisor
M. en C. C. Rafael Cruz Reyes

Asesora
Dra. en C. C. Yulia Nikolaevna Ledeneva

Dra. en Admón. Adriana Fonseca Mungaia
Jefa del Departamento Académico de la UAP Tianguistenco
Vo.Bo.

AGRADECIMIENTOS

A mis padres, por apoyarme en mis estudios, gracias a sus regaños y consejos han formado en mí una persona con valores.

A mi abuela Tere, por encontrar en ella un refugio cálido, reflexión y sobre todo aceptar las cosas.

A la Dra. Yulia Nikolaevna Ledeneva, por orientarme en este trabajo de tesis y compartirme sus conocimientos.

Al Dr. Rene Arnulfo García Hernandez, por su disposición y haberme apoyado con el módulo de transliteración.

Al Mtro. Rafael Cruz Reyes y al Mtro. Gerardo Ávila Vilchis por su tiempo para la revisión de esta tesis.

A la Mtra. Griselda Areli Matias Mendoza por su apoyo en todo momento y brindarme el método de (Matias, 2016) para elaborar los experimentos.

A todo los compañeros y docentes de seminario por brindarme sus críticas constructivas para mejorar esta tesis.

A la Mtra. Emma, Mtro. Esteban, Mtra. Tere, Mtro. Álvaro, Ing. Alexis, gracias por su tiempo y confianza que me brindaron.

A Erik y amigos por estar conmigo en los mejores, malos y peores momentos.

DEDICATORIA

A Dios por permitirme vivir estos momentos.

A mis padre y hermanos por su amor, comprensión y paciencia.

A mis abuelos, con mucho amor.

RESUMEN

Hoy en día la información digital crece de manera exponencial. Por esto, cuando se realiza una investigación sobre un tema específico en un motor de búsqueda (*Google Search, Yahoo! Search*) nos genera demasiados resultados, por lo cual se complica revisar todos los documentos recuperados que contengan las palabras de la consulta. Uno de los recursos más eficientes utilizados por los usuarios para condensar el volumen de información es el uso de resúmenes.

Un resumen es un texto corto producido a partir de uno o más documentos, clasificado en abstractivo o extractivo. El resumen extractivo se crea a partir de la selección de oraciones sobresalientes del texto original, por otro lado, el resumen abstractivo consiste en interpretar el texto en menos palabras.

Además, existen dos tareas en la generación de un resumen: a partir de un solo documento o a partir de múltiples documentos. El resumen generado de un solo documento consiste en generar un texto corto, mientras que el resumen generado por múltiples documentos consiste en generar un texto corto con los elementos relevantes de éstos.

En este trabajo de tesis se utiliza el resumen de tipo extractivo y con la tarea de un solo documento.

Se han elaborado diversos trabajos que determinan el desempeño de las herramientas comerciales y métodos del estado del arte para la generación automática de resúmenes en el idioma inglés, español, portugués y ruso; utilizando conjuntos de documentos como entrada llamados corpus, los cuales son orientados al dominio de noticias. Sin embargo, en el caso del idioma ruso no se han utilizado diversos métodos del estado del arte.

En este trabajo de tesis se determina el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos de un solo documento utilizando el corpus TEXTRUSS, por medio de la herramienta de evaluación ROUGE (Lin, 2004), utilizando la medida *F-measure* como indicador de evaluación.

Se realizaron experimentos con diferentes configuraciones de parámetros de los métodos del estado del arte para la generación automática de resúmenes en el idioma ruso. Además, se comparan los resultados de los métodos del estado del arte para determinar su desempeño.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

CAPÍTULO 1. INTRODUCCIÓN

1.1 Relevancia de la información

Hoy en día la información digital crece de manera exponencial. Por esto, cuando se investiga sobre un tema específico en un motor de búsqueda (*Google Search, Yahoo! Search*) nos genera demasiados resultados, lo cual se complica revisar todos los documentos recuperados por el hecho que la población no cuenta con el tiempo suficiente para revisarlos dado que son demasiados (Miranda, 2013).

Según un estudio realizado por el Instituto Nacional de Estadística y Geografía (INEGI) se confirma que el uso del internet va en incremento, puesto que el 63.9% de la población mexicana es usuaria de internet, y siendo la búsqueda de información la actividad principal por los internautas (INEGI, 2018).

1.2 Definición del resumen

El resumen se puede definir como:

- Documento corto que transmite la información más importante del documento de origen (Ledeneva, 2008).
- Forma abreviada y concreta el contenido principal de un documento (Castillo, 2009).
- Información en una versión corta preservando el contenido del documento fuente (Vlainic, 2013).
- Reducción a términos breves y precisos de lo esencial de una fuente de información (Barrios, 2015).

1.3 Clasificación del resumen

Para varios autores (Alfonseca, 2003) (Lloret, 2008) (Ledeneva, 2008) (García et. al, 2009) (Montiel, 2009) (Ibáñez, 2013) (Matias, 2013a) (Matias, 2013b) (Mendoza, 2014) (González, 2016) (Matias, 2016) (Ledeneva & García, 2017) los resúmenes se clasifican en extractivos o abstractivos:

- Los resúmenes extractivos se crean a partir de la selección de un conjunto de oraciones sobresalientes del texto original.
- Los resúmenes abstractivos consisten en interpretar el texto en menos palabras, similar al modo en que lo harían las personas.

También existen dos tareas en la generación automática de resúmenes de un solo documento y múltiples documentos (Ledeneva, 2008) (Ledeneva & García, 2017). Las características que describen estos tipos de resúmenes son las siguientes:

- El resumen generado de un solo documento consiste en generar un texto corto.
- El resumen generado por múltiples documentos consiste en generar un texto corto con los elementos relevantes de éstos.

Los resúmenes extractivos de un solo documento son los más utilizados ya que son generados a partir de la selección de oraciones, y no utilizan recursos lingüísticos (Ledeneva, 2008). Además, dan buenos resultados en las investigaciones para la generación automática de resúmenes (Miranda, 2013). Por lo tanto, en este trabajo de tesis se utilizan los resúmenes extractivos de un solo documento, teniendo como antecedente varios trabajos en el estado del arte.

1.4 Métodos del estado del arte

Los métodos del estado del arte más sobresalientes para la generación automática de resúmenes extractivos de un solo documento son:

- Métodos de grafos se probaron en los idiomas inglés y portugués (Mihalcea, 2004) (Mihalcea, 2005).
- Secuencias Frecuentes Maximales (SFM) se probaron en el idioma inglés (Ledeneva, 2008).
- Algoritmos genéticos se probaron en los idiomas inglés, portugués, español, también en ruso (Matias, 2013a) (Matias, 2016). En este último caso no se utilizaron todas las configuraciones de parámetros del método.
- Regresión simbólica se probó en el idioma inglés (Vázquez, 2015).

1.5 Corpus

Para experimentar con los métodos de estado del arte se hace uso de corpus especializados para la generación automática de resúmenes, los corpus más utilizados y del dominio de noticias son:

- DUC-2001 y DUC-2002 (*Document Understanding Conference*) en el idioma inglés
- TER (Textos en español para resúmenes) en el idioma español (Matias, 2016)
- TeMário (*Textos com Sumários* o Textos con resúmenes) en el idioma portugués
- TEXTRUSS (*Translation of Texts in Russian*) en el idioma ruso (Rojas, 2016)

En esta tesis se realizan los experimentos para el corpus TEXTRUSS.

Se han elaborado trabajos para la generación automática de resúmenes extractivos de un solo documento como: (Ledeneva, 2008), (García et. al, 2009), (Montiel, 2009), (Matias, 2013a), (Matias, 2016), (Vázquez, 2015) para el inglés utilizando el corpus

DUC-2002; (Ibáñez, 2013) y (Matias, 2016) para el portugués utilizando el corpus TeMário; (Matias, 2016) y (Gonzáles, 2016) para el español utilizando el corpus TER. Para el idioma ruso solo se sabe del trabajo de (Rojas, 2016) donde se propone y describe el corpus TEXTRUSS, además el método del estado del arte con el que se experimentó, no se realizaron todas sus configuraciones de parámetros.

El método del estado del arte de (Matias, 2016) se probó con el corpus TEXTRUSS en el trabajo de (Rojas, 2016), pero no se ha experimentado con diferentes configuraciones de parámetros correspondientes para la generación automática de resúmenes extractivos de un solo documento, entonces no se terminó de comprobar el desempeño del método al generar un resumen. Por lo tanto, no se sabe el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos de un solo documento utilizando diferentes configuraciones de parámetros de los métodos (Matias, 2016) y (Matias, 2016) + (Vázquez, 2015).

1.6 Planteamiento del problema

¿Cuál es el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos de un solo documento utilizando el corpus TEXTRUSS?

1.7 Objetivos

1.7.1 Objetivo general

Determinar el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos en el idioma ruso usando el corpus TEXTRUSS.

1.7.2 Objetivos específicos

- Utilizar diferentes combinaciones de parámetros de los métodos del estado del arte.
- Transliterar el corpus TEXTRUSS para ser utilizado en los métodos del estado del arte.
- Generar resúmenes extractivos de un solo documento con los métodos del estado del arte, de acuerdo con la longitud del *gold standard* (resumen generado por un humano) y a 100 palabras.
- Calcular las heurísticas *Baseline y Baseline Random*.
- Evaluar con ROUGE (Lin, 2004) los resúmenes generados por los métodos del estado del arte.
- Comparar resultados para determinar el desempeño de los métodos del estado del arte.

1.8 Hipótesis

Si se evalúan los métodos del estado del arte con diferentes configuraciones de parámetros para la generación automática de resúmenes extractivos de un solo documento con los resúmenes hechos por un humano (*gold standard*) del corpus TEXTRUSS, entonces se podrá saber el desempeño de estos métodos.

1.9 Delimitación del problema

- Solo se trabaja con el corpus TEXTRUSS en el idioma ruso.
- Los resúmenes extractivos de un solo documento generados automáticamente por los métodos del estado del arte serán de acuerdo con la longitud del *gold standard* y a 100 palabras.

1.10 Estructura de la tesis

En este primer capítulo, se habla sobre el aumento de la información en Internet, la importancia de los resúmenes, el problema, el objetivo general, los objetivos específicos, la hipótesis y la delimitación del problema.

El resto de la tesis está organizado de la siguiente manera:

En el segundo capítulo, se definen conceptos para entender mejor el desarrollo de esta tesis. Entre ellos se encuentra procesamiento de lenguaje natural, resumen, resumen automático, pasos para generar resúmenes de tipo extractivo, transliteración, herramienta de evaluación ROUGE, heurísticas para la generación de resúmenes: *Baseline* y *Baseline random*, algoritmos genéticos y esquema general del mismo.

En el tercer capítulo se presentan trabajos del estado del arte relacionados a la generación automática de resúmenes extractivos.

En el cuarto capítulo, se realiza una metodología para el desarrollo de resúmenes automáticos las cuales consta de: 1) Análisis del corpus, 2) Pruebas de los métodos del estado del arte, 3) Determinación de los parámetros, 4) Generación de resúmenes, 5) Calculo de heurísticas, 6) Evaluación, 7) Comparación de los resultados.

En el quinto capítulo, se muestran los resultados de los experimentos con los métodos del estado del arte de acuerdo con la longitud del *gold standard* y a 100 palabras.

En el sexto capítulo, se presentan las conclusiones en donde se determina el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos de un solo documento utilizando el corpus TEXTRUSS.

CAPÍTULO 2. MARCO TEÓRICO

En el presente capítulo, se describen los conceptos fundamentales para entender mejor este trabajo de tesis. Inicialmente, se define Procesamiento de Lenguaje Natural (PLN) y sus aplicaciones. Posteriormente se define al resumen y resumen automático de acuerdo con varios autores. Enseguida se describen los cuatro pasos para generar resúmenes de tipo extractivo. Se definen la transliteración y las heurísticas para la generación de resúmenes: *Baseline* y *Baseline random*. Se presenta la definición de ROUGE. Finalmente, Algoritmos genéticos y esquema general del mismo.

2.1 Procesamiento de Lenguaje Natural (PLN)

Gelbukh (2010) define al PLN como la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje.

De acuerdo con Pino et al. (2001) el PLN es permitir la interacción de las personas con las computadoras en el lenguaje cotidiano y así poder hacer consultas a las bases de datos mediante el tipo de pregunta o peticiones que se usarían con otras personas y no teniendo que escribir complejas instrucciones. Por otro lado, consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo ésta entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje (Cortez et al., 2009).

Según Hernandez et al. (2013) y Gelbukh (2010) las tareas principales del PLN son:

- Recuperación y extracción de información
- Traducción automática
- Sistemas de búsquedas de respuestas

- Generación automática de resúmenes
- Desambiguación del sentido de las palabras
- Análisis de sentimientos, entre otras.

2.2 Definición de resumen

Según Ledeneva (2008) es un documento corto que transmite la información más importante del documento de origen.

Mientras que Matias (2013a) menciona que es un texto corto que puede ser ubicado al inicio o final del documento original. Este texto contiene la información más importante sin cambiar el contexto y el orden.

Castillo (2009) lo define como una forma abreviada y concreta del contenido principal de un documento.

Desde el punto de vista de Montiel (2009) el resumen concreta las ideas principales del texto original.

Por lo tanto, un resumen es generar un texto corto por el humano derivado de uno o más documentos originales el cual contiene la información esencial.

2.3 Generación automática de resúmenes

Según Hassel (2007), la generación automática de resúmenes es considerada como una técnica en la que una computadora crea automáticamente un resumen de uno o más textos.

Para Matias (2013a) menciona que implica utilizar una aplicación (software) que tome un texto original, y extraiga la información más importante presentándole al usuario un texto corto.

La generación automática de resúmenes busca seleccionar las oraciones más relevantes en un documento, por lo que es importante establecer las características que ayudan a identificar estas oraciones y así mejorar la calidad de los resúmenes generados (Mendoza, 2014).

En cambio, para Plaza (2010) reduce el texto de uno o varios documentos utilizando un software, de tal forma que el resumen producido automáticamente condense la información importante del documento de entrada.

En conclusión, la generación automática de resúmenes utiliza una herramienta o método computacional creada por el humano con base en ciertos requisitos para obtener un texto corto con la información más importante del texto original.

2.4 Pasos para generar resúmenes extractivos

Ledeneva (2008) propone una lista de cuatro pasos para generar automáticamente resúmenes extractivos y los cuales son:

1. Selección de términos: Decidir qué unidades contarán como términos, por ejemplo, pueden ser palabras, n-gramas u oraciones.
2. Pesado o ponderación de términos: Proceso de ponderación (o estimación) de los términos individuales con respecto al contenido del documento.
3. Pesado o ponderación de oraciones: Proceso de asignación de una medida numérica de utilidad o de la oración, por ejemplo, sumar los pesos de la utilidad de los términos individuales de los cuales se compone la oración.
4. Selección de las oraciones: Se seleccionan oraciones u otras unidades como partes finales del resumen. Una forma sencilla es asignar a las oraciones alguna medida numérica que refleje su utilidad dentro del texto original, y solo seleccionar las mejores al elaborar el resumen.

2.5 Transliteración al idioma ruso

Strutunnof (2012) define al alfabeto ruso como una variante del alfabeto cirílico, compuesto por 33 letras (ver Tabla 1):

Tabla 1. Alfabeto ruso transliterado (Strutunnof, 2012).

Alfabeto Ruso					
Mayúscula	Minúscula	Transliteración	Mayúscula	Minúscula	Transliteración
А	а	a	Р	р	r
Б	б	b	С	с	s
В	в	v	Т	т	t
Г	г	g	У	у	u
Д	д	d	Ф	ф	f
Е	е	e	Х	х	x
Ё	ё	ë	Ц	ц	c
Ж	ж	ż	Ч	ч	ċ
З	з	z	Ш	ш	ṡ
И	и	i	Щ	щ	ṡċ
Й	й	j	Ъ	ъ	''
К	к	k	Ы	ы	y
Л	л	l	Ь	ь	''
М	м	m	Э	э	ė
Н	н	n	Ю	ю	ju
О	о	o	Я	я	ja
П	п	p			

El alfabeto cirílico surge desde hace muchos años como lo menciona Escandell (2016) que en el año 2000 a.C. la escritura estaba ya extinguiendo en el Oriente Próximo. Un milenio más tarde apareció el alfabeto, un sistema de escritura cuyos grafemas tienen valor fonológico y no ideográfico. Según Florian Coulmas (1999), la escritura jeroglífica está en el remoto origen del alfabeto fenicio (hacia 1200 a.C.) del que derivaron el alfabeto hebreo, el árabe y el griego, del cual provienen, a su vez, los alfabetos latino y cirílico.

El alfabeto cirílico (Bulgaria, primera mitad el s. X d.C.) debe su nombre al monje Cirilo, misionero enviado por Bizancio para escribir los textos sagrados en lenguas eslavas, hasta entonces ágrafas. Aunque proviene el alfabeto griego, tiene también semejanza con el latino. Se utiliza para el ruso, otras lenguas eslavas (ucraniano, bielorruso, serbio, búlgaro y macedonio) y de las repúblicas exsoviéticas (turkmeno, azerbaiyano, checheno, kirguís, yakuto, etc.).

Para tratar estos textos del alfabeto cirílico se hace uso de la transliteración.

La transliteración se define como representar los signos de un sistema de escritura mediante los signos de otro (DRAE, 2016).

Es especialmente utilizada por las bibliotecas o para el tratamiento de datos de texto.

La Organización Internacional para la Estandarización define a la transliteración como la acción de representar los caracteres o los signos de un alfabeto por los de otro alfabeto, bajo el principio de letra por letra (Orozco, 2016). En la Tabla 2 se encuentra una lista de letras en ruso transliteradas a letras latinas, y en la Figura 1 se da un ejemplo de transliteración.

Tabla 2. Transliteración de letras cirílicas a letras latinas (Translit, 2016).

A	B	V	G	D	E	Jo	Zh	Z	I	J	K	L	M	N	O	P	R	S	T	U	F	H	C	Ch	Sh	Shh	##	Y	"	Je	Ju	Ja
a	b	v	g	d	e	jo,yo,ö	zh	z	i	j	k	l	m	n	o	p	r	s	t	u	f	h,x	c	ch	sh	shh,w	#	y	'	je,ä	ju,yu,ü	ja,ya,q
а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

Figura 1. Ejemplo de transliteración (elaboración propia).

2.6 Heurísticas para la generación automática de resúmenes

2.6.1 *Baseline*

Selecciona las primeras *n* oraciones del texto original, hasta alcanzar el número de palabras deseadas. Esta configuración da muy buenos resultados en los textos de dominio de noticias (Ledeneva, 2008).

2.6.2 *Baseline random*

Selecciona aleatoriamente *n* oraciones del texto original (Ledeneva, 2008).

2.7 Herramienta de evaluación para la generación automática de resúmenes

En un principio la evaluación de resúmenes se hacía manual, esto quiere decir que los humanos juzgaban directamente la calidad de los resúmenes, por lo tanto, implicaba tiempo. Posteriormente se desarrollaron métodos de evaluación automática (Lin, 2004).

El método de evaluación de resúmenes intrínseco más utilizado hoy en día es ROUGE (Lin, 2004), el cual compara un resumen candidato (generado automáticamente) con un resumen modelo (creado por un humano).

Las métricas con las que trabaja ROUGE son:

Precisión

Probabilidad de que una oración relevante sea detectada (ver Ecuación 1):

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

Ecuación 1. Fórmula de precisión utilizada por ROUGE.

Recuerdo

Probabilidad de encontrar oraciones importantes (ver Ecuación 2):

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

Ecuación 2. Fórmula de recuerdo utilizada por ROUGE.

Haciendo una explicación de las palabras correctas, olvidadas e incorrectas utilizadas en las ecuaciones anteriores podemos decir que:

$correctas$: Número total de oraciones extraídas por el sistema y el humano.

$olvidadas$: Número total de oraciones extraídas por el sistema, pero no por el humano.

$incorrectas$: Número total de oraciones extraídas por el humano, pero no por el sistema.

F-measure

Métrica que combina el recuerdo y la precisión (ver Ecuación 3), donde $recuerdo * precisión$ son los resultados arrojados de cada una de sus operaciones (Precisión y Recuerdo):

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

Ecuación 3. Fórmula de *F-measure* utilizada por ROUGE.

2.8 Algoritmos genéticos

Cito textualmente la definición de algoritmo genéticos:

> *"Los algoritmos genéticos son métodos adaptativos, generalmente usados en problemas de búsqueda y optimización de parámetros, basados en la reproducción sexual y en el principio supervivencia del más apto (Gestal, 2013)"*.

2.8.1 Esquema general de un algoritmo genético

El funcionamiento de un algoritmo genético simple se puede apreciar en el siguiente segmento de pseudo-código (Coello, 1995):

```
generar población inicial, G(0);
evaluar G(0);
t=0;
repetir
        t:=t+1;
        generar G(t) usando G(t-1);
        evaluar G(t);
hasta encontrar una solución;
```

Antes de describir cada una de las etapas es importante tener la representación del individuo. La representación del individuo depende del problema que se desea resolver. Desde los primeros trabajos de John Holland la representación suele hacerse mediante valores binarios (Gestal, 2013). Pero también se puede representar mediante números enteros o incluso cadenas de palabras (Arranz, 2018).

2.8.2 Generar la población inicial

Se genera aleatoriamente la población inicial, que estará constituida por un conjunto de cromosomas o cadena de caracteres que presentan las soluciones posibles del problema (Coello, 1995). Los cromosomas de la población inicial suelen ser cadenas de cero y uno generados de forma aleatoria.

2.8.3 Función de aptitud

La función de aptitud depende del problema que se quiera resolver (Matias, 2013a). A cada uno de los cromosomas de la población inicial se le aplica la función de aptitud con el fin de saber que tan buena es la solución que está codificando.

Para determinar que cromosoma es una buena solución es necesario calificarlo de alguna manera. Cada cromosoma de cada generación de un algoritmo genético recibe una calificación o, para usar el término biológico, una medida de su grado de adaptación (*fitness*).

2.8.4 Condición de parada

Se refiere a la condición que se debe cumplir para que el algoritmo deje de evolucionar y encontrar la mejor solución. En esta tesis se utiliza la condición de parada como número de generaciones.

2.8.5 Selección

Los operadores de selección son los encargados de escoger el conjunto de individuos con mejores condiciones a reproducirse. En esta tesis se utiliza los siguientes operadores:

- Selección por ruleta: cada individuo tiene una parte de esa ruleta mayor o menor en función a la puntuación que tenga cada uno. Se hace girar la ruleta y se selecciona el individuo en el que se para la ruleta. El individuo con mayor puntuación saldrá con mayor probabilidad (Arranz, 2018).
- Selección por torneo: se eligen subgrupos de individuos de la población, y los miembros de cada subgrupo compiten entre ellos. Sólo se elige a un individuo de cada subgrupo para la reproducción (Mateos, 2017).

2.8.6 Cruza

Se requiere de dos individuos seleccionados en la etapa anterior (Gestal, 2013). Y los sobrevivientes a la selección combinan sus propiedades y sus hijos o descendientes formaran la población de la siguiente generación (Coello, 1995).

2.8.7 Mutación

El operador de mutación crea un nuevo subconjunto de individuos con pequeñas modificaciones de características para preservar la diversidad de los individuos en una población (Arranz, 2018).

CAPÍTULO 3. ESTADO DEL ARTE

En este capítulo, se exponen los trabajos que están relacionados con el tema de generación automática de resúmenes.

3.1 Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization

Ledeneva (2008) generó resúmenes de tipo extractivo, trabajó con el corpus DUC-2002 para uno y multi-documentos, los resúmenes generados fueron evaluados con ROUGE.

En el trabajo se expone cuatro pasos para generar un resumen de texto los cuales son: selección de términos, pesado de términos, pesado de oraciones, selección de oraciones. Utilizó secuencias frecuentes maximales como técnica principal y calcula *Topline*. El resultado más bajo que obtuvo fue *Baseline random*.

3.2 Comparing Commercial Tools and State-of-the-Art Methods for Generating Text Summaries

García et al. (2009) realizó la comparación de herramientas comerciales y métodos del estado del arte para saber cuál de ellos fue la mejor para generar resúmenes automáticos; los resúmenes que se generaron con las herramientas comerciales fueron de tipo extractivo y utilizó el corpus DUC-2002 en el idioma inglés. Para evaluar los resúmenes utilizó la herramienta de evaluación ROUGE. Las herramientas comerciales utilizadas fueron: *Copernic Summarizer, Microsoft Office Word 2003, Microsoft Office Word 2007* y los métodos del estado del arte utilizados fueron: *TextRank* (Mihalcea, 2004)*, Maximal Frequent Sequences (SFMs)* (Ledeneva, 2008)*, Clustering con SFMs* (Ledeneva, 2011) y las heurísticas *Baseline* y *Baseline random*. El resultado más bajo fue para la heurística *Baseline random* y el mejor para el método del estado del arte *Clustering* con SFMs.

3.3 Generación automática de resúmenes usando algoritmos genéticos

Matias (2013a) realizó resúmenes extractivos para un solo documento independiente del lenguaje, utilizó el modelo de texto de bi-gramas mediante la implementación de un algoritmo genético. Los documentos que utilizó como entrada para probar el método propuesto fue la colección DUC-2002, la cual ha sido diseñada para este tipo de experimentos. Los resúmenes resultantes son evaluados con la herramienta de evaluación ROUGE. El mejor resultado que obtuvo fue con el método propuesto utilizando un algoritmo genético y el peor resultado fue *Baseline random*.

3.4 Evaluación de las herramientas comerciales y métodos del estado del arte para la generación de resúmenes extractivos individuales

Matias (2013b) comparó las herramientas comerciales con los métodos del estado del arte para generar resúmenes. Utilizó el corpus DUC-2002 en el idioma inglés y evaluó los resúmenes con la herramienta de evaluación ROUGE. El mejor resultado que se obtuvo fue implementando un algoritmo genético con el modelo de texto bi-gramas.

3.5 Modelo de relevancia de la posición de las oraciones en resúmenes de texto, mediante regresión simbólica

Vázquez (2015) realizó un método para generar modelos para determinar la posición de las oraciones en resúmenes de texto mediante regresión simbólica. Los datos que se utilizaron como el conjunto de entrada es la colección DUC-2002. Las etapas de este método son: en la primera etapa, una entrada compuesta por un conjunto de datos, los cuales describen con qué frecuencia fueron utilizadas las oraciones de un conjunto de documentos para generar los resúmenes de éste. En la segunda etapa, se utiliza la técnica de programación genética, en esta etapa se generan modelos que describan la relación entre los datos de entrada. Tercera, se evalúa el modelo

obtenido en la etapa anterior con el conjunto de datos de entrada, con el fin de averiguar qué tan acertado es el modelo resultante.

3.6 Generación Automática de Resúmenes Independientes del Lenguaje

Matias (2016) realizó un algoritmo genético para generar resumes independiente del lenguaje y realiza experimentos para tres idiomas: inglés, portugués y español. Las colecciones de documentos que utiliza en este trabajo son: para inglés-la colección DUC2002, para portugués-la colección TeMário y para el idioma español-TER. La colección TER fue la aportación en este trabajo (Matias, 2016), la cual es una colección de noticias de un periódico mexicano (La crónica) especialmente para el uso de resúmenes. Los resúmenes resultantes son evaluados con la herramienta de evaluación ROUGE.

Los resultados fueron los siguientes. Para el idioma inglés (DUC2002): el método propuesto supera a las herramientas comerciales y métodos del estado del arte incluyendo el de (Matias, 2013a). Para el idioma portugués (TeMário): el método propuesto supera a todas las herramientas comerciales en línea y obtiene los mejores resultados en el estado del arte. Para el idioma español (TER): el método propuesto supera los resultados obtenidos con las herramientas comerciales instalables y en línea.

En la siguiente figura (ver Figura 2) se muestran los resultados de los métodos del estado del arte antes descritos con la colección DUC-2002 para el idioma inglés:

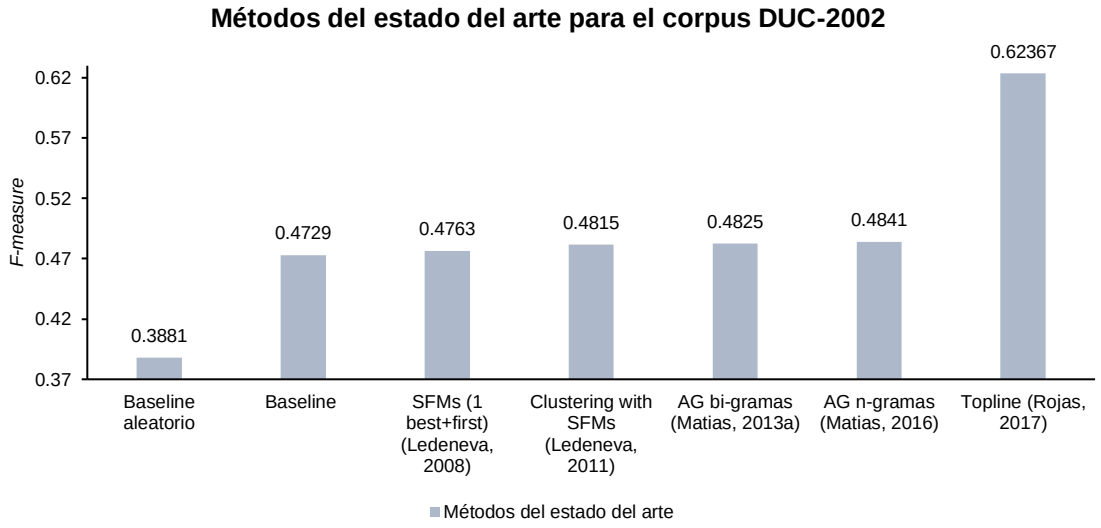

Figura 2. Resultados del estado del arte con el corpus DUC-2002 para el idioma inglés (Montiel, 2009) (Ledeneva, 2008) (Ledeneva, 2011) (Matias, 2013a) (Matias, 2016) y (Rojas, 2017).

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

Rojas (2016) creó el corpus TEXTRUSS en idioma ruso en el dominio de noticias y generó resúmenes extractivos con herramientas comerciales y el método de (Matias, 2016) a 100 palabras.

Llegó a las conclusiones que fueron las siguientes: la mejor herramienta comercial fue *t-CONSPCTUS* y el método del estado del arte de (Matias, 2016) supera a las herramientas comerciales en la generación automática de resúmenes. El *Baseline* supera a las herramientas comerciales y al método del estado del arte de (Matias, 2016). Los resultados pueden apreciarse en la Figura 3.

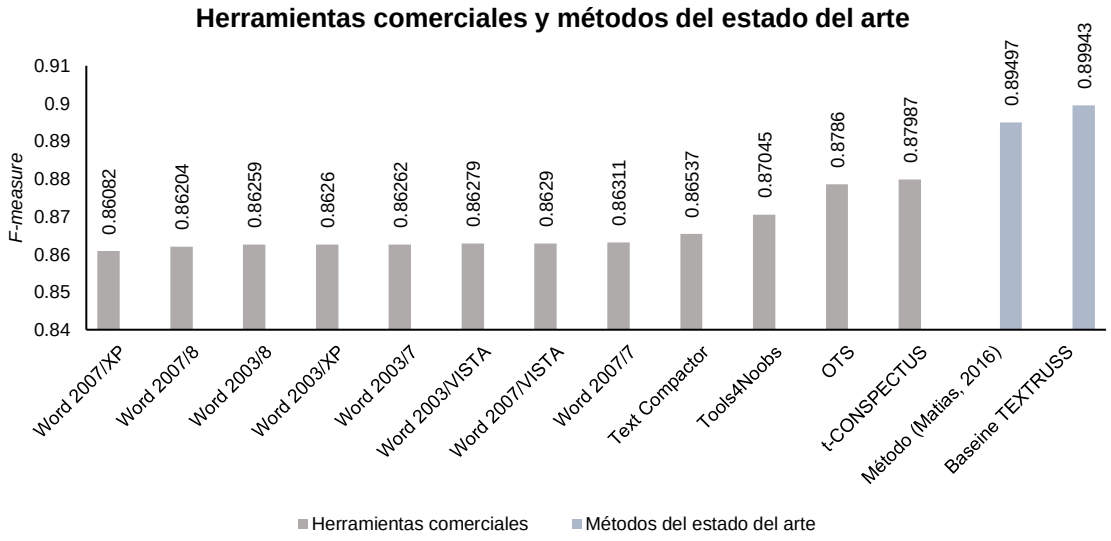

Figura 3. Compracion de las herramientas comerciales y los métodos del estado del arte (Rojas, 2016).

CAPÍTULO 4. METODOLOGÍA PROPUESTA

En este capítulo, se describe la metodología propuesta para la generación automática de resúmenes extractivos con el corpus TEXTRUSS en el idioma ruso.

A continuación, se muestra la metodología propuesta (ver Figura 4):

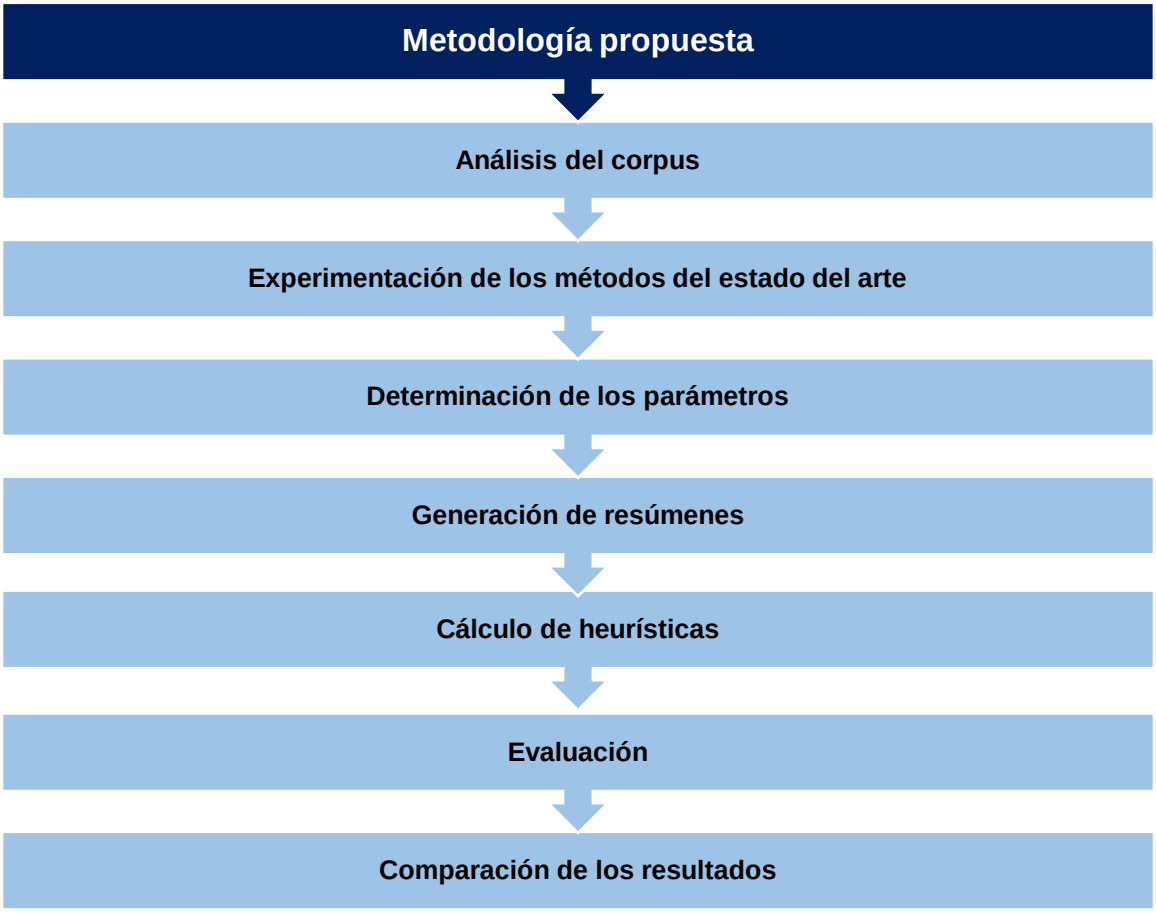

Figura 4. Metodología propuesta (elaboración propia).

4.1 Análisis del corpus

El corpus TEXTRUSS es un compendio de noticias en idioma ruso, está conformado por tres capetas: la primera carpeta llamada Texto completo con título contiene las noticias con título, descripción de foto ilustrativa de la noticia, fecha y hora de publicación, nombre del autor de la noticia y la noticia completa. La segunda carpeta llamada Texto sin título donde solo contienen la noticia sin título. La tercera carpeta llamada Resúmenes *gold standard* contiene un resumen elaborado por un humano de cada una de las noticias (ver Figura 5).

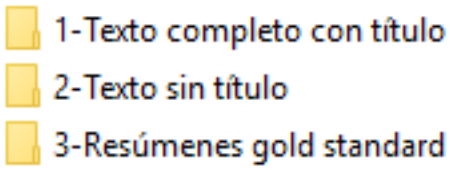

Figura 5. Carpetas que conforman el corpus TEXTRUSS.

A su vez, las tres carpetas contienen 11 categorías: política, negocios, compañía, criticas, cultura, ciencia, tecnología, bienes inmuebles, auto, estilo de vida, deportes y un Readme.txt donde se explica mejor la sección de categorías (ver Figura 6).

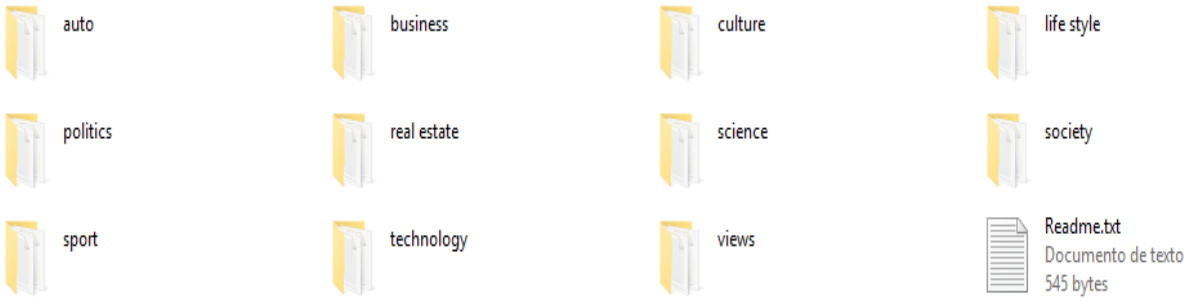

Figura 6. Once categorías con que trabaja el corpus TEXTRUSS.

Y cada una de las 11 categorías contiene 22 noticias (archivos), haciendo un total de 242 noticias de todo el corpus (archivos) (ver Figura 7).

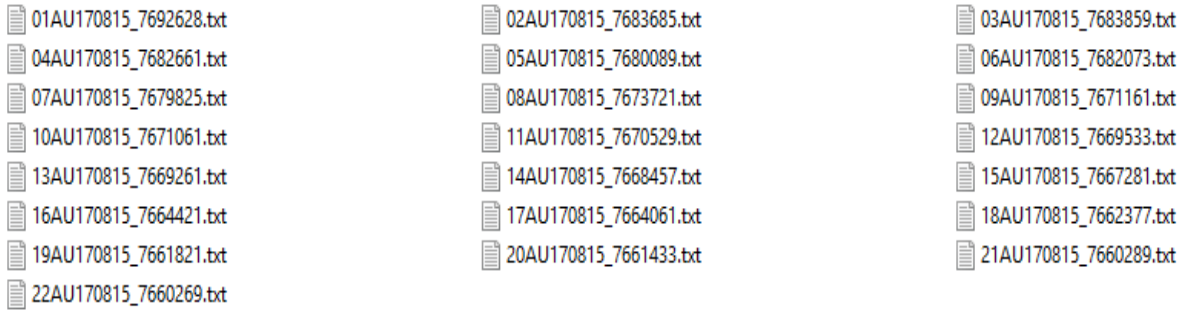

Figura 7. Cada categoría del corpus TEXTRUSS contiene veintidós noticias.

4.2 Experimentación de los métodos del estado del arte

Como se mencionó en el capítulo 1, existen varios métodos del estado del arte, pero no se sabe el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos de un solo documento utilizando diferentes configuraciones de parámetros de los métodos (Matias, 2016) y (Vázquez, 2015) para el idioma ruso. Por lo tanto, los métodos del estado del arte para la generación automática de resúmenes extractivos de un documento utilizados en esta tesis son:

- AG (Matias, 2016)
- AG (Matias, 2016) + (Vázquez, 2015)

4.3 Determinación de los parámetros

El método de (Matias, 2016) utiliza los siguientes parámetros para generar resúmenes extractivos de un solo documento:

- Pre-procesamiento: sí o no
- Modelos de texto:
    - Bolsa de palabras
    - Bi-gramas
    - Tri-gramas
    - Tetra-gramas
    - Penta-gramas

- Importancia de las oraciones: son los valores de la pendiente (ver Tabla 3):

Tabla 3. Valores de la pendiente.

-0.25	-0.3	-0.375	-0.45	-0.5	-0.55	-0.6	-0.625	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	-1

- Función de aptitud
- Operador de selección: Ruleta

También se realizan experimentos con relevancia de la posición de las oraciones, que son modelos que se generan con el método de (Vázquez, 2015) aplicados al método de (Matias, 2016).

- Pre-procesamiento: sí o no
- Modelos de texto:
    - Bolsa de palabras
    - Bi-gramas
    - Tri-gramas
    - Tetra-gramas
    - Penta-gramas
- Importancia de las oraciones: Modelos de (Vázquez, 2015)
- Operador de selección: Ruleta

4.4 Generación de resúmenes

Como se aprecia en el estado del arte, los resúmenes generados con las herramientas comerciales y métodos del estado del arte son a 100 palabras para el idioma inglés. En este trabajo de tesis se generan los resúmenes de acuerdo con la longitud del *gold standard* y a 100 palabras con los métodos del estado del arte en diferentes configuraciones de parámetros.

4.5 Cálculo de heurísticas

Se calculan las heurísticas *Baseline* y *Baseline random* a dos longitudes:

1) Longitud del *gold standard*

El *Baseline* del corpus se calcula extrayendo las primeras N palabras del texto original.

El *Baseline random* del corpus se calcula extrayendo N palabras aleatoriamente.

Donde N = cantidad de palabras del *gold standard* (cantidad de palabras de un resumen elaborado por un humano).

2) A 100 palabras

El *Baseline* del corpus se calcula extrayendo las primeras 100 palabras.

El *Baseline random* del corpus se calcula extrayendo 100 palabras aleatoriamente.

4.6 Evaluación

Los resúmenes extractivos generados automáticamente con los métodos del estado del arte se evalúan con la herramienta de evaluación ROUGE haciendo uso de la métrica *F-measure*.

4.7 Comparación de los resultados

En esta etapa se analizan y comparan los resultados de los experimentos con la finalidad de determinar el desempeño de los métodos del estado del arte.

CAPÍTULO 5. EXPERIMENTACIÓN Y RESULTADOS

En este capítulo, se describen los resultados obtenidos por los métodos del estado del arte de acuerdo con dos diferentes longitudes de un resumen. Primero, la longitud que se calcula a partir de la longitud del resumen del *gold standard*. Segundo, la longitud definida por la tarea que es de 100 palabras.

5.1 Resultados de los métodos del estado del arte a la longitud del *gold standard*

A continuación, se presentan los experimentos realizados con el método del estado del arte de (Matias, 2016) a la longitud del *gold standard,* con la siguiente configuración de parámetros utilizados (ver Tabla 4).

Tabla 4. Parámetros utilizados en el método del estado del arte de (Matias, 2016) a la longitud del *gold standard*.

Parámetros	
Pre-procesamiento	No
Modelo de texto	*n*-gramas (*n* = 1,2,3,4,5)
Importancia de las oraciones	Valor de la pendiente
Función de aptitud	0.5δ+0.5β
Operador de selección	Ruleta

En la Figura 8, se observan los resultados con el modelo de texto Bolsa de palabras y el mejor resultado fue con la pendiente -0.625 obteniendo un desempeño de 0.90863.

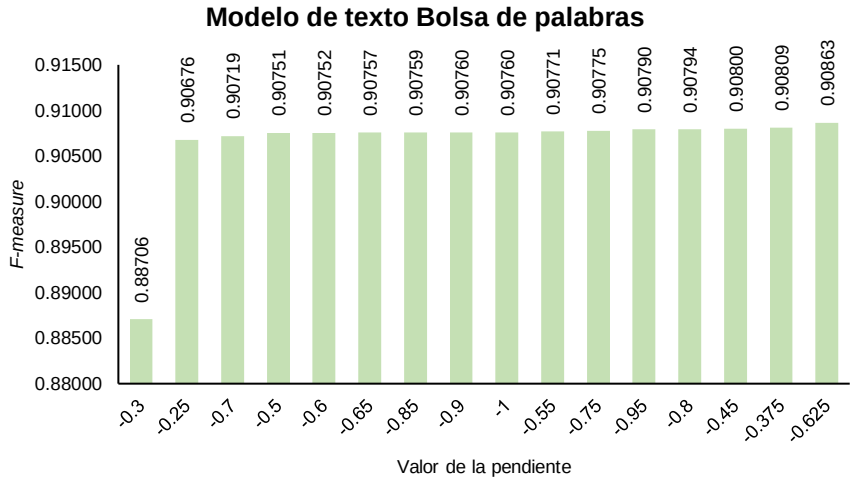

Figura 8. Resultados con el modelo de texto Bolsa de palabras con las 16 pendientes a la longitud del *gold standard.*

En la Figura 9, se observan los resultados con el modelo de texto Bi-gramas y el mejor resultado fue con la pendiente -0.45 obteniendo un desempeño de 0.90746.

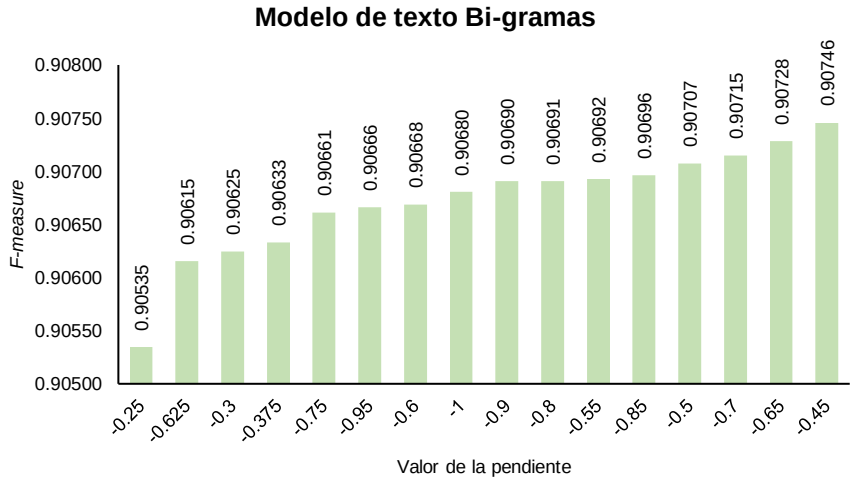

Figura 9. Resultados con el modelo de texto Bi-gramas con las 16 pendientes a la longitud del *gold standard*.

En la Figura 10, se observan los resultados con el modelo de texto Tri-gramas y el mejor resultado fue con la pendiente -0.65 obteniendo un desempeño de 0.90647.

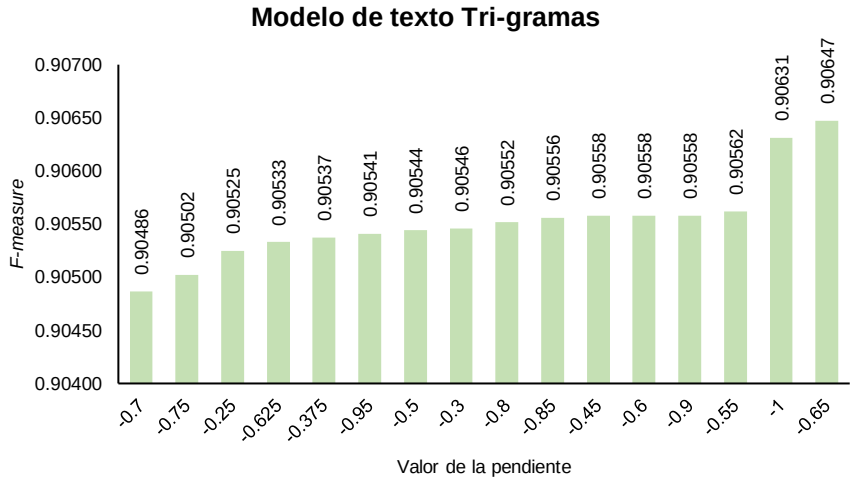

Figura 10. Resultados con el modelo de texto Tri-gramas con las 16 pendientes a la longitud del *gold standard*.

En la Figura 11, se observan los resultados con el modelo de texto Tetra-gramas y el mejor resultado fue con la pendiente -1 obteniendo un desempeño de 0.90645.

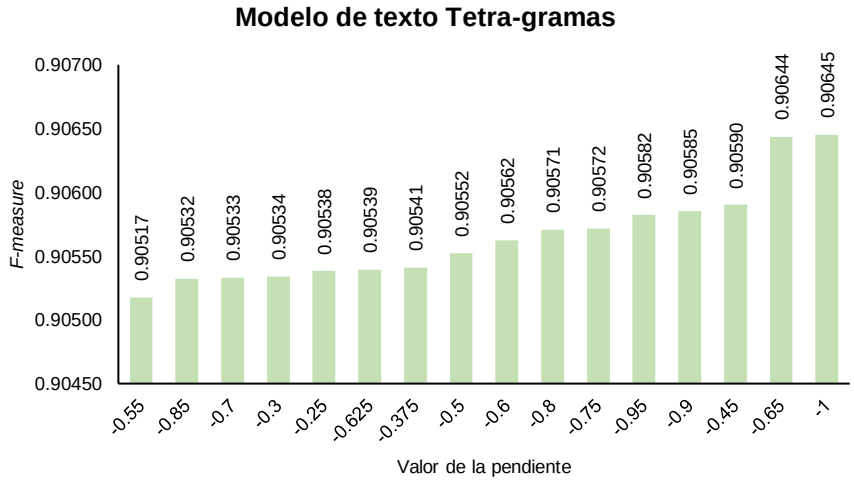

Figura 11. Resultados con el modelo de texto Tetra-gramas con las 16 pendientes a la longitud del *gold standard*.

En la Figura 12, se observan los resultados con el modelo de texto Penta-gramas y el mejor resultado fue con la pendiente -1 obteniendo un desempeño de 0.90649.

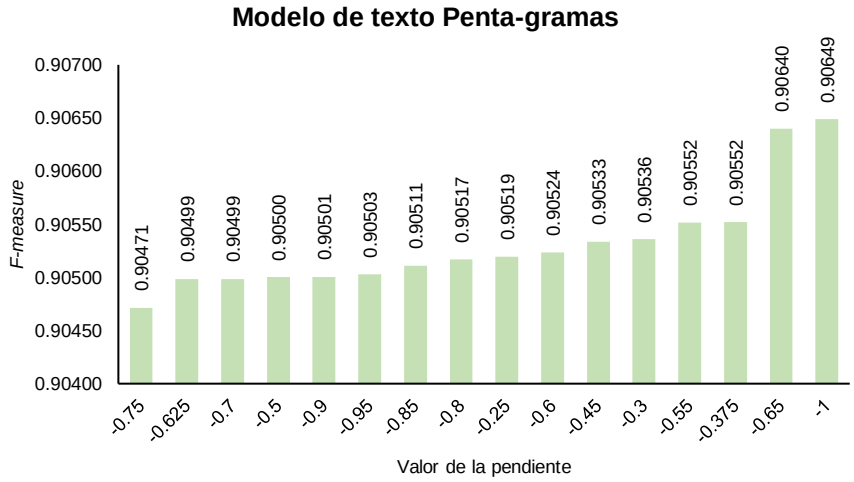

Figura 12. Resultados con el modelo de texto Penta-gramas con las 16 pendientes a la longitud del *gold standard*.

En la Figura 13, se observan los mejores resultados de los cinco modelos de texto y el valor de la pendiente que mejor se adapta a cada uno de ellos, del método del estado del arte de (Matias, 2016) a la longitud del *gold standard*. El mejor resultado que se obtuvo fue con el modelo de texto Bolsa de palabras con la pendiente de -0.625 con un desempeño de 0.90863.

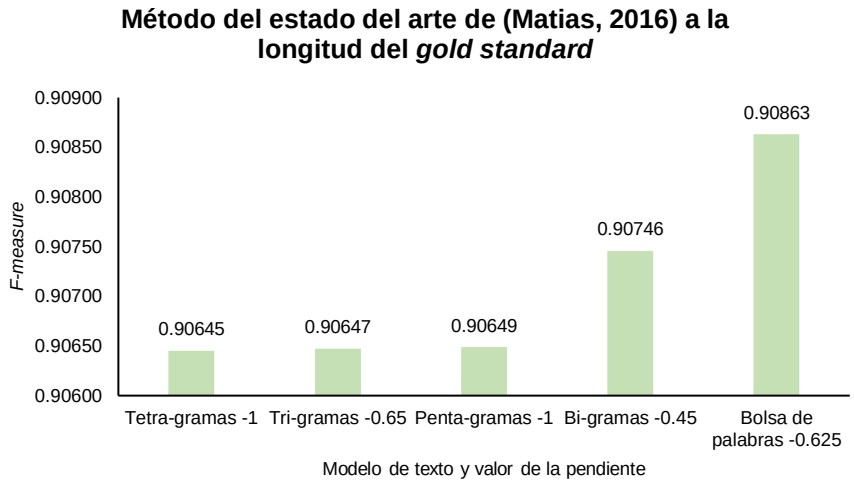

Figura 13. Resultados del método del estado del arte de (Matias, 2016) a la longitud del *gold standard*.

Los siguientes experimentos se realizaron con el método del estado de arte de (Matias, 2016) +(Vázquez, 2015) y la configuración de parámetros utilizados se muestran en la Tabla 5.

Tabla 5. Parámetros utilizados en el método del estado del arte de (Matias, 2016) +(Vázquez, 2015) a la longitud del *gold standard*.

Parámetros	
Pre-procesamiento	No
Modelo de texto	*n*-gramas (*n* = 1,2,3,4,5)
Importancia de las oraciones	Modelos de (Vázquez, 2015)
Operador de selección	Ruleta

En la Figura 14, se observan los resultados que se generaron con el método del estado de arte de (Matias, 2016) +(Vázquez, 2015) y el mejor resultado es con el modelo de texto Bolsa de palabras con un desempeño de 0.90791.

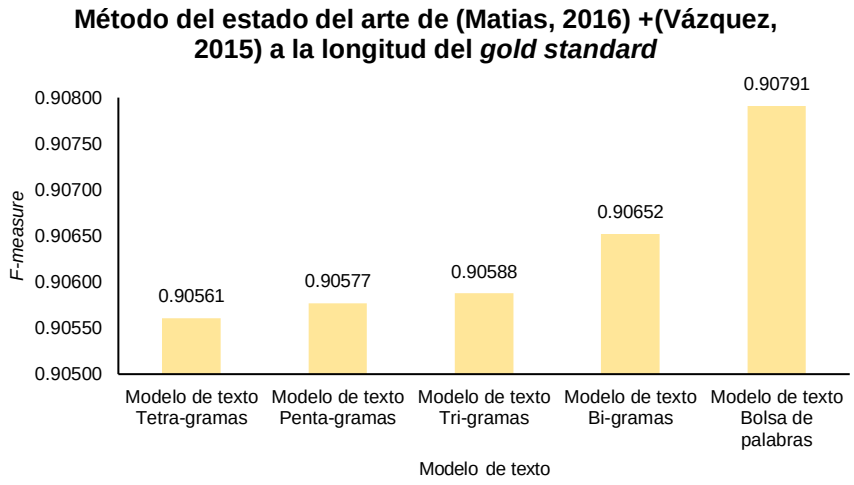

Figura 14. Resultados del método del estado del arte de (Matias, 2016) +(Vázquez, 2015) a la longitud del *gold standard*.

Cabe mencionar que los resultados de las Figuras 13 y 14 se utilizan para la comparación de los métodos del estado del arte a la longitud del *gold standard*.

5.2 Comparación de los métodos del estado del arte a la longitud del *gold standard*

En la Figura 15, se comparan los resultados de los métodos del estado del arte. La heurística *Baseline random* obtiene el resultado más bajo obteniendo un desempeño de 0.89999 y la heurística *Baseline* obtiene un desempeño de 0.90459. El mejor resultado que se obtuvo con el método del estado del arte de (Matias, 2016) +(Vázquez, 2015) es con un desempeño de 0.90791. El método del estado del arte de (Matias, 2016) obtiene el mejor desempeño para la generación automática de resúmenes con 0.90863.

En la Tabla 6, se observa la mejor configuración de parámetros con el método de (Matias, 2016) a la longitud del *gold standard*.

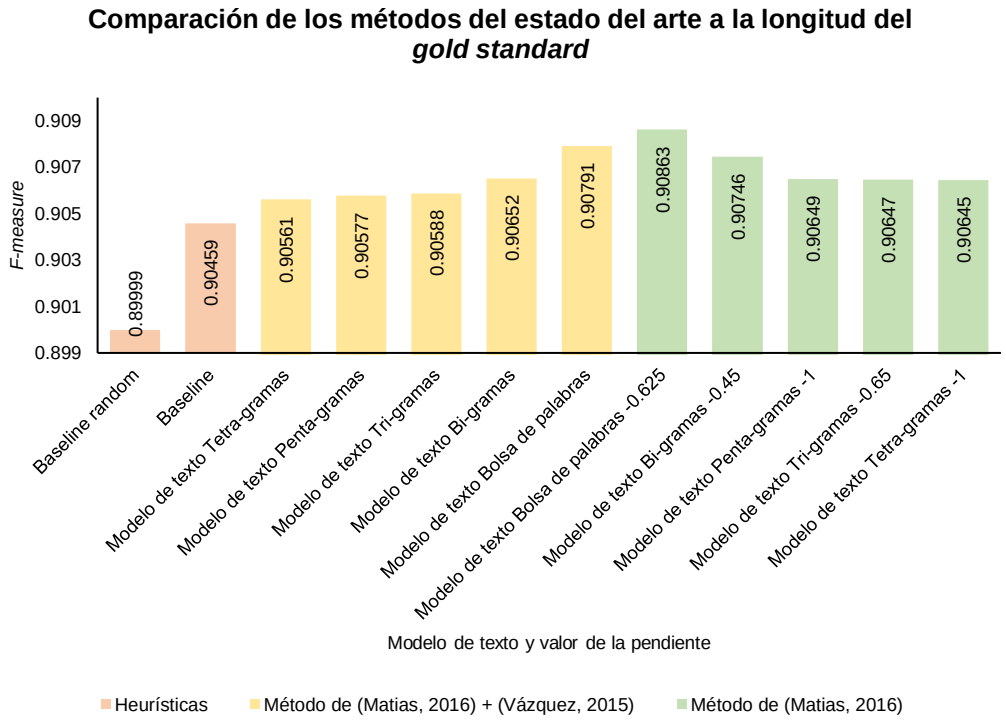

Figura 15. Comparación de los métodos de estado del arte la longitud del *gold standard*.

Tabla 6. Configuración de parámetros para el método del estado del arte de (Matias, 2016) a longitud del *gold standard*.

Parámetros a la longitud del *gold standard*	
Pre-procesamiento	No
Modelo de texto	Bolsa de palabras ó $n = 1$
Importancia de las oraciones (valor de la pendiente)	-0.625
Función de aptitud	0.5δ+0.5β
Operador de selección	Ruleta

5.3 Resultados de los métodos del estado del arte a la longitud de 100 palabras

A continuación, se presentan los experimentos realizados con el método del estado del arte (Matias, 2016) a la longitud 100 palabras, con la siguiente configuración de parámetros utilizados (ver Tabla 7).

Tabla 7. Parámetros utilizados en el método del estado del arte de (Matias, 2016) a la longitud de 100 palabras.

Parámetros	
Pre-procesamiento	No
Modelo de texto	*n*-gramas (*n* = 1,2,3,4,5)
Importancia de las oraciones	Valor de la pendiente
Función de aptitud	0.5δ+0.5β
Operador de selección	Ruleta

En la Figura 16, se observan los resultados con el modelo de texto Bolsa de palabras y el mejor resultado fue con la pendiente -0.85 obteniendo un desempeño de 0.89660.

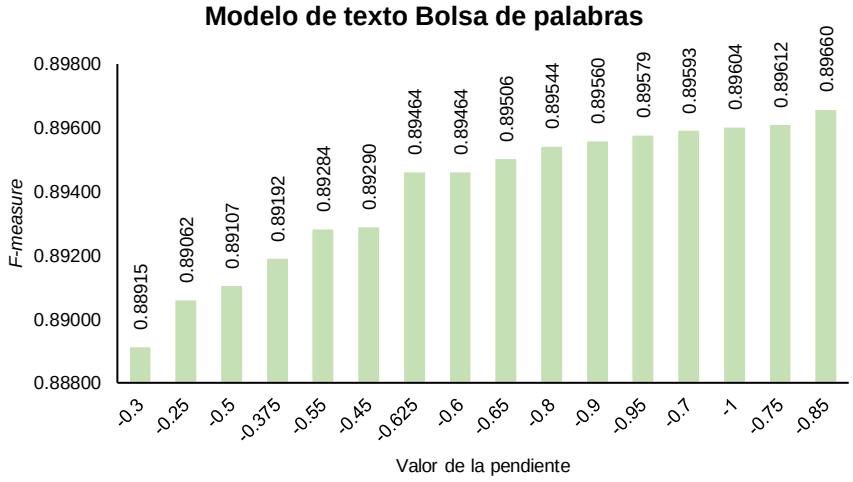

Figura 16. Resultados con el modelo de texto Bolsa de palabras con las 16 pendientes a la longitud de 100 palabras.

En la Figura 17, se observan los resultados con el modelo de texto Bi-gramas y el mejor resultado fue con la pendiente -0.95 obteniendo un desempeño de 0.89783.

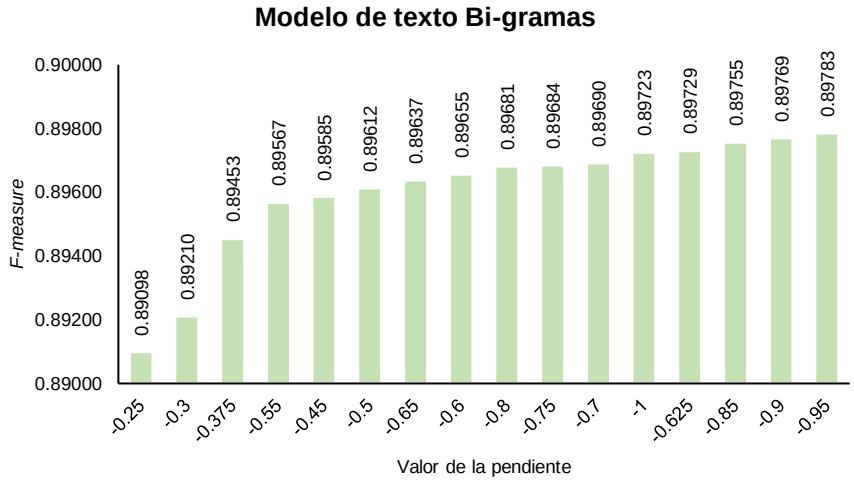

Figura 17. Resultados con el modelo de texto Bi-gramas con las 16 pendientes a la longitud de 100 palabras.

En la Figura 18, se observan los resultados con el modelo de texto Tri-gramas y el mejor resultado fue con la pendiente -0.85 obteniendo un desempeño de 0.89960.

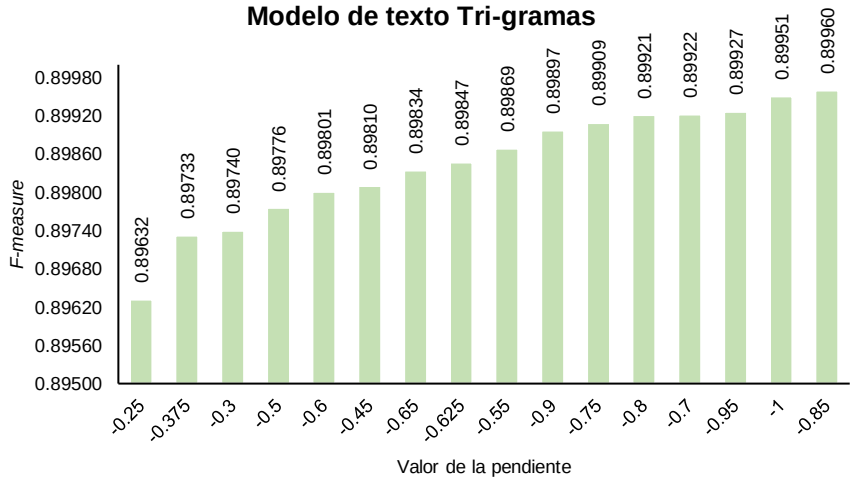

Figura 18. Resultados con el modelo de texto Tri-gramas con las 16 pendientes a la longitud de 100 palabras.

En la Figura 19, se observan los resultados con el modelo de texto Tetra-gramas y el mejor resultado fue con la pendiente -0.5 obteniendo un desempeño de 0.90010.

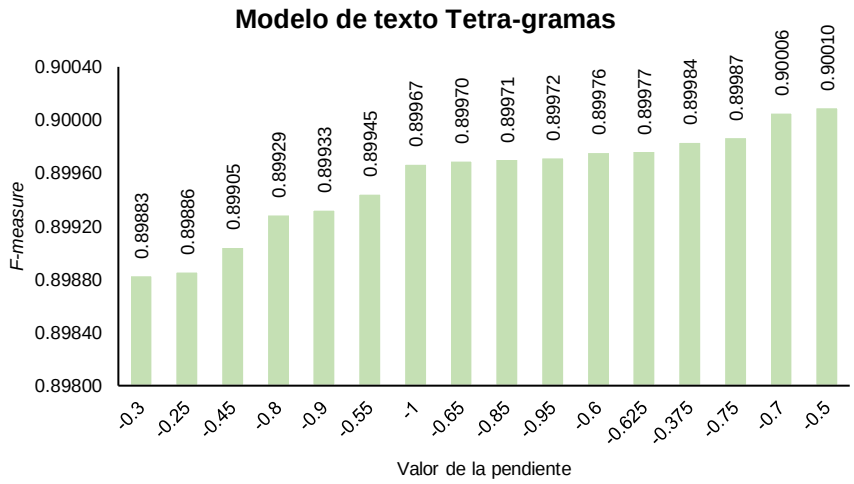

Figura 19. Resultados con el modelo de texto Tetra-gramas con las 16 pendientes a la longitud de 100 palabras.

En la Figura 20, se observan los resultados con el modelo de texto Penta-gramas y el mejor resultado fue con la pendiente -1 obteniendo un desempeño de 0.89938.

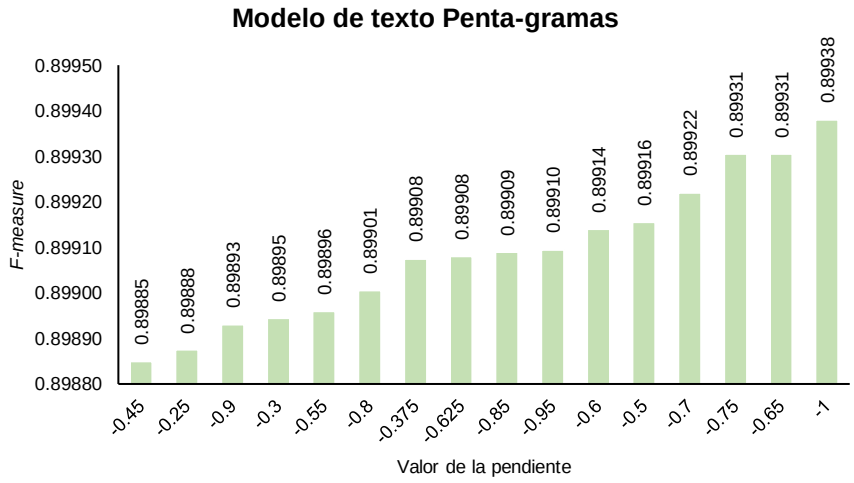

Figura 20. Resultados con el modelo de texto Penta-gramas con las 16 combinaciones a la longitud de 100 palabras.

En la Figura 21, se observan los mejores resultados de los cinco modelos de texto y el valor de la pendiente que mejor se adapta a cada uno de ellos, del método del estado del arte de (Matias, 2016) a la longitud de 100 palabras. El mejor resultado que se obtuvo fue con el modelo de texto Tetra-gramas con la pendiente de -0.5 con un desempeño de 0.90010.

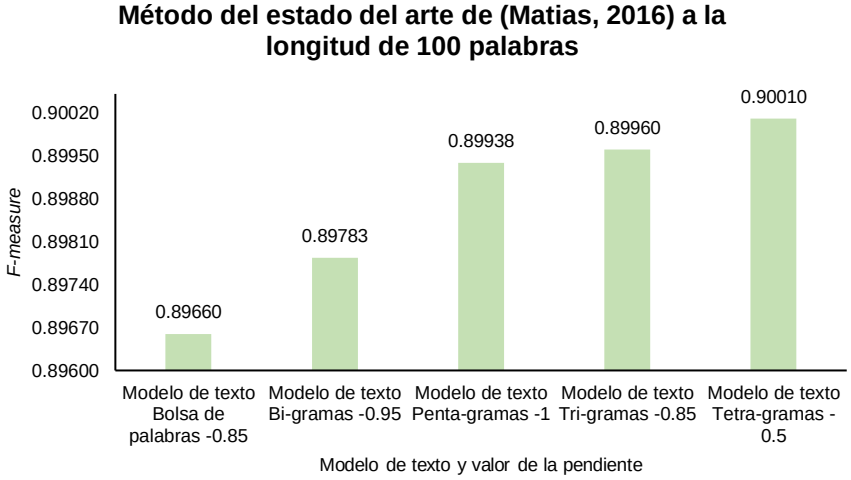

Figura 21. Resultados del método del estado del arte de (Matias, 2016) a la longitud de 100 palabras.

Los siguientes experimentos se realizaron con el método del estado de arte de (Matias, 2016) +(Vázquez, 2015) y la configuración de parámetros utilizados se muestran en la Tabla 8.

Tabla 8. Parámetros utilizados en el método del estado del arte de (Matias, 2016) +(Vázquez, 2015) a la longitud de 100 palabras.

Parámetros	
Pre-procesamiento	No
Modelo de texto	*n*-gramas (*n* = 1,2,3,4,5)
Importancia de las oraciones	Modelos de (Vázquez, 2015)
Operador de selección	Ruleta

En la Figura 22, se observan los resultados que se generaron con el método del estado de arte de (Matias, 2016) +(Vázquez, 2015) y el mejor resultado es con el modelo de texto Tetra-gramas con un desempeño de 0.89981.

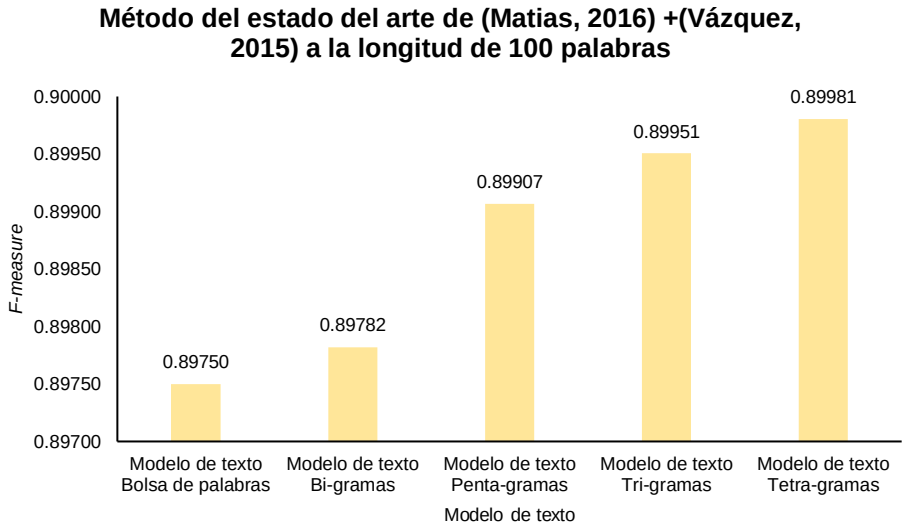

Figura 22. Resultados del método del estado del arte de (Matias, 2016) +(Vázquez, 2015) a la longitud de 100 palabras.

Cabe mencionar que los resultados de las Figuras 21 y 22 se utilizan para la comparación de los métodos del estado del arte a la longitud de 100 palabras.

5.4 Comparación de los métodos del estado del arte a la longitud de 100 palabras

En la Figura 23, se comparan los resultados de los métodos del estado del arte. La heurística *Baseline random* obtiene el resultado más bajo obteniendo un desempeño de 0.87333 y la heurística *Baseline* obtiene un desempeño de 0.89946. El mejor resultado que se obtuvo con el método del estado del arte (Matias, 2016) +(Vázquez, 2015) es con un desempeño de 0.89981. El método del estado del arte de (Matias, 2016) obtiene el mejor desempeño para la generación automática de resúmenes con 0.90010.

En la Tabla 9, se observa la mejor configuración de parámetros con el método de (Matias, 2016) a la longitud de 100 palabras.

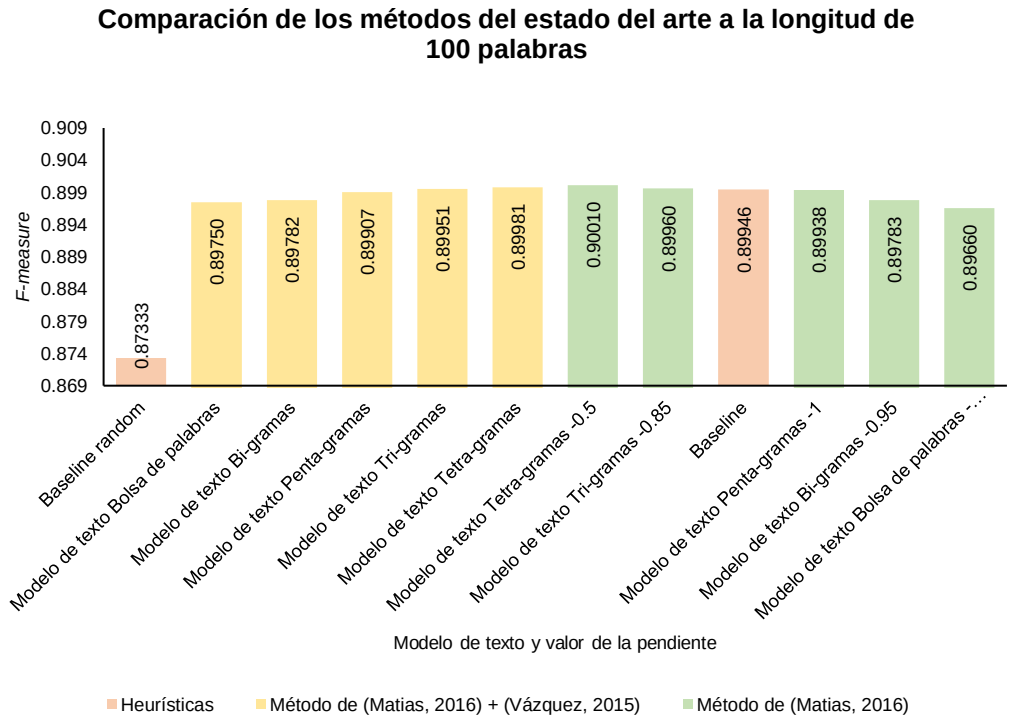

Figura 23. Comparación de los métodos de estado del arte a la longitud de 100 palabras.

Tabla 9. Configuración de parámetros para el método del estado del arte de (Matias, 2016) a longitud de 100 palabras.

Parámetros a la longitud de 100 palabras	
Pre-procesamiento	No
Modelo de texto	Tetra-gramas ó $n = 4$
Importancia de las oraciones (valor de la pendiente)	-0.5
Función de aptitud	0.5δ+0.5β
Operador de selección	Ruleta

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

En este capítulo, se dan las conclusiones de esta tesis y el trabajo futuro.

6.1 Conclusiones

En esta sección se presentan conclusiones generales:

- Se determinó el desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos en el idioma ruso usando el corpus TEXTRUSS.
- Se utilizó un módulo de transliteración para transliterar el corpus TEXTRUSS y utilizado en los métodos del estado del arte.
- Se generaron los resúmenes con los métodos del estado del arte a dos longitudes: longitud del *gold standard* y a 100 palabras.
- Se probó el método de (Matias, 2016) y el método de (Matias, 2016) +(Vázquez, 2015) con diferentes configuraciones de parámetros.
- Se obtuvieron las heurísticas del corpus TEXTRUSS.

6.2 Análisis de resultados obtenidos

- El método de (Matias, 2016) se desempeña mejor para la generación automática de resúmenes, superando a las heurísticas *Baseline random* y *Baseline*.
- En el método de (Matias, 2016) se realizaron experimentos con el operador de Ruleta y Torneo, por lo que se concluyó que el método se desempeña mejor con el operador Ruleta para la generación automática de resúmenes (ver Anexo 6).
- En la Tabla 10, se muestra la mejor configuración de parámetros del método de (Matias, 2016) a la longitud del *gold standard* y a 100 palabras.

Tabla 10. Mejor configuración de parámetros para el método del estado del arte de (Matias, 2016) a longitud del *gold standard* y a 100 palabras.

Parámetros	**A la longitud del *gold standard***	**A la longitud de 100 palabras**
Pre-procesamiento	No	No
Modelo de texto	Bolsa de palabras ó $n = 1$	Tetra-gramas ó $n = 4$
Importancia de las oraciones (valor de la pendiente)	-0.625	-0.5
Función de aptitud	0.5δ+0.5β	0.5δ+0.5β
Operador de selección	Ruleta	Ruleta

- No se recomienda utilizar herramientas comerciales ya que su desempeño es bajo en comparación con los métodos del estado del arte (ver Anexo 2 y 4), por eso se recomienda utilizar métodos del estado del arte para la generación automática de resúmenes para el idioma ruso (ver Figura 24 y 25).

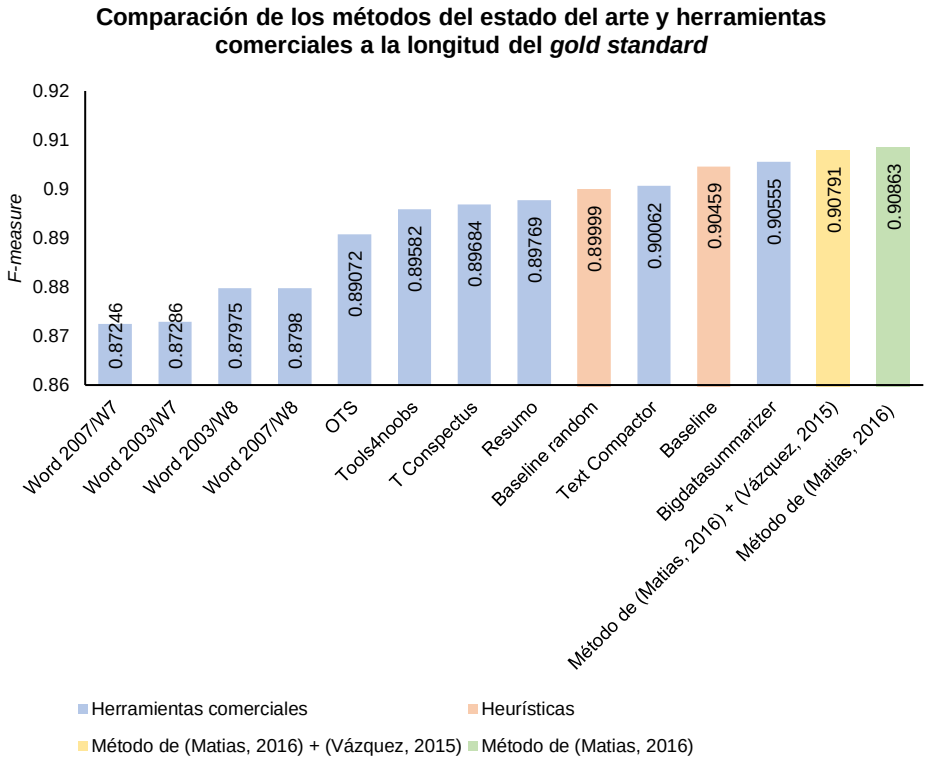

Figura 24. Comparación de los métodos del estado del arte y herramientas comerciales a la longitud del *gold standard*.

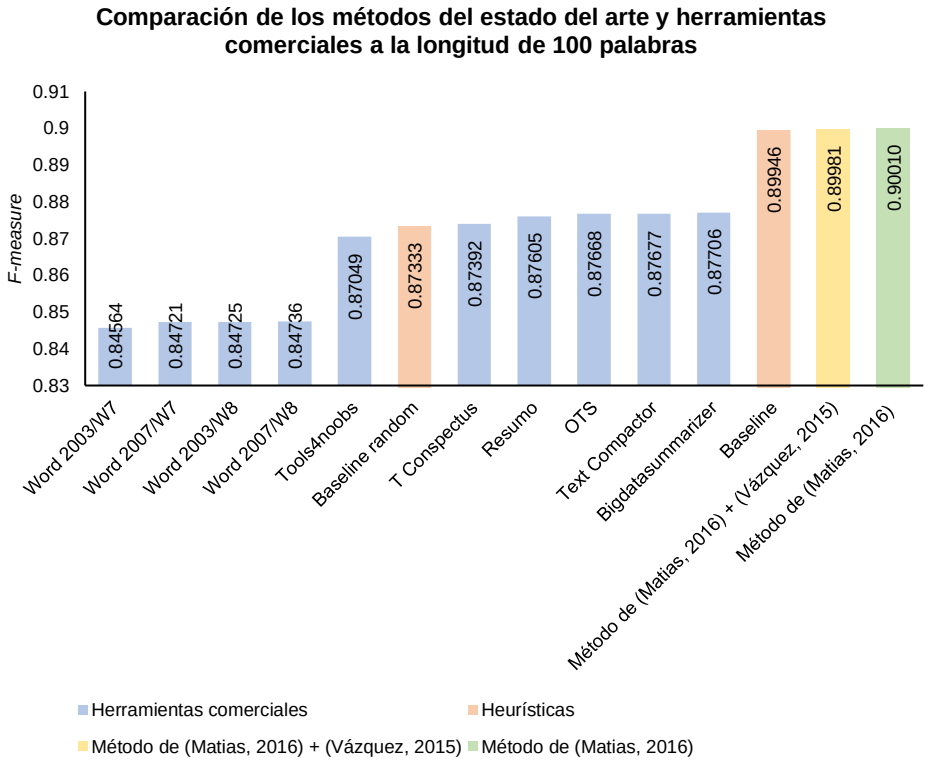

Figura 25. Comparación de los métodos del estado del arte y herramientas comerciales a la longitud de 100 palabras.

6.2 Trabajos futuros

- Agregar un segundo resumen *gold standard* al corpus TEXTRUSS, como se tiene para los corpus DUC-2002, TeMário y TER. Evaluar tomando la referencia los dos *gold standard.*
- Crear un corpus para la generación automática de resúmenes de múltiples documentos en el idioma ruso.

REFERENCIAS BIBLIOGRÁFICAS

(Alfonseca, 2003) Alfonseca E. & Pilar, R. (2003). Generating extracts with genetic algorithms. Universidad Autónoma de Madrid, España, pp. 511-519.

(Arranz, 2018) Arranz de la Peñas, J., Parra Truyol, A. (2018). Algoritmos genéticos. Universidad Carlos III. 8 páginas. Obtenido de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/05.pdf

(Barrios, 2015) Barrios, F., López, F. (2015). Módulo de resúmenes automáticos basado en TextRank con integración de GENSIM (Trabajo profesional). Universidad de Buenos Aires.

(Bigdatasum, 2017) Bigdatasummarizer (2017). Herramienta comercial. Obtenido de https://bigdatasummarizer.com/summarizer/online/advanced.jsp?ui.lang=es

(Castillo, 2009) Castillo, E. (2009). Resumen Automático. 7 páginas. Obtenido de http://resumenautomatico.50webs.com/docs/resumen_automatico.pdf

(Coello, 1995) Coello, C. (1995). Algoritmo genético y sus aplicaciones, Soluciones avanzadas. Tecnologías de información y estrategias de negocios, p.p. 5-11.

(Cortez et al., 2009) Cortez, A., Vega, H., Pariona, J. (2009). Procesamiento de Lenguaje Natural. Revista de ingeniería de sistemas e informática. Universidad Nacional Mayor de San Marcos. Vol. 6, pp. 45-54.

(DRAE, 2016) Diccionario de la Real Academia Española. (2016). Obtenido de http://dle.rae.es/?w=transliterar&origen=REDLE

(Escandell, 2016) Escandell, V., Marrero, V., Casado, C., Gutiérrez, E., Polo, N., & Ruiz-Va, P. (2016). Claves del lenguaje natural.429 páginas. Editorial Universitaria Ramón Areces. Obtenido de https://books.google.com.mx/books?id=GHWnDAAAQBAJ&pg=PA32&lpg=PA32&dq=claves+del+lenguaje+natural+escandell&source=bl&ots=9mCHBYogBk&sig=2cqBjqSneG25LGiiCxaMWGma9Yw&hl=es&sa=X&ved=0ahUKEwjEovnZsPzZAhUT7GMKHRI9BDkQ6AEIQzAE#v=onepage&q=claves%20del%20lenguaje%20natural%20escandell&f=false

(García et. al, 2009) García-Hernández, R. A., Ledeneva, Y., Matías, G., Hernández, A., Chávez, J., Gelbukh, A., & Tapia, J. L. (2009). Comparing Commercial Tools and State-of-the-Art Methods for Generating Text Summaries. Eighth Mexican International Conference on Artificial Intelligence, pp. 92-96.

(Gelbukh, 2010) Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus aplicaciones. Sociedad Mexicana de Inteligencia Artificial. ISSN 2007-0691, vol. I, pp. 6-11.

(Gestal, 2013) Gestal, M. (2013). Introducción a los algoritmos genéticos. University of A Coruña, pp. 2-16. Obtenido de https://www.researchgate.net/publication/237812449_Introduccion_a_los_Algoritmos_Geneticos

(Gonzales, 2016) González, N. (2016). Determinación del desempeño de resúmenes generados automáticamente para el idioma español (Tesis de Licenciatura). UAEMex, Estado de México.

(Hassel, 2007) Hassel, M. (2007). Resource Lean and Portable Automatic Text Summarization (Tesis doctoral). KTH School of Computer Science and Communication, Suecia.

(Hernandez et al. 2013) Hernandez, M., Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. Revista Politécnica. Vol. 32, pp. 87-96.

(Ibáñez, 2013) Ibañez, D. Y. (2013). Evaluación de las Herramientas Comerciales de generación automática de resúmenes de textos para el idioma portugués (Tesis de Licenciatura). UAEMex, Estado de México.

(INEGI, 2018) Instituto Nacional de Estadística y Geografía (2016). Estadísticas a propósito del día mundial de internet “17 de mayo”. Aguascalientes, pp. 1-13. Obtenido de http://www.inegi.org.mx/saladeprensa/aproposito/2017/internet2017_Nal.pdf

(Ledeneva & García, 2017) Ledeneva, Y. & García-Hernández, R.A. (2017). Generación automática de resúmenes. Retos, propuestas y experimentos. Automatic Generation of Text Sumaries. Challenges, Proposals and Experiments. Universidad Autónoma del Estado de México (Ed.) /Cigome, ISBN 978-607-422-782-6, (en inglés e español, 1000 ejemplares). 285 páginas.

(Ledeneva, 2008) Ledeneva, Y. N. (2008). Automatic Language-Independent Detection of Multi Word Descriptions for Text Summarization (Tesis de Doctorado). Instituto Politécnico Nacional, Ciudad de México.

(Ledeneva, 2011) Ledeneva, Y., García, R., Montiel, R., Cruz, R. & Gelbukh, A. (2011). EM Clustering Algorithm for Automatic Text Summarization. Springer-Verlag LNAI 7094, pp. 305-315.

(Lin, 2004) Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL. Barcelona, España, pp. 1-9.

(Lloret, 2008) Lloret, E., Ferrández, O., Muñoz, R., & Palomar., M. (2008). Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Sociedad Española para el procesamiento del Lenguaje Natural. España. ISSN 1135-5948, pp 183-190.

(Mateos, 2017) Mateos, A. (2017). Algoritmos evolutivos y Algoritmos genéticos. Universidad Carlos III de Madrid. N.I.A 100027597,14 páginas. Obtenido de http://www.it.uc3m.es/~jvillena/irc/practicas/estudios/ae

(Matias, 2013a) Matias, G. A. (2013a). Generación automática de resúmenes usando algoritmos Genéticos (Tesis de Licenciatura). UAEMex, Estado de México.

(Matias, 2013b) Matias, G., Ledeneva, Y., García-Hernández, R. A., & Sidorov G. (2013b). Evaluación de las herramientas comerciales y métodos del estado del arte para la generación de resúmenes extractivos individuales. Research in Computing Science, 70, pp. 265-274.

(Matias, 2016) Matias, G. A. (2016). Generación Automática de Resúmenes Independientes del Lenguaje (Tesis de Maestría). UAEMex, Estado de México.

(Mendoza, 2014) Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. Expert systems with applications, pp. 4158-4169.

(Mihalcea, 2004) Mihalcea, R. Tarau, P. (2004). TextRank: Bringing Order into Texts. Departamento of Computer Science University of North Texas, pp 1-8.

(Mihalcea, 2005) Mihalcea, R. Tarau, P. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization. Department of Computer Science and Engineering Texas. Vol. 1, pp. 602-607.

(Miranda, 2013) Miranda, S. (2013). Modelo para la generación automática de resúmenes abstractivo basado en grafos conceptuales (Tesis de doctorado). Instituto Politécnico Nacional, Ciudad de México.

(Montiel, 2009) Montiel, R. (2009). Generación automática de resúmenes mediante aprendizaje no supervisado (Tesis de licenciatura). Metepec, Edo. Mex.

(MOW, 2016) Microsoft Office Word 2003 y 2007. (2016). Herramienta commercial. Obtenido de https://www.microsoft.com/es-mx/download

(Orozco, 2016) Orozco-Aguirre, A. (2016). Las RCA2 y la transliteración de nombres de autores personales rusos. Dirección General de Biblioteca (UNAM). Obtenido de http://www.dgbiblio.unam.mx/servicios/dgb/publicdgb/bole/fulltext/volIV2/rca2.htm

(OTS, 2017) Open Text Summarizer. (2017). Herramienta commercial. Obtenido de https://www.splitbrain.org/services/ots

(Pino et al. 2001) Pino, Raúl., Gómez, Alberto., De Abajo, Nicolás., (2001). Introducción a la inteligencia Artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva. Servicio de Publicaciones, Universidad de Oviedo. Obtenido de https://books.google.com.mx/books?id=RKqLMCw3IUkC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

(Plaza, 2010) Plaza, L. (2010). Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo (Tesis de doctorado). Madrid.

(Resumo, 2017) Generador de Texto Automático – Resumo. (2017). Herramienta commercial. Obtenido de http://www.geradordetextoautomatico.com.br/Resumo/

(Rojas, 2016) Rojas, J. M. (2016). Evaluación de herramientas comerciales y métodos del estado del arte para la generación de resúmenes en idioma ruso (Tesis de Licenciatura). UAEMex, Estado de México.

(Rojas, 2017) Rojas, J. (2017). Cálculo de *TOPLINE* para la generation automática de resúmenes usando algoritmos genéticos. (Tesis de Licenciatura). UAEMex, Estado de México.

(Strutunnof, 2012) Strutunnof, I. (2012). Gramática del ruso. De Vecchi Ediciones. Barcelona. Obtenido de https://books.google.com.mx/books?id=al-8BNVXUZsC&pg=PA12&dq=Strutunnof+alfabeto+ruso&hl=es&sa=X&ved=0ahUKEwjpvt-J4dvSAhUQ-GMKHaLqDTAQ6AEIGjAA#v=onepage&q=Strutunnof%20alfabeto%20ruso&f=false

(T4N, 2017) Tools4noobs. (2017). Herramienta commercial. Obtenido de https://www.tools4noobs.com/summarize/

(TCom, 2017) Text Compactor. (2017). Herramienta commercial. Obtenido de http://textcompactor.com/

(TCons, 2017) T Conspectus. (2017). Herramienta commercial. Obtenido de http://tconspectus.pythonanywhere.com/summarization

(Translit, 2016) Translit.ru (2016). Obtenido de http://translit.net/

(Vázquez, 2015) Vázquez, E. (2015). Modelo de relevancia de la posición de las oraciones en resúmenes de texto, mediante regresión simbólica. (Tesis de Licenciatura). UAEMex, Estado de México.

(Vlainic, 2013) Vlainić, M., Mikelić, N. (2013). A Comparative Study of Automatic Text Summarization System Performance. Recent Advances in Information Science. ISBN: 978-960-474-304-9, pp. 222- 227.

ANEXOS

Anexo 1 Descripción de las herramientas comerciales

Las siguientes herramientas comerciales son utilizadas para la generación automática de resúmenes.

Herramientas en línea

Bigdatasummarizer (Bigdatasum, 2017)

Es una de las nuevas herramientas que se implementó en este trabajo para generar resúmenes en idioma ruso.

Es una herramienta que realiza resúmenes de textos en 21 idiomas como: chino, inglés, francés, alemán, italiano, ruso, español, etc. Trabaja con 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% y 100% de umbral.

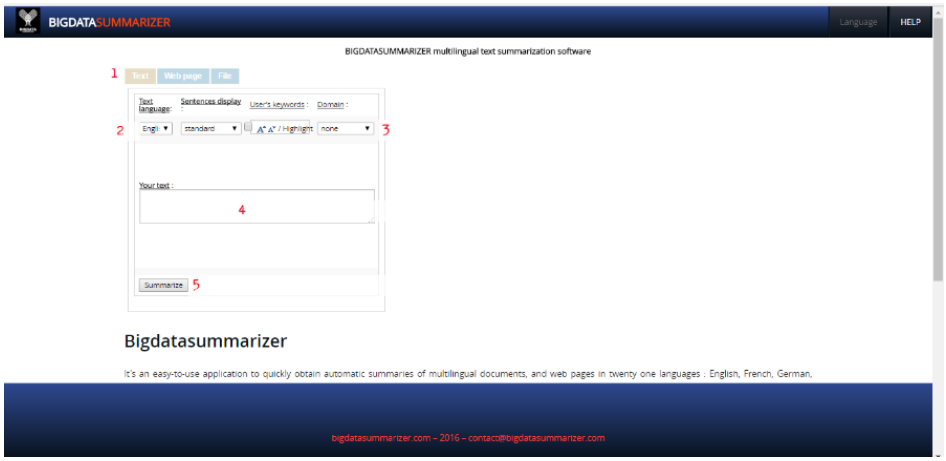

Figura 26. Herramienta comercial en línea Bigdatasummarizer.

Para realizar los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Tiene la opción de ingresar el texto de forma manual, a través de un link o subiendo un documento.
2. Seleccionar el idioma del texto origen.
3. Seleccionar el dominio del texto.
4. Si ingresas el texto de forma manual está el espacio para pegar el texto.

5. El resumen es generado presionando el botón Summarizer.

Open Text Summarizer (OTS, 2017)

Esta herramienta analiza automáticamente textos en varios idiomas e intenta identificar las partes más importantes del texto.

Sólo se tiene que pegar el texto o cargarlo desde una URL para obtenerlo resumido.

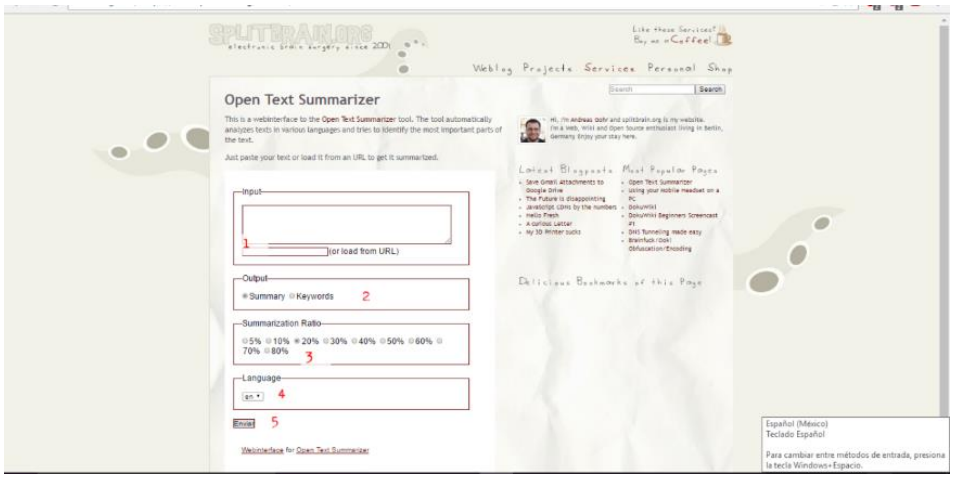

Figura 27. Herramienta comercial en línea Open Text Summarizer.

Para realizar los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Tiene la opción de ingresar el texto de forma manual o a través de un link.
2. Tiene la opción de seleccionar la tarea ya sea resumir texto o que genere palabras clave.
3. Selecciona el porcentaje de 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% de umbral.
4. Seleccionar el idioma del texto origen.
5. El resumen es generado presionando el botón Enviar.

T Conspectus (TCons, 2017)

Es una aplicación web para resumir artículos en inglés, alemán y ruso.

Ofrece servicio de resumen automático multilingüe de artículos de noticias. El resumen utiliza algunas técnicas de PNL para extraer automáticamente las frases más informativas de un texto sin formato insertado en el cuadro de texto, cargado por el usuario o desde una URL.

Adicionalmente, basándose en algunos módulos del resumen, se pueden utilizar otros dos servicios para el procesamiento de texto como la segmentación de texto y la ponderación de frecuencia de término.

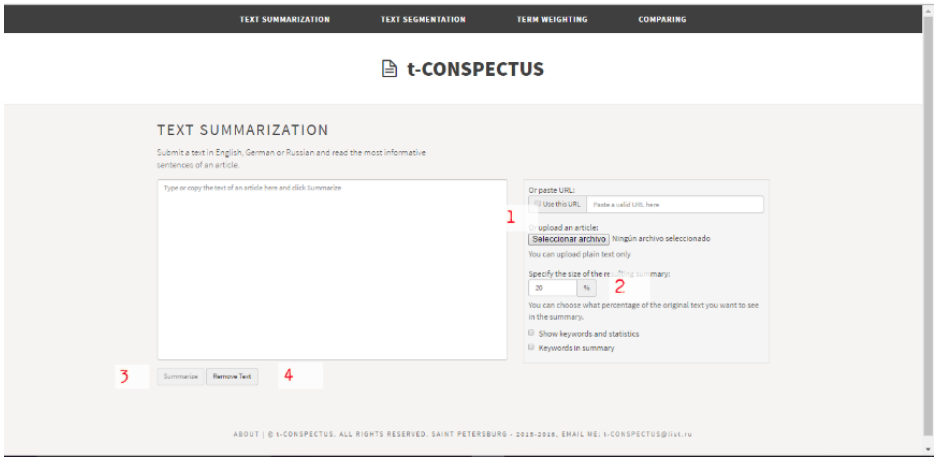

Figura 28. Herramienta comercial en línea T Conspectus.

Para realizar la tares de los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Tiene la opción de ingresar el texto de forma manual, a través de una URL o subiendo el archivo.
2. Selecciona el umbral 5%, 10%,15%, 20%, 25%, …,70%.
3. El resumen es generado presionando el botón Summarize.
4. También cuenta con la opción de eliminar el texto presionando el botón Remove Text.

Text Compactor (TCom, 2017)

Fue creado por Keith Edyburn para Knowledge by Design, Inc. Se basó en Open Text Summarizer.

Esta herramienta gratuita para resumir textos en línea fue creada para ayudar a los lectores a procesar cantidades abrumadoras de información. Sin embargo, el enfoque general ayudará a cualquier estudiante, maestro o profesionista.

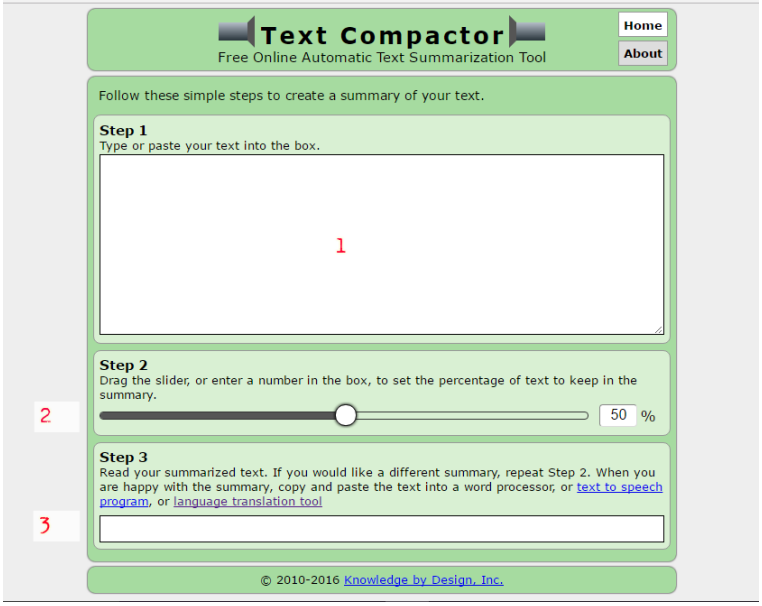

Figura 29. Herramienta comercial en línea Text Compactor.

Para realizar la tares de los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Ingresar los textos de forma manual.
2. Colocar un porcentaje para generar el resumen.
3. Muestra el resumen del texto.

Tools4noobs (T4N, 2017)

Esta herramienta comercial realiza varias tareas aparte de la generación de resúmenes: muestra la relevancia de la oración, las mejores palabras, palabras clave y oraciones. Soporta textos en formato UTF-8.

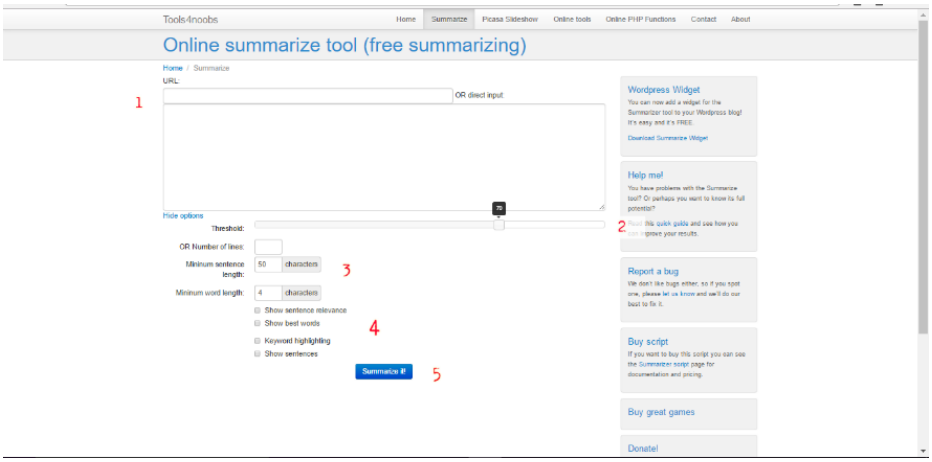

Figura 30. Herramienta comercial en línea Tool4noobs.

Para realizar la tares de los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Ingresar los textos de forma manual o mediante una URL.
2. Colocar un límite o porcentaje para generar el resumen.
3. Si es necesario se puede modificar número de líneas, longitud mínima de oraciones, longitud mínima de palabras.
4. Se puede seleccionar una tarea distinta a la de resúmenes como: relevancia de la oración, mejores palabras, palabras clave y oraciones.
5. Por último, resumir el texto presionando el botón nombrado *Summarize it!*

Resumo (Resumo, 2017)

Esta herramienta es un generador de resúmenes de textos, la interfaz es en idioma portugués. Es una de las nuevas herramientas que se implementó en este trabajo para generar resúmenes en idioma ruso. Tiene la opción para generar resúmenes a diferentes porcentajes. En la Figura 31 se puede apreciar la interfaz y más adelante se describe los pasos para generar un resumen.

Figura 31. Herramienta comercial en línea Generador de Texto Automático/Resumo.

Para realizar la tares de los resúmenes con esta herramienta se siguen los siguientes pasos:

1. Insertar texto para resumirlo.
2. Seleccionar el umbral (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% e indiferente).
3. Seleccionar el número de líneas (1, 2, 3 ,4 ,5 e indiferente) al cual se va a resumir.
4. Seleccionamos el botón con nombre “CREAR RESUMO”.
5. Área donde aparece el resumen del texto original resumido.

Herramientas Instalables

Microsoft Office Word (MOW, 2016)

Es una herramienta comercial editor de textos que contiene un sin fin de tareas para el procesamiento de texto, a continuación, se describen los pasos para generar resúmenes en las versiones de Microsoft Office Word 2003 y 2007.

Microsoft Office Word 2003

Para esta versión los pasos con los siguientes:

1. Seleccionamos la pestaña de Herramientas y de ahí nos deslizamos a la opción de Autorresumen.
2. A continuación, aparecerá una pantalla que está conformado en dos secciones. Primeramente, el tipo de resumen como se quiera generar: Resaltar puntos principales, Insertar un resumen ejecutivo o extractor al principio del documento, Crear un documento nuevo para colocar el resumen, Ocultar todo excepto el resumen sin salir del documento original.
3. En seguida se selecciona un porcentaje para generar el resumen las opciones que tiene son: 10 oraciones, 20 oraciones, 100 palabras o menos, 500 palabras o menos, 10 %, 25%, 50% y 75%. En este apartado más que poder seleccionar una de estas opciones predeterminadas podemos modificar su valor introduciendo una desde teclado, ya que para este trabajo de tesis se realizaron los resúmenes a 100 palabras y a la longitud del *gold standard*.

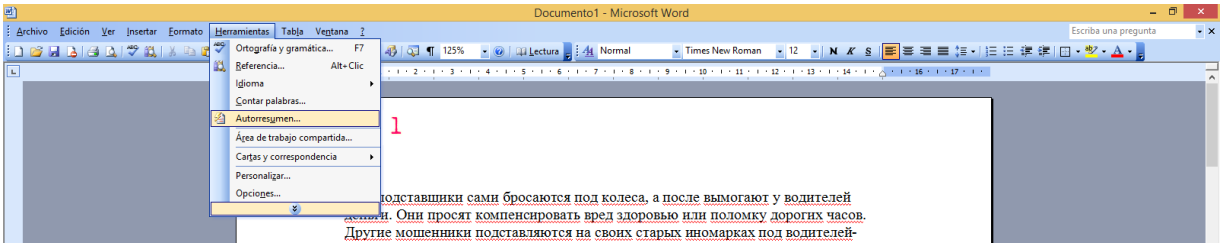

Figura 32. Interfaz con la opción de Autorresumen en Microsoft Office Word 2003.

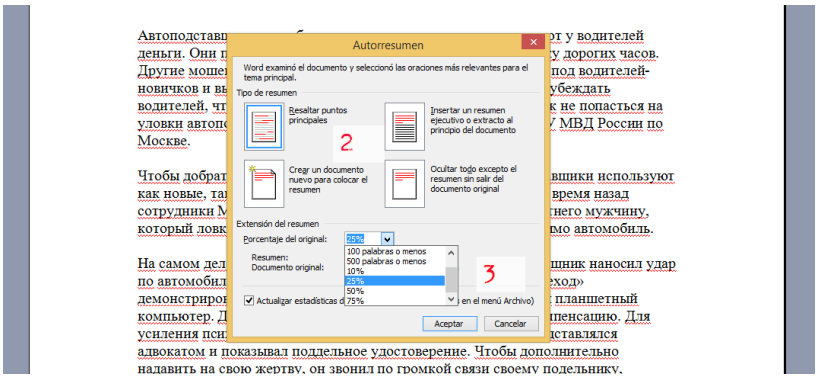

Figura 33. Interfaz con los parámetros del Autorresumen en Microsoft Office Word 2003.

Microsoft Office Word 2007

Para esta versión los pasos son los siguientes:

Una vez abierto el Office Word procedemos a colocar el acceso rápido Resumen automático en la barra de herramientas.

1. En el botón de inicio nos dirigimos a Opciones de Word.
2. En la nueva ventana del lado izquierdo nos dirigimos a la opción de Personalizar.
3. En la pestaña que dice Comandos disponibles seleccionamos Todos los comandos.

Seleccionamos la que dice Herramientas de resumen automático >> Agregar >> valga la redundancia se agregue a la barra de herramientas y por último le damos en el botón Aceptar que está ubicado en la parte inferior derecha.

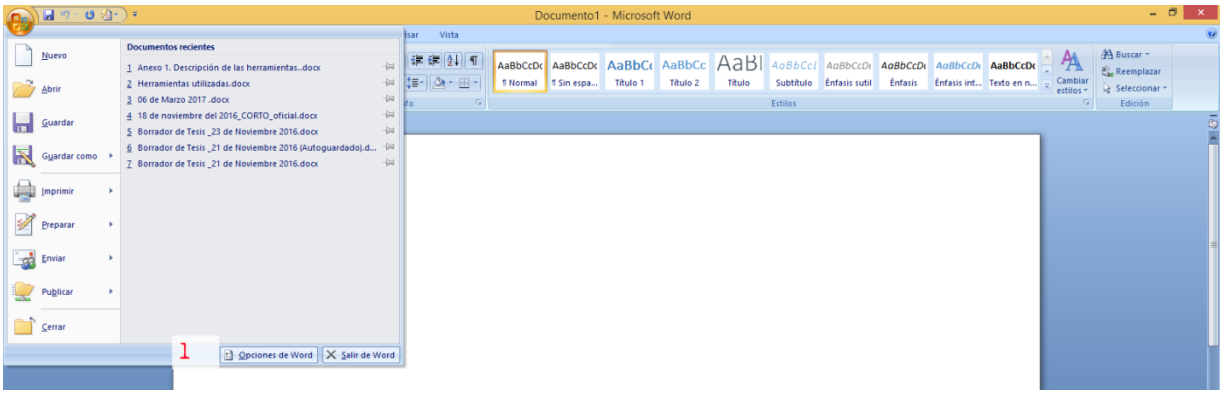

Figura 34. Interfaz para activar Autorresumen en Microsoft Office Word 2007.

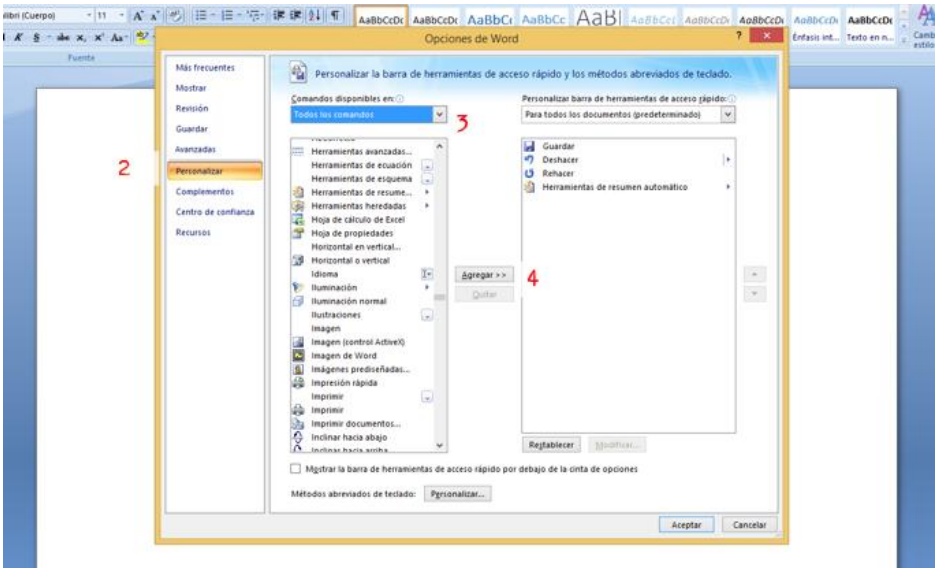

Figura 35. Interfaz que muestra el proceso para activar Autorresumen en Microsoft Office Word 2007.

Una vez colocado el acceso directo procedemos a realizar los resúmenes y los pasos son los siguientes:

1. El icono aparecerá en la parte superior izquierda.
2. A continuación, aparecerá una pantalla que está conformado en dos secciones. Primeramente, el tipo de resumen como se quiera generar: Resaltar puntos principales, Insertar un resumen ejecutivo o extractor al principio del documento, Crear un documento nuevo para colocar el resumen, Ocultar todo excepto el resumen sin salir del documento original.

3. En seguida seleccionar un porcentaje para generar el resumen las opciones que tiene son: 10 oraciones, 20 oraciones, 100 palabras o menos, 500 palabras o menos, 10 %, 25%, 50% y 75%. Recalcando que en este apartado más que poder seleccionar una de estas opciones predeterminadas podemos modificar su valor introduciendo una desde teclado, ya que para este trabajo de tesis se realizaron los resúmenes a 100 palabras y a la longitud del *gold standard*.

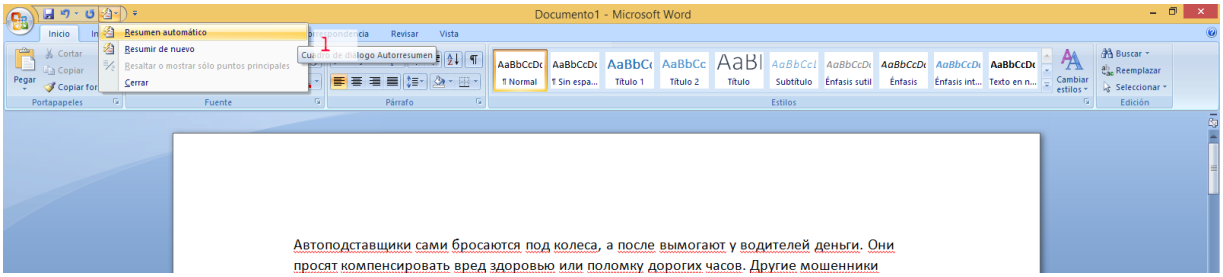

Figura 36. Interfaz con la opción de Autorresumen en Microsoft Office Word 2007.

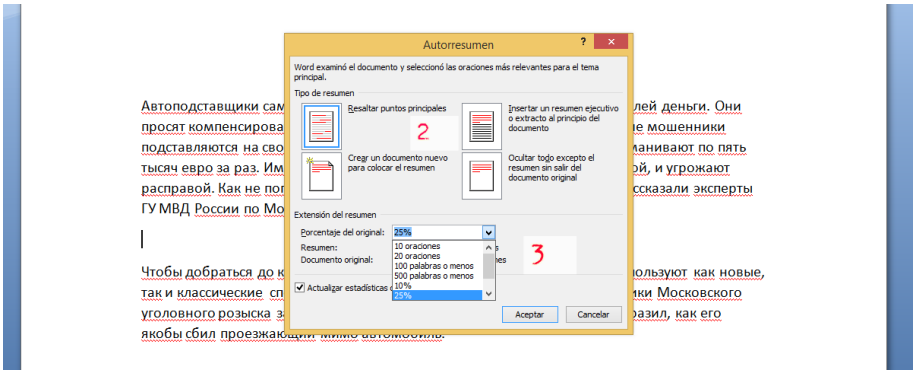

Figura 37. Interfaz con los parámetros del Autorresumen en Microsoft Office Word 2007.

Anexo 2 Resultados de las herramientas comerciales a la longitud del *gold standard*

A continuación, se presentan la evaluación por ROUGE de las herramientas comerciales a la longitud del *gold standard*, mostrando el promedio de los resultados en tablas y en gráficas.

Herramientas comerciales en línea

En la Tabla 11, se observan los resultados de las herramientas comerciales en línea a la longitud del *gold standard*, el mejor resultado que se obtuvo fue con la herramienta *Bigdatasummarizer* con un desempeño de 0.90555 (ver Figura 38).

Tabla 11. Evaluación de las herramientas comerciales en línea a la longitud del *gold standard*.

Nº de experimento	Herramienta comercial en línea	*F-measure*
1	*Bigdatasummarizer*	0.90555
2	*Text compactor*	0.90062
3	*Resumo*	0.89769
4	*T Conspectus*	0.89684
5	*Tool4noobs*	0.89582
6	*Open Text Summarizer (OTS)*	0.89072

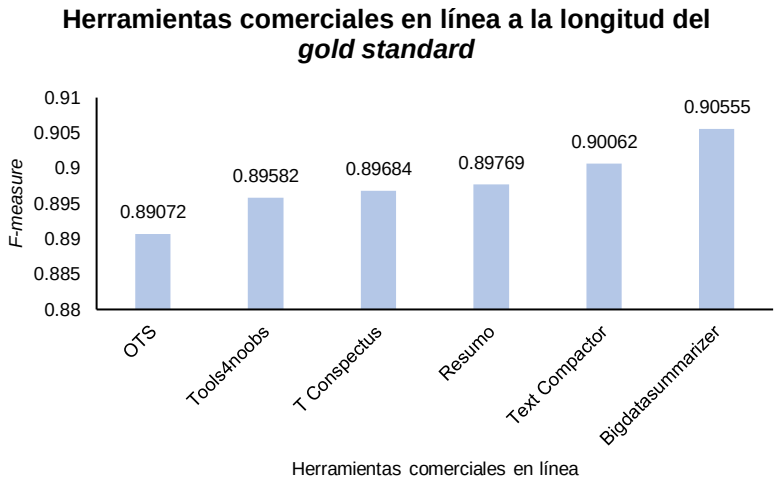

Figura 38. Resultados de las herramientas comerciales en línea a la longitud del *gold standard*.

Herramientas comerciales instalables

En la Tabla 12, se observan los resultados de las herramientas comerciales instalables a la longitud del *gold standard*, el mejor resultado que se obtuvo fue con la herramienta *Microsoft Office Word* 2007 con el sistema operativo Windows 8 con un desempeño de 0.87980 (ver Figura 39).

Tabla 12. Evaluación de las herramientas comerciales instalables a la longitud del *gold standard*.

Nº de experimento	Herramienta comercial instalable	Sistema operativo en el que se ejecuto	*F-measure*
7	*Microsoft Office Word 2007*	Windows 8	0.87980
8	*Microsoft Office Word 2003*	Windows 8	0.87975
9	*Microsoft Office Word 2003*	Windows 7	0.87286
10	*Microsoft Office Word 2007*	Windows 7	0.87246

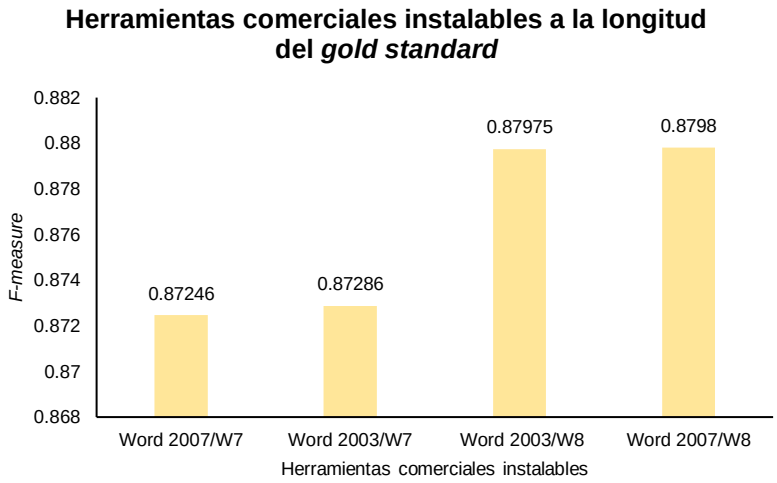

Figura 39. Resultados de las herramientas comerciales instalables a la longitud del *gold standard*.

En la Figura 40, se comparan las herramientas comerciales a la longitud del *gold standard*, la herramienta comercial en línea *Bigdatasummarizer* obtiene el mejor desempeño para la generación automática de resúmenes obteniendo 0.90555.

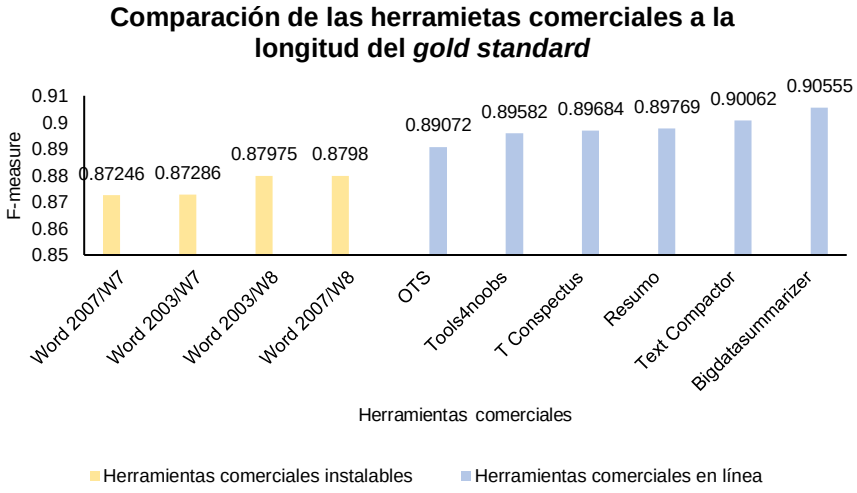

Figura 40. Comparación de las herramientas comerciales a la longitud del *gold standard*.

Anexo 3 Promedio de los métodos del estado del arte a la longitud del *gold standard*

A continuación, se presentan la evaluación por ROUGE de los métodos del estado del arte a la longitud del *gold standard*, mostrando el promedio de los resultados en la siguiente tabla (ver Tabla 13).

Tabla 13. Evaluación del método de (Matias, 2016) con modelo de texto y valor de la pendiente a la longitud del *gold standard*.

N° de experimento	Modelo de texto	Valor de la pendiente	*F-measure*
1	Bolsa de palabras	-0.25	0.90676
2	Bi-gramas	-0.25	0.90535
3	Tri-gramas	-0.25	0.90525
4	Tetra-gramas	-0.25	0.90538
5	Penta-gramas	-0.25	0.90519
6	Bolsa de palabras	-0.3	0.88706
7	Bi-gramas	-0.3	0.90625
8	Tri-gramas	-0.3	0.90546
9	Tetra-gramas	-0.3	0.90534
10	Penta-gramas	-0.3	0.90536
11	Bolsa de palabras	-0.375	0.90809
12	Bi-gramas	-0.375	0.90633
13	Tri-gramas	-0.375	0.90537
14	Tetra-gramas	-0.375	0.90541
15	Penta-gramas	-0.375	0.90552
16	Bolsa de palabras	-0.45	0.90800
17	Bi-gramas	-0.45	0.90746
18	Tri-gramas	-0.45	0.90558
19	Tetra-gramas	-0.45	0.90590
20	Penta-gramas	-0.45	0.90533
21	Bolsa de palabras	-0.5	0.90751
22	Bi-gramas	-0.5	0.90707
23	Tri-gramas	-0.5	0.90544
24	Tetra-gramas	-0.5	0.90552
25	Penta-gramas	-0.5	0.90500

26	Bolsa de palabras	-0.55	0.90771
27	Bi-gramas	-0.55	0.90692
28	Tri-gramas	-0.55	0.90562
29	Tetra-gramas	-0.55	0.90517
30	Penta-gramas	-0.55	0.90552
31	Bolsa de palabras	-0.6	0.90752
32	Bi-gramas	-0.6	0.90668
33	Tri-gramas	-0.6	0.90558
34	Tetra-gramas	-0.6	0.90562
35	Penta-gramas	-0.6	0.90524
36	Bolsa de palabras	-0.625	0.90863
37	Bi-gramas	-0.625	0.90615
38	Tri-gramas	-0.625	0.90533
39	Tetra-gramas	-0.625	0.90539
40	Penta-gramas	-0.625	0.90499
41	Bolsa de palabras	-0.65	0.90757
42	Bi-gramas	-0.65	0.90728
43	Tri-gramas	-0.65	0.90647
44	Tetra-gramas	-0.65	0.90644
45	Penta-gramas	-0.65	0.90640
46	Bolsa de palabras	-0.7	0.90719
47	Bi-gramas	-0.7	0.90715
48	Tri-gramas	-0.7	0.90486
49	Tetra-gramas	-0.7	0.90533
50	Penta-gramas	-0.7	0.90499
51	Bolsa de palabras	-0.75	0.90775
52	Bi-gramas	-0.75	0.90661
53	Tri-gramas	-0.75	0.90502
54	Tetra-gramas	-0.75	0.90572
55	Penta-gramas	-0.75	0.90471
56	Bolsa de palabras	-0.8	0.90794
57	Bi-gramas	-0.8	0.90691
58	Tri-gramas	-0.8	0.90552
59	Tetra-gramas	-0.8	0.90571
60	Penta-gramas	-0.8	0.90517
61	Bolsa de palabras	-0.85	0.90759

62	Bi-gramas	-0.85	0.90696
63	Tri-gramas	-0.85	0.90556
64	Tetra-gramas	-0.85	0.90532
65	Penta-gramas	-0.85	0.90511
66	Bolsa de palabras	-0.9	0.90760
67	Bi-gramas	-0.9	0.90690
68	Tri-gramas	-0.9	0.90558
69	Tetra-gramas	-0.9	0.90585
70	Penta-gramas	-0.9	0.90501
71	Bolsa de palabras	-0.95	0.90790
72	Bi-gramas	-0.95	0.90666
73	Tri-gramas	-0.95	0.90541
74	Tetra-gramas	-0.95	0.90582
75	Penta-gramas	-0.95	0.90503
76	Bolsa de palabras	-1	0.90760
77	Bi-gramas	-1	0.90680
78	Tri-gramas	-1	0.90631
79	Tetra-gramas	-1	0.90645
80	Penta-gramas	-1	0.90649

En la siguiente tabla se muestran los resultados del método de (Matias, 2016) + (Vázquez, 2015) (ver Tabla 14) a la longitud del *gold standard*.

Tabla 14. Evaluación del método de (Matias, 2016) + (Vázquez, 2015) a la longitud del *gold standard*.

N° de experimento	Modelo de texto	*F-measure*
1	Bolsa de palabras	0.90791
2	Bi-gramas	0.90652
3	Tri-gramas	0.90588
4	Tetra-gramas	0.90559
5	Penta-gramas	0.90577
6	Bolsa de palabras	0.90760
7	Bi-gramas	0.90623
8	Tri-gramas	0.90527
9	Tetra-gramas	0.90561
10	Penta-gramas	0.90546

Anexo 4 Resultados de las herramientas comerciales a la longitud de 100 palabras

A continuación, se presentan la evaluación por ROUGE de las herramientas comerciales a la longitud de 100 palabras.

Herramientas comerciales en línea

A continuación, se observan los resultados de las herramientas comerciales en línea a 100 palabras, el mejor resultado que se obtuvo fue con la herramienta *Bigdatasummarizer* con un desempeño de 0.87706 (ver Figura 41).

Experimento 1

Bigdatasummarizer

---

1 ROUGE-1 Average_R: 0.87472 (95%-conf.int. 0.86830 - 0.88105)

1 ROUGE-1 Average_P: 0.88329 (95%-conf.int. 0.87694 - 0.88880)

1 ROUGE-1 Average_F: 0.87706 (95%-conf.int. 0.87351 - 0.88038)

---

1 ROUGE-2 Average_R: 0.61390 (95%-conf.int. 0.60583 - 0.62130)

1 ROUGE-2 Average_P: 0.62022 (95%-conf.int. 0.61227 - 0.62843)

1 ROUGE-2 Average_F: 0.61570 (95%-conf.int. 0.60867 - 0.62267)

---

1 ROUGE-SU4 Average_R: 0.76205 (95%-conf.int. 0.75550 - 0.76841)

1 ROUGE-SU4 Average_P: 0.76967 (95%-conf.int. 0.76303 - 0.77536)

1 ROUGE-SU4 Average_F: 0.76413 (95%-conf.int. 0.75982 - 0.76844)

Experimento 2

Text Compactor

---

1 ROUGE-1 Average_R: 0.91096 (95%-conf.int. 0.90597 - 0.91568)

1 ROUGE-1 Average_P: 0.84839 (95%-conf.int. 0.84208 - 0.85441)

1 ROUGE-1 Average_F: 0.87677 (95%-conf.int. 0.87348 - 0.87989)

---

1 ROUGE-2 Average_R: 0.65044 (95%-conf.int. 0.64180 - 0.65831)

1 ROUGE-2 Average_P: 0.60533 (95%-conf.int. 0.59748 - 0.61329)

1 ROUGE-2 Average_F: 0.62580 (95%-conf.int. 0.61842 - 0.63341

---

1 ROUGE-SU4 Average_R: 0.79718 (95%-conf.int. 0.79095 - 0.80318)

1 ROUGE-SU4 Average_P: 0.74183 (95%-conf.int. 0.73558 - 0.74796)

1 ROUGE-SU4 Average_F: 0.76691 (95%-conf.int. 0.76235 - 0.77145)

Experimento 3

Open Text Summarizer

---

1 ROUGE-1 Average_R: 0.91855 (95%-conf.int. 0.91401 - 0.92307)

1 ROUGE-1 Average_P: 0.84139 (95%-conf.int. 0.83452 - 0.84813)

1 ROUGE-1 Average_F: 0.87668 (95%-conf.int. 0.87275 - 0.88060)

---

1 ROUGE-2 Average_R: 0.66910 (95%-conf.int. 0.66041 - 0.67762)

1 ROUGE-2 Average_P: 0.61332 (95%-conf.int. 0.60365 - 0.62244)

1 ROUGE-2 Average_F: 0.63884 (95%-conf.int. 0.63013 - 0.64676)

---

1 ROUGE-SU4 Average_R: 0.80974 (95%-conf.int. 0.80405 - 0.81558)

1 ROUGE-SU4 Average_P: 0.74131 (95%-conf.int. 0.73418 - 0.74852)

1 ROUGE-SU4 Average_F: 0.77259 (95%-conf.int. 0.76696 - 0.77774)

Experimento 4

Resumo

---

1 ROUGE-1 Average_R: 0.91241 (95%-conf.int. 0.90637 - 0.91785)

1 ROUGE-1 Average_P: 0.84627 (95%-conf.int. 0.83931 - 0.85345)

1 ROUGE-1 Average_F: 0.87605 (95%-conf.int. 0.87209 - 0.87984)

---

1 ROUGE-2 Average_R: 0.66663 (95%-conf.int. 0.65620 - 0.67671)

1 ROUGE-2 Average_P: 0.61769 (95%-conf.int. 0.60831 - 0.62744)

1 ROUGE-2 Average_F: 0.63976 (95%-conf.int. 0.63099 - 0.64837)

---

1 ROUGE-SU4 Average_R: 0.80607 (95%-conf.int. 0.79895 - 0.81305)

1 ROUGE-SU4 Average_P: 0.74687 (95%-conf.int. 0.73994 - 0.75429)

1 ROUGE-SU4 Average_F: 0.77351 (95%-conf.int. 0.76805 - 0.77892)

Experimento 5

T Conspectus

---

1 ROUGE-1 Average_R: 0.90819 (95%-conf.int. 0.90246 - 0.91387)

1 ROUGE-1 Average_P: 0.84590 (95%-conf.int. 0.83813 - 0.85328)

1 ROUGE-1 Average_F: 0.87392 (95%-conf.int. 0.86969 - 0.87797)

---

1 ROUGE-2 Average_R: 0.66000 (95%-conf.int. 0.65094 - 0.66903)

1 ROUGE-2 Average_P: 0.61512 (95%-conf.int. 0.60486 - 0.62444)

1 ROUGE-2 Average_F: 0.63532 (95%-conf.int. 0.62673 - 0.64380)

---

1 ROUGE-SU4 Average_R: 0.79917 (95%-conf.int. 0.79268 - 0.80567)

1 ROUGE-SU4 Average_P: 0.74405 (95%-conf.int. 0.73560 - 0.75195)

1 ROUGE-SU4 Average_F: 0.76883 (95%-conf.int. 0.76294 - 0.77449)

Experimento 6

Tools4noobs

---

1 ROUGE-1 Average_R: 0.90735 (95%-conf.int. 0.90187 - 0.91278)

1 ROUGE-1 Average_P: 0.84022 (95%-conf.int. 0.83327 - 0.84723)

1 ROUGE-1 Average_F: 0.87049 (95%-conf.int. 0.86679 - 0.87437)

---

1 ROUGE-2 Average_R: 0.63754 (95%-conf.int. 0.62871 - 0.64649)

1 ROUGE-2 Average_P: 0.59032 (95%-conf.int. 0.58188 - 0.59880)

1 ROUGE-2 Average_F: 0.61160 (95%-conf.int. 0.60356 - 0.61987)

---

1 ROUGE-SU4 Average_R: 0.79051 (95%-conf.int. 0.78404 - 0.79702)

1 ROUGE-SU4 Average_P: 0.73153 (95%-conf.int. 0.72449 - 0.73841)

1 ROUGE-SU4 Average_F: 0.75810 (95%-conf.int. 0.75301 - 0.76315)

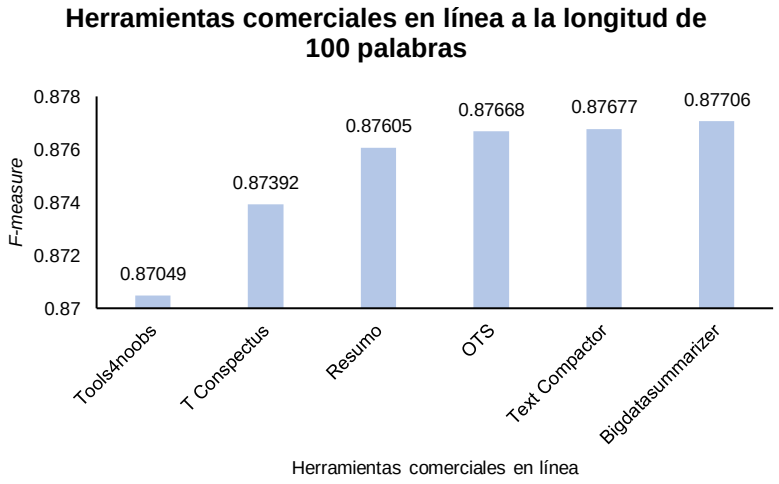

Figura 41. Resultados de las herramientas comerciales en línea a la longitud de 100 palabras.

Herramientas comerciales instalables

A continuación, se observan los resultados de las herramientas comerciales instalables a 100 palabras, el mejor resultado que se obtuvo fue con la herramienta *Microsoft Office Word* 2007 con el sistema operativo Windows 8 con un desempeño de 0.84736 (ver Figura 42).

Experimento 7

Microsoft Office Word 2003 en el sistema operativo Windows 7

---

1 ROUGE-1 Average_R: 0.94068 (95%-conf.int. 0.93686 - 0.94445)
1 ROUGE-1 Average_P: 0.77221 (95%-conf.int. 0.76269 - 0.78146)
1 ROUGE-1 Average_F: 0.84564 (95%-conf.int. 0.83970 - 0.85154)

---

1 ROUGE-2 Average_R: 0.67188 (95%-conf.int. 0.66363 - 0.68075)
1 ROUGE-2 Average_P: 0.55180 (95%-conf.int. 0.54169 - 0.56191)
1 ROUGE-2 Average_F: 0.60418 (95%-conf.int. 0.59588 - 0.61323)

---

1 ROUGE-SU4 Average_R: 0.82626 (95%-conf.int. 0.82075 - 0.83217)
1 ROUGE-SU4 Average_P: 0.67716 (95%-conf.int. 0.66787 - 0.68657)
1 ROUGE-SU4 Average_F: 0.74208 (95%-conf.int. 0.73569 - 0.74885)

Experimento 8

Microsoft Office Word 2007 en el sistema operativo Windows 7

---

1 ROUGE-1 Average_R: 0.94064 (95%-conf.int. 0.93686 - 0.94484)
1 ROUGE-1 Average_P: 0.77426 (95%-conf.int. 0.76548 - 0.78258)
1 ROUGE-1 Average_F: 0.84721 (95%-conf.int. 0.84201 - 0.85220)

---

1 ROUGE-2 Average_R: 0.67152 (95%-conf.int. 0.66280 - 0.68048)
1 ROUGE-2 Average_P: 0.55252 (95%-conf.int. 0.54326 - 0.56202)

1 ROUGE-2 Average_F: 0.60470 (95%-conf.int. 0.59636 - 0.61337)

---

1 ROUGE-SU4 Average_R: 0.82600 (95%-conf.int. 0.82063 - 0.83201)

1 ROUGE-SU4 Average_P: 0.67864 (95%-conf.int. 0.67001 - 0.68661)

1 ROUGE-SU4 Average_F: 0.74318 (95%-conf.int. 0.73742 - 0.74909)

Experimento 9

Microsoft Office Word 2003 en el sistema operativo Windows 8

---

1 ROUGE-1 Average_R: 0.94055 (95%-conf.int. 0.93667 - 0.94449)

1 ROUGE-1 Average_P: 0.77438 (95%-conf.int. 0.76601 - 0.78298)

1 ROUGE-1 Average_F: 0.84725 (95%-conf.int. 0.84206 - 0.85247)

---

1 ROUGE-2 Average_R: 0.67162 (95%-conf.int. 0.66321 - 0.68026)

1 ROUGE-2 Average_P: 0.55278 (95%-conf.int. 0.54348 - 0.56146)

1 ROUGE-2 Average_F: 0.60490 (95%-conf.int. 0.59623 - 0.61295)

---

1 ROUGE-SU4 Average_R: 0.82594 (95%-conf.int. 0.82041 - 0.83181)

1 ROUGE-SU4 Average_P: 0.67878 (95%-conf.int. 0.67061 - 0.68668)

1 ROUGE-SU4 Average_F: 0.74323 (95%-conf.int. 0.73753 - 0.74877)

Experimento 10

Microsoft Office Word 2007 en el sistema operativo Windows 8

---

1 ROUGE-1 Average_R: 0.94060 (95%-conf.int. 0.93692 - 0.94482)

1 ROUGE-1 Average_P: 0.77453 (95%-conf.int. 0.76576 - 0.78272)

1 ROUGE-1 Average_F: 0.84736 (95%-conf.int. 0.84209 - 0.85233)

---

1 ROUGE-2 Average_R: 0.67172 (95%-conf.int. 0.66300 - 0.68084)

1 ROUGE-2 Average_P: 0.55292 (95%-conf.int. 0.54347 - 0.56246)

1 ROUGE-2 Average_F: 0.60503 (95%-conf.int. 0.59646 - 0.61375)

---

1 ROUGE-SU4 Average_R: 0.82604 (95%-conf.int. 0.82062 - 0.83216)

1 ROUGE-SU4 Average_P: 0.67894 (95%-conf.int. 0.67037 - 0.68688)

1 ROUGE-SU4 Average_F: 0.74338 (95%-conf.int. 0.73733 - 0.74917)

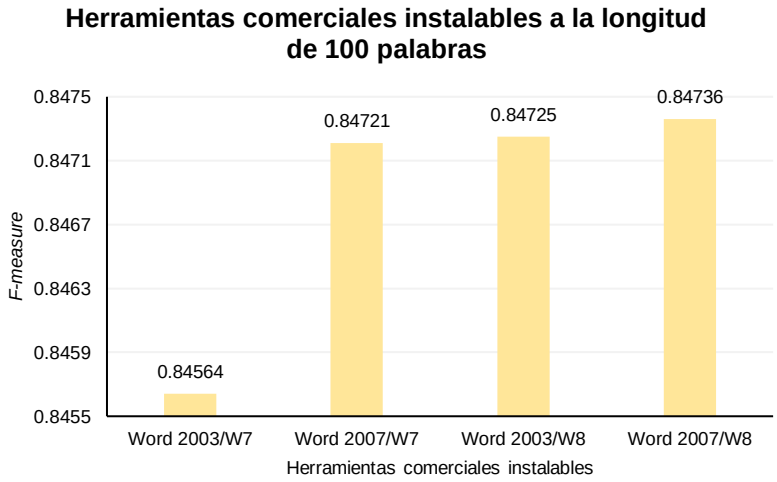

Figura 42. Resultados de las herramientas comerciales instalables a la longitud de 100 palabras.

En la Figura 43, se comparan las herramientas comerciales a 100 palabras, la herramienta comercial en línea *Bigdatasummarizer* obtiene el mejor desempeño para la generación automática de resúmenes obteniendo 0.87706.

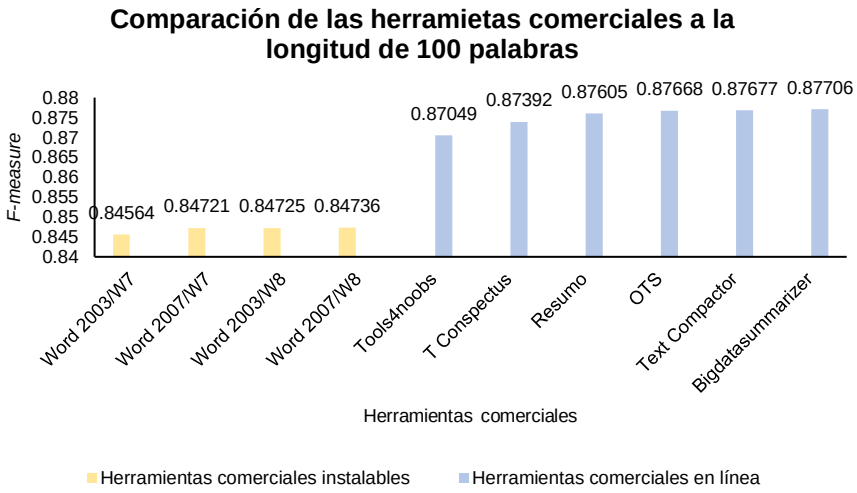

Figura 43. Comparación de las herramientas comerciales a la longitud de 100 palabras.

Anexo 5 Resultados de los métodos del estado del arte a la longitud de 100 palabras

A continuación, se presentan la evaluación por ROUGE de los métodos del estado del arte a la longitud de 100 palabras (ver Tabla 15).

Tabla 15. Evaluación del método de (Matias, 2016) con modelo de texto y valor de la pendiente a la longitud de 100 palabras.

N° de experimento	Modelo de texto	Valor de la pendiente	*F-measure*
1	Bolsa de palabras	-0.25	0.89062
2	Bi-gramas	-0.25	0.89098
3	Tri-gramas	-0.25	0.89632
4	Tetra-gramas	-0.25	0.89886
5	Penta-gramas	-0.25	0.89888
6	Bolsa de palabras	-0.3	0.88915
7	Bi-gramas	-0.3	0.89210
8	Tri-gramas	-0.3	0.89740
9	Tetra-gramas	-0.3	0.89883
10	Penta-gramas	-0.3	0.89895
11	Bolsa de palabras	-0.375	0.89192
12	Bi-gramas	-0.375	0.89453
13	Tri-gramas	-0.375	0.89733
14	Tetra-gramas	-0.375	0.89984
15	Penta-gramas	-0.375	0.89908
16	Bolsa de palabras	-0.45	0.89290
17	Bi-gramas	-0.45	0.89585
18	Tri-gramas	-0.45	0.89810
19	Tetra-gramas	-0.45	0.89905
20	Penta-gramas	-0.45	0.89885
21	Bolsa de palabras	-0.5	0.89107
22	Bi-gramas	-0.5	0.89612
23	Tri-gramas	-0.5	0.89776
24	Tetra-gramas	-0.5	0.90010
25	Penta-gramas	-0.5	0.89916
26	Bolsa de palabras	-0.55	0.89284
27	Bi-gramas	-0.55	0.89567

28	Tri-gramas	-0.55	0.89869
29	Tetra-gramas	-0.55	0.89945
30	Penta-gramas	-0.55	0.89896
31	Bolsa de palabras	-0.6	0.89464
32	Bi-gramas	-0.6	0.89655
33	Tri-gramas	-0.6	0.89801
34	Tetra-gramas	-0.6	0.89976
35	Penta-gramas	-0.6	0.89914
36	Bolsa de palabras	-0.625	0.89464
37	Bi-gramas	-0.625	0.89729
38	Tri-gramas	-0.625	0.89847
39	Tetra-gramas	-0.625	0.89977
40	Penta-gramas	-0.625	0.89908
41	Bolsa de palabras	-0.65	0.89506
42	Bi-gramas	-0.65	0.89637
43	Tri-gramas	-0.65	0.89834
44	Tetra-gramas	-0.65	0.89970
45	Penta-gramas	-0.65	0.89931
46	Bolsa de palabras	-0.7	0.89593
47	Bi-gramas	-0.7	0.89690
48	Tri-gramas	-0.7	0.89922
49	Tetra-gramas	-0.7	0.90006
50	Penta-gramas	-0.7	0.89922
51	Bolsa de palabras	-0.75	0.89612
52	Bi-gramas	-0.75	0.89684
53	Tri-gramas	-0.75	0.89909
54	Tetra-gramas	-0.75	0.89987
55	Penta-gramas	-0.75	0.89931
56	Bolsa de palabras	-0.8	0.89544
57	Bi-gramas	-0.8	0.89681
58	Tri-gramas	-0.8	0.89921
59	Tetra-gramas	-0.8	0.89929
60	Penta-gramas	-0.8	0.89901
61	Bolsa de palabras	-0.85	0.89660
62	Bi-gramas	-0.85	0.89755
63	Tri-gramas	-0.85	0.89960

64	Tetra-gramas	-0.85	0.89971
65	Penta-gramas	-0.85	0.89909
66	Bolsa de palabras	-0.9	0.89560
67	Bi-gramas	-0.9	0.89769
68	Tri-gramas	-0.9	0.89897
69	Tetra-gramas	-0.9	0.89933
70	Penta-gramas	-0.9	0.89893
71	Bolsa de palabras	-0.95	0.89579
72	Bi-gramas	-0.95	0.89783
73	Tri-gramas	-0.95	0.89927
74	Tetra-gramas	-0.95	0.89972
75	Penta-gramas	-0.95	0.89910
76	Bolsa de palabras	-1	0.89604
77	Bi-gramas	-1	0.89723
78	Tri-gramas	-1	0.89951
79	Tetra-gramas	-1	0.89967
80	Penta-gramas	-1	0.89938

En la Tabla 16, se muestran los resultados del método de (Matias, 2016) + (Vázquez, 2015) a la longitud de 100 palabras.

Tabla 16. Evaluación del método de (Matias, 2016) +(Vázquez, 2015) a la longitud de 100 palabras.

Nº de experimento	Modelo de texto	*F-measure*
1	Bolsa de palabras	0.89630
2	Bi-gramas	0.89588
3	Tri-gramas	0.89824
4	Tetra-gramas	0.89981
5	Penta-gramas	0.89907
6	Bolsa de palabras	0.89750
7	Bi-gramas	0.89782
8	Tri-gramas	0.89951
9	Tetra-gramas	0.89935
10	Penta-gramas	0.89834

Anexo 6 Experimentos del método de (Matias, 2016) con los operadores Ruleta y Torneo

Los siguientes experimentos muestran la evaluación y comparación de los resúmenes generados con el método de (Matias, 2016) utilizando diferentes configuraciones de parámetros (ver Tabla 17) con el corpus TEXTRUSS.

Tabla 17. Parámetros utilizados en el método de (Matias, 2016).

Parámetros	
Pre procesamiento	No
Modelo de texto	*n*-gramas (*n* = 1,2,3,4,5)
Importancia de las oraciones	Valor de la pendiente
Función de aptitud	0.5δ+0.5β
Operador de selección	Ruleta y Torneo
Generaciones	Igual al número de oraciones de la noticia (archivo) utilizada del corpus TEXTRUSS.

Los valores de la pendiente se aprecian en la tabla 18.

Tabla 18. Valores de la pendiente (importancia de las oraciones).

-0.25	-0.3	-0.375	-0.45	-0.5	-0.55	-0.6	-0.625	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	-1

Los siguientes resultados muestran una comparación de los resultados obtenidos en el método de (Matias, 2016) a 100 palabras utilizando dos operadores de selección: Ruleta y Torneo (ver tabla 19 y figura 44).

Tabla 19. Parámetros utilizados en el método de (Matias, 2016) a la longitud de 100 palabras.

n-grama	Valor de la pendiente	Operador	Matias (2016) a la longitud de 100 palabras
4	-0.5	Ruleta	0.9001
4	-0.5	Torneo	0.88622
3	-0.85	Ruleta	0.8996
3	-0.85	Torneo	0.884965

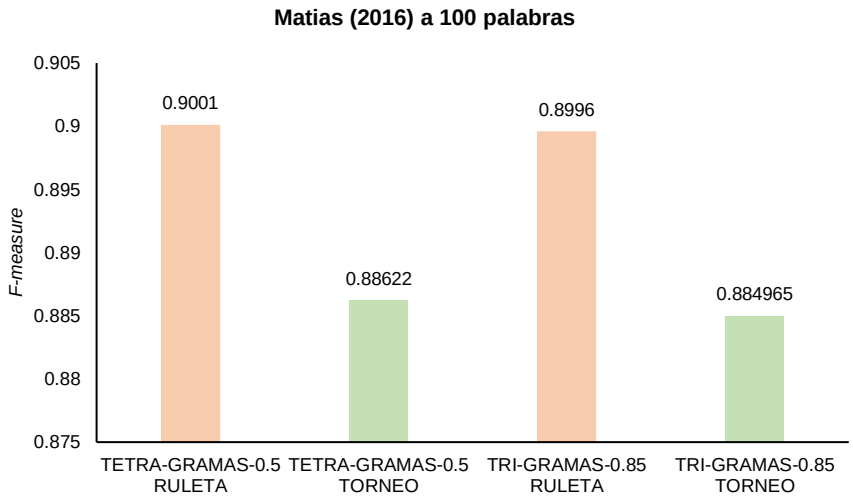

Figura 44. Resultados del método del estado del arte (Matias, 2016) a la longitud de 100 palabras.

Los siguientes resultados muestran una comparación de los resultados obtenidos en el método de (Matias, 2016) a la longitud del *gold standard* utilizando dos operadores de selección: Ruleta y Torneo (ver tabla 20 y figura 45).

Tabla 20. Parámetros utilizados en el método de (Matias, 2016) a la longitud del *gold standard*.

n-grama	Valor de la pendiente	Operador	Matias (2016) a la longitud tomada del *gold standard*
1	-0.625	Ruleta	0.90863
1	-0.625	Torneo	0.902469
2	-0.45	Ruleta	0.90746
2	-0.45	Torneo	0.902469

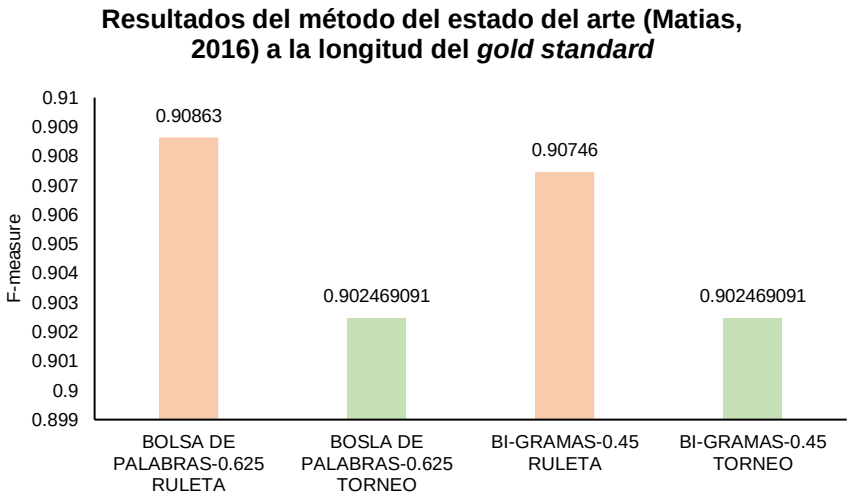

Figura 45. Resultados del método del estado del arte (Matias, 2016) a la longitud del *gold standard*.

Como se puede apreciar en la Figura 46, el método del estado del arte (Matias, 2016) se desempeña mejor con el operador de Ruleta que con el operador Torneo.

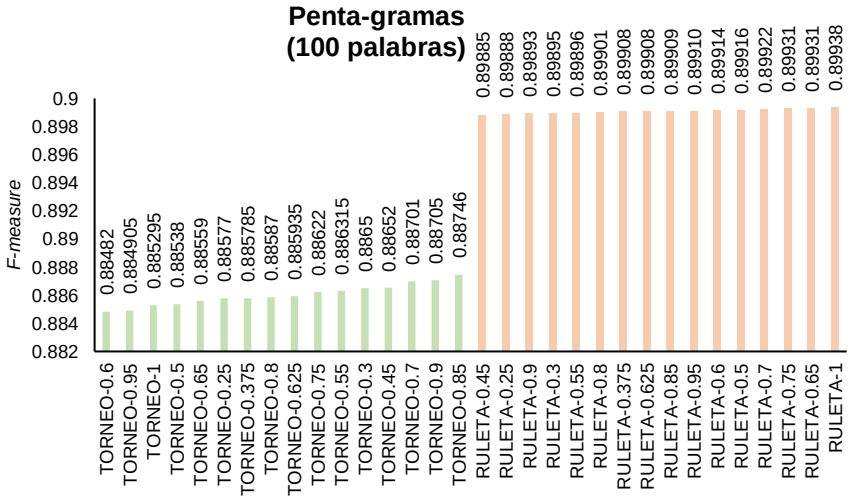

Figura 46. Evaluación del método (Matias, 2016) utilizando penta-gramas con los operadores de selección Torneo y Ruleta a la longitud de 100 palabras.