Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso

60 años de investigación

Griselda Areli Matias Mendoza
Yulia Ledeneva
René Arnulfo García Hernández

Alfaomega

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso.

60 años de investigación

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso.

60 años de investigación

Griselda Areli Matias Mendoza

Yulia Ledeneva

René Arnulfo García Hernández

Alfaomega

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación.
Griselda Areli Matias Mendoza, Yulia Ledeneva y René Arnulfo García Hernández.

Primera edición, febrero de 2020

Datos catalográficos

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación.

Primera edición

Alfaomega Grupo Editor, S.A. de C.V. México
ISBN: 978-607-538-615-7

Formato: 17 x 23 cm. Páginas 256

Imagen de la cubierta: Dreamstime

ISBN 978-607-633-151-4 (PDF UAEM)
ISBN 978-607-538-615-7 (PDF Alfaomega Grupo Editor)

El presente libro cuenta con la revisión y aprobación de dos pares doble ciego externos a la Universidad Autónoma del Estado de México. El arbitraje fue vigilado por la Secretaría de Investigación y Estudios Avanzados, según consta en el expediente número 143/2018.

El contenido de esta publicación es responsabilidad de las personas autoras.

Hecho en México

PRÓLOGO

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación, es un libro que puede ser leído por cualquier persona. Sin embargo, al ser un texto que presenta una tarea de Procesamiento del Lenguaje Natural (PLN) está más enfocado a investigadores, estudiantes de posgrado, estudiantes de doctorado, ingenieros y para todos los interesados en problemas del PLN y generación del conocimiento.

Una de las tareas inteligentes que realiza el ser humano es la creación de resúmenes de documentos, esto con el objetivo de que los lectores puedan conocer rápidamente la información contenida en ellos. Sin embargo, con el crecimiento exponencial de la información electrónica se ha dificultado esta labor, ya que se tiene mucha información y el acceso a ella requiere tiempo y por consecuencia recursos. La Generación Automática de Resúmenes de Textos (GART) es una tarea que cumple 60 años de investigación, desde la publicación de Lunh en 1958. Desde entonces existe un gran avance en las indagaciones sobre la GART especialmente en el lenguaje inglés, lo cual se puede ver reflejado en libros y artículos científicos que muestran la calidad de los métodos y técnicas de forma cuantitativa mediante su evaluación. No obstante, ha faltado realizar pruebas cualitativas que permitan saber si un resumen hecho por una máquina ha alcanzado la calidad para confundir a un humano y que no se dé cuenta si el resumen lo hizo una máquina o un humano. Para ello, se presentan los *Test de Turing* realizados con las máquinas que actualmente generan resúmenes de forma automática en los lenguajes más hablados y escritos como lo son: el inglés, el español, el portugués y el ruso.

El capítulo I de este libro comienza con una observación muy interesante que ha sido causa de debate entre los investigadores ¿*Una máquina puede ser inteligente*?, esta pregunta se aborda desde la problemática de GART. Para responder se presentan dos pruebas del *Test de Turing* para la tarea de GART: una para el lenguaje español y la otra para el lenguaje inglés. El objetivo de las pruebas es que el humano identifique de entre 6 resúmenes cuáles son los 2 hechos por personas. Además, se muestran los resultados obtenidos y se da una breve introducción a la problemática de la GART.

Para que la tarea de generación automática de resúmenes pueda ser estudiada se debe contar con ciertos recursos, por lo que en el capítulo II se presentan los elementos principales para estudiar y resolver la tarea de GART, como son: *corpus*, heurísticas y métricas de evaluación.

En el capítulo III se presentan los dos principales tipos de resúmenes según su estrategia de condensación: abstractivos y extractivos. Se da una descripción de cada uno, así como de algunos métodos científicos novedosos que trabajan de forma abstractiva. Además se muestra una tabla con las principales características que se usan para la tarea de GART, esencialmente para los extractivos.

En el capítulo IV se hace un análisis de las herramientas comerciales para la GART. Se describe el método que utiliza cada una de ellas y los pasos que se deben realizar para su funcionamiento. Esto con el objetivo de tener un panorama de la calidad de las herramientas comerciales con respecto a las heurísticas y los métodos científicos novedosos.

En los capítulos V, VI,VII y VIII se aborda la tarea de GART para los lenguajes inglés, español, portugués y ruso respectivamente, pero ¿Por qué en estos lenguajes? En inglés porque es el lenguaje más estudiado y hay más recursos y trabajos con los cuales hacer una comparación, además de que aún hay mucho por investigar. En español porque es nuestra lengua materna y obviamente estamos interesados en hacer resúmenes de textos de nuestro propio lenguaje. En portugués porque está dentro de las lenguas romances al igual que el español; hay investigaciones acerca de este lenguaje que aún no obtienen resultados deseables y, finalmente, en ruso primero porque no hay ningún tipo de investigación reportada para este lenguaje en la tarea de GART y posteriormente porque para poder generar los recursos para este lenguaje tenemos a un experto nativo. Además, en este capítulo, se muestran los resultados de evaluaciones realizadas con *corpus* especializados para

cada lenguaje. Se prueban los mejores métodos científicos novedosos, las herramientas comerciales y las principales heurísticas para la GART. Finalmente, en el capítulo IX se muestran las conclusiones y discusiones.

Entre las aportaciones que se destacan de este libro están: el reporte de seis pruebas del *Test de Turing*, con lo que se demuestra que una máquina puede engañar a un humano y presentar un resumen mejor que el realizado por éste; la integración y el reporte de los métodos novedosos desarrollados hasta el momentos; la comparación con los sistemas, la integración y reporte en español y ruso de la GART, ya que para estos lenguajes no se tenía una pesquisa formal y, finalmente, los resultados mostrados son una fuente de referencia para saber en qué punto está la investigación de la GART en los cuatro lenguajes.

Agradecimientos

La publicación de este libro fue posible gracias al financiamiento de la Secretaría de Investigación y Estudios Avanzados, la Sociedad Mexicana de Inteligencia Artificial y los autores.

Contenido

Capítulo I

Introducción

En este capítulo se introduce al lector en la problemática de la Generación Automática de Resúmenes de Texto (GART). Se presentan dos pruebas del *Test de Turing*, una en español y la otra en inglés, con el objetivo de saber si el humano es capaz de identificar cuáles son los resúmenes generados por los humanos y cuáles por la máquina. También se plantea si el humano puede replicar el conocimiento necesario para generar un resumen de manera automática en una máquina, de forma particular en una computadora.

¿Una máquina puede ser inteligente?

Esta simple pero profunda pregunta llevó a varios científicos a debatir sobre qué es la inteligencia, por lo que saltaron cuestionamientos tan básicos como: para desplazarme de un lugar a otro, ¿necesito ser inteligente? o quien realiza una suma de dos números, ¿se puede considerar inteligente? Para responder, Alan Turing, considerado uno de los padres de la computación, presentó una prueba que podía resolver de manera indirecta la pregunta original sobre cuándo una máquina, en particular una computadora, puede considerarse inteligente.

El *Test de Turing* consiste en un juego de imitación, lo realizan tres personas: un hombre (A), una mujer (B) y un interrogador (C) que puede ser de cualquier sexo. El interrogador se queda en una habitación separada de los otros dos. El objetivo del juego para el interrogador es determinar cuál de los otros dos es el hombre y cuál es la mujer. Él los conoce por las etiquetas X, Y, y al final del juego dice "X es A y Y es B" o "X es B y Y es A"; puede formar preguntas para A y B, todas las respuestas se dan de modo escrito para que la voz no ayude al interrogador a emitir una respuesta. La variante introducida por Turing consiste en sustituir a uno de los interrogados por una máquina, entonces el interrogador deberá determinar de la misma forma quién es A y quién es B, sin que sepa del aparato que hace las veces de uno de los interrogados. La máquina podría pasar el *Test de Turing* cuando el interrogador no lograra distinguir con quién está hablando (Turing, 1950). Es decir, cuando el humano se confunde, frecuentemente la máquina muestra inteligencia. Es por esta prueba que también se conoce a Alan Turing como el padre de la inteligencia artificial.

La inteligencia artificial tiene como objetivo emular algunas de las facultades intelectuales humanas en sistemas artificiales (Benítez *et al.*, 2014). Por emular se entiende la aplicación de modelos teóricos en una máquina (computadora) con el fin de obtener resultados satisfactorios para el humano. Las áreas de estudio más emergentes de la inteligencia artificial son: robótica, sistemas expertos, problemas de percepción, aprendizaje, Procesamiento del Lenguaje Natural (PLN), entre otras (Coarite Choque, 2008).

Basada en el *Test de Turing*, la empresa IBM desarrolló sistemas que pudieran competir claramente contra el humano en tareas que se pueden considerar muy inteligentes. Para ello, se enfocó en el juego del ajedrez: se puso a competir al campeón de ajedrez Garry Kasparov con el sistema desarrollado por IBM,

conocido como Deep Blue, donde después de varios encuentros la computadora pudo vencer al jugador más inteligente (Hsu, 1999).

Recientemente, en el año 2011, una de las más famosas pruebas del mismo *test* fue aplicado por IBM para la tarea de responder a preguntas formuladas en lenguaje natural. La empresa puso a competir a dos humanos campeones del juego Jeopardy con el sistema Watson, capaz de reconocer la voz, formular la pregunta y reproducir la respuesta también en voz. Watson es un sistema de inteligencia artificial y, en la competencia, fue capaz de emular y superar al humano (Gliozzo *et al.*, 2017).

Más allá de los juegos, hoy en día es común tener noticias sobre cómo la inteligencia artificial (y todas sus subdisciplinas) apoyan o superan actividades que realiza el ser humano en ámbitos como la medicina, la seguridad y la educación; por mencionar algunos.

1.1 *Test de Turing* aplicado a la generación automática de resúmenes

Una de las tareas inteligentes que realiza el ser humano es la creación de resúmenes de documentos, con el objetivo de que los lectores puedan conocer rápidamente la información contenida en ellos. Sin embargo, con el crecimiento exponencial de la información electrónica se ha dificultado esta tarea pues se tiene mucha información y el acceso requiere de tiempo y, por consecuencia, de recursos.

La Generación Automática de Resúmenes de Textos (GART) es una tarea que cumple 60 años de investigación desde su primera publicación (Luhn, 1958). Existe un gran avance en las investigaciones sobre la GART en inglés, reflejado en libros especializados y artículos científicos, que se muestra a través de la calidad de los métodos y las técnicas cuantitativas mediante su evaluación (Mani, 2001), (Mihalcea *and* Radev, 2011), (Ledeneva *and* García-Hernández, 2013), (Torres-Moreno, 2014), (Ledeneva *and* García-Hernández, 2017), etc. Sin embargo, ha faltado realizar el *Test de Turing* a las máquinas que actualmente generan resúmenes de forma automática, en los lenguajes más hablados y escritos, como son: el inglés, el español, el portugués y el ruso.

Para introducir los conceptos fundamentales de la GART se recomienda al lector que realice el siguiente *Test de Turing* para distinguir, a partir de una noticia

en español (del periódico mexicano *La Crónica*[1]), cuáles de los resúmenes fueron realizados por humanos y cuáles por máquinas.

La noticia que se presenta a continuación se tomó tal como aparece en el portal.

Diseñan casa que resbalaría vientos de huracán y tornados

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. ¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera requisitos de la región, como la nieve, por ejemplo, desarrolló una idea sencilla pero práctica. "Si uno parte de la base de que algo que no es plano no recibe viento a diferencia de algo vertical que lo recibe totalmente, algo que tenga 45 grados de inclinación deberá de ser sólo afectado en el 50 por ciento", explica en entrevista el arquitecto. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Este diseño piramidal ha sido patentado ya por el mexicano en el país y en EU, puesto que no existe algo parecido en la industria de la construcción hasta ahora. La tarea ahora es promoverla porque el papel por sí mismo "no sirve de nada". Si las afectaciones a la población por este tipo de fenómenos naturales son una constante, por qué no se había pensado en soluciones distintas para disminuir el riesgo o los daños. "Porque hay mucho poder económico en medio de esto. Pero aún así se puede modificar y hacer el esfuerzo". CONFORTABLE. De acuerdo con el arquitecto, el diseño que propone se puede componer de dos formas: elaborando las viviendas con piezas precoladas o con colado en el mismo lugar. Lo preferente sería la segunda opción, añade. "De esta forma, la estructura sería más íntegra y resistente a tornados, pero ni huracanes

[1]Periódico de alto renombre en México. Se puede acceder de manera electrónica en: http://www.cronica.com.mx/noticias.php

ni sismos le afectarían. Tenemos una opción para evitar más fallecimientos y millones de dólares en pérdidas materiales". La estructura de concreto armado podría ser armado con aditivos que la hicieran impermeable, pero incluso resistente al fuego y también confortable, porque sería menos afectada por aumento o baja de temperaturas. Díaz Zubieta tiene, por otra parte, una solución de este tipo para México en el contexto de los ciclones que nos afectan y la situación económica de nuestra población. "Es una más sencilla y económica porque tenemos otro contexto (también sería aplicable a países de Latinoamérica). Ésta representaría sólo el 15% del costo del diseño que propongo para EU; sería de fácil construcción y pensada para población de bajos recursos". Si bien apunta que sería totalmente diferente a su primer diseño, las modificaciones que se realizarían serían sin detrimento a su calidad y seguridad.

De la noticia anterior se le pidió tanto a los humanos como a las máquinas que generaran resúmenes con una longitud de cien palabras, por lo que se tuvo que hacer un recorte de los mismos para tener una evaluación objetiva. En particular, hay que identificar de los resúmenes que se presentan cuáles son los dos generados por humanos.

Resumen 1

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera requisitos de la región, como la nieve, por ejemplo, desarrolló una idea sencilla pero práctica. CONFORTABLE. La estructura de concreto armado podría ser armado con aditivos que la hicieran impermeable, pero incluso resistente al fuego y también confortable, porque sería menos afectada por aumento o baja de temperaturas. Díaz Zubieta…

Resumen 2

¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Si

las afectaciones a la población por este tipo de fenómenos naturales son una constante, por qué no se…

Resumen 3

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. ¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera…

Resumen 4

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. “Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto”. De acuerdo con el arquitecto, el diseño que propone se puede componer de dos formas: elaborando las viviendas con piezas precoladas o con colado en…

Resumen 5

CONFORTABLE. “Es una más sencilla y económica porque tenemos otro contexto (también sería aplicable a países de Latinoamérica) Díaz Zubieta tiene, por otra parte, una solución de este tipo para México en el contexto de los ciclones que nos afectan y la situación económica de nuestra población Tenemos una opción para evitar más fallecimientos y millones de dólares en pérdidas materiales” “Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto” Ésta representaría sólo el 15% del costo del…

Resumen 6

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes ¿Por qué no hacer casas más resistentes?, Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los

> ciclones: casas resistentes a tornados y huracanes. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Este diseño piramidal ha sido patentado…

La prueba anterior y otras dos más (ver anexo A) en español se aplicaron a setenta y tres estudiantes y profesores de nivel superior y de posgrado, quienes tienen como lengua nativa el español. Los resultados se muestran en la **tabla 1.1.**

Tabla 1.1 Resultados del *Test de Turing* respecto al humano para el lenguaje español

Pares de resúmenes elegidos por el humano	Porcentaje de confusión entre los resúmenes seleccionados (%)
Humano – Máquina	56
Máquina – Máquina	36
Humano – Humano	8

En la fila tercera de la **tabla 1.1** se muestra cómo en sólo 8% de las ocasiones las personas acertaron en identificar correctamente los dos resúmenes hechos por el humano. La mayor confusión se dio en la primera fila, con 56%, donde se seleccionó un resumen elaborado por el humano y uno por la máquina. Sin embargo, de modo interesante, 36% de las personas pensaron que los resúmenes generados automáticamente fueron los creados por humanos. Con estos resultados es posible ver que la máquina no sólo ha pasado el *Test de Turing*, sino que ha superado al humano.

Es posible que después de estas cifras el lector pueda dudar de su elección, pero la confusión o duda se debe a que los resúmenes son muy parecidos. A lo largo de este libro, mostraremos cuáles son los resúmenes hechos por el humano y cuáles son los realizados por la máquina.

Motivados por los resultados anteriores y debido a que el grueso de la investigación en la GART se ha realizado en inglés, se decidió hacer el *test* anterior para este lenguaje. Se recurrió a sesenta y ocho personas que tienen dominio en lectura del inglés. De igual forma, se invita al lector a que haga el *test* sobre la siguiente noticia.

Hurricane Gilbert Heads Toward Dominican Coast

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. ``There is no need for alarm,'' Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather'' rotating around the center of the storm. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast. There were no reports of casualties. San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night. On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 80 mph winds and sheets of rain. Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

Los resúmenes correspondientes al texto anterior se presentan a continuación; todos aquellos en inglés tienen la misma longitud de cien palabras.

Resumen 1

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. An estimated 100,000 people live in the province, including 70,000 in the city

of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with…

Resumen 2

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo…

Resumen 3

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto…

Resumen 4

Tropical Storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15mph with a broad area of cloudiness and heavy weather with sustained winds of 75mph gusting to 92mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at..

Resumen 5

The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. An estimated 100,000 people live in the province, including 70,000

in the city of Barahona, about 125 miles west of Santo Domingo. The National Hurricane Center in Miami reported its position at 2 a m Sunday at latitude 16 1 north, longitude 67 5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. Residents returned home, happy to find little damage from 80…

Resumen 6

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical…

Los resultados del *Test de Turing* para el lenguaje inglés se presentan en la tabla 1.2.

Tabla 1.2 Resultados del *Test de Turing* respecto al humano para el lenguaje inglés

Pares de resúmenes elegidos por el humano	Porcentaje de confusión entre los resúmenes seleccionados (%)
Humano – Máquina	46
Máquina – Máquina	41
Humano – Humano	13

Ahora tenemos que solamente se pudo identificar 13% de los casos de manera correcta, es decir, Humano-Humano. La mayor parte de la confusión estuvo en las selecciones Humano-Máquina, con 46%. Sin embargo, en 41% de los casos se prefirieron resúmenes hechos por una máquina, es decir, de cada 4 pruebas 3 fueron confundidas y se optó por el resumen de la máquina. Nuevamente, el *Test de Turing* fue superado al menos en el dominio[2] de las noticias. Cabe mencionar que, para los lenguajes portugués y ruso aún no se presentan pruebas con dicho *test*. Sin embargo, se considera como un trabajo futuro. Para el lenguaje inglés se realizaron dos pruebas más, las cuales se presentan en el anexo B.

[2]El dominio hace referencia al ámbito en el que están escritos los documentos, por ejemplo, noticias, artículos científicos, poemas, tweets, correos electrónicos, entre otros.

Seguramente a partir de las pruebas anteriores surgen varias observaciones, pero serán resueltas a lo largo del libro; además, se mencionarán cuáles son los resúmenes realizados por la máquina y cuáles por el humano. El primer aspecto es sobre la manera de hacer los resúmenes, los cuales por su forma de condensación pueden ser extractivos o abstractivos. Se considera que los resúmenes extractivos sólo sustraen un conjunto de oraciones (pueden ser párrafos o frases) del documento original. Sin embargo, los abstractivos pueden modificar las oraciones del documento original e incluir ideas u opiniones del humano que hace el resumen. Por lo anterior el resumen extractivo se considera más objetivo y el abstractivo más subjetivo, lo importante es diferenciar el uso que se quiere dar.

En las pruebas anteriores y a lo largo del libro sólo se tratan los resúmenes extractivos porque es de nuestro interés mantener las ideas originales del autor. Por otro lado, se considera que en realidad el humano genera un resumen más del tipo extractivo porque únicamente copia las partes que considera importantes (Jing, 2002).

1.2 ¿Cómo el humano hace el resumen?

En el caso de los humanos, a quienes se inicia en la enseñanza de nivel básico siendo niños, se les instruye en los procesos para la generación de resúmenes. Muchos autores coinciden en que se deben seguir siempre una serie de pasos (Vivaldi, 2000), (Maqueo, 2004), (UNE 50-103-90, 1990), (Kaufman & Perelman, 1999). Sin embargo, cada autor tiene un planteamiento diferente por lo que no se conoce un método estandarizado. No obstante, lo que se busca es extraer las características más importantes de un texto original para ser integradas en uno más corto que contenga las ideas principales.

Entre algunos de los pasos para realizar un resumen, según propone Ana María Maqueo (Maqueo, 2004), son los siguientes:

1. Leer con atención un texto.
2. Separar en bloques de ideas.
3. Subrayar las ideas principales.
4. Redactar el resumen enlazando las ideas principales con los nexos correspondientes.

En un estudio hecho por Kaufman & Perelman (1999), se encontró que a ciento ochenta niños se les indicó que los resúmenes se generan según los siguientes pasos:

1. Subrayar el texto que se considere importante.
2. Borrar todo aquello que se piense no debe estar en su resumen.
3. Textualizar: se le proporciona una hoja rayada (con trece renglones) y se le pide que escriba el resumen en el espacio asignado, sin exceder la extensión.

En su libro, Vivaldi (2000) menciona que los pasos para realizar el resumen de un libro son los siguientes:

1. Averiguar cuáles son los capítulos más importantes y tomar notas sobre los conceptos fundamentales.
2. Resumir lo más interesante dejando para el final lo menos relevante. Respecto a este paso, el autor menciona que si se tiene el tiempo para leer despacio se pueden hacer notas. Sin embargo, si el tiempo es insuficiente, se suele subrayar el texto y presentarlo como parte del resumen.
3. El orden del resumen debe reflejar lo que significa el libro.

Si observamos los pasos propuestos por los autores, es posible darse cuenta de que los procedimientos que se enseñan para generar resúmenes están guiados a partir de la generación de resúmenes extractivos. Aunque, cabe mencionar que autores como Maqueo (2004) recomiendan textualizar y reescribir el texto, que es donde el humano implementa la abstracción.

Considerando lo anterior, no es de sorprender que existan estudios donde se comprueba que los humanos generalmente realizan resúmenes mediante la selección de oraciones del documento original y sólo en casos donde se hace una revisión exhaustiva y se exige una comprensión muy elevada del texto, utilizan técnicas de reescritura y abstracción (Banko *and* Vanderwende, 2004). El trabajo de Jing (2002) presenta una descomposición de un conjunto de trescientos resúmenes realizados por humanos, con el objetivo de determinar cómo se generaron. En el resultado se identificaron seis tipos de operaciones que realizan los humanos a la hora de hacer los resúmenes:

1. Simplificación de oraciones.
2. Combinación de oraciones.
3. Transformación sintáctica.

4. Parafraseado del léxico.
5. Generalización.
6. Especialización.

Los experimentos mostraron que 81% de las oraciones correspondían a una copia idéntica de aquéllas presentes en los documentos originales (Jing, 2002).

Tomando como referencia el estudio anterior, los pasos propuestos para la creación de un resumen y los resultados obtenidos en el *Test de Turing*, se puede concluir que los humanos hacen resúmenes extractivos porque sólo copian las partes que consideran importantes; entre ellas están las primeras oraciones. Al respecto, se tienen varias hipótesis: que el humano se cansa y no lee el texto completo, que el autor escribe lo más importante al principio del texto y que el dominio del documento influye de manera relevante. Por estas razones, los resúmenes que hace el humano son de las primeras oraciones.

Como se mencionó, el dominio del documento influye en la construcción de un resumen; de manera particular para la tarea de GART, los textos más estudiados son las noticias. En el ámbito de noticias se usa la denominada regla de la pirámide invertida, que consiste en colocar la información más relevante al principio del texto (**figura 1.1**). Aunque existen otras heurísticas como la estructura tradicional del texto narrativo, donde se suelen tener los elementos de mayor importancia en el final, o la de pirámide trunca también llamada reloj

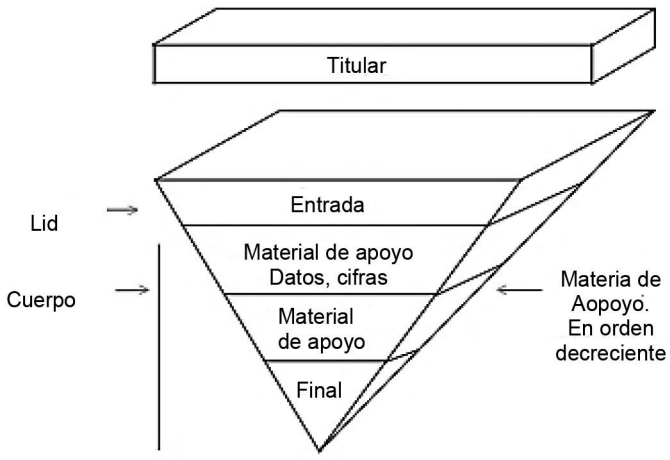


Figura 1.1 Estructura de una noticia (Briones et al., 2012)

de arena donde se comienza con la presentación de los datos más importantes al principio, el resto de la información en forma decreciente y, dado un momento, se comienza nuevamente a redactar información importante (Briones *et al.*, 2012).

En Vázquez (2015) se hace un análisis de un *corpus* en inglés (DUC02) para identificar cómo los humanos seleccionan las oraciones. El documento más largo tiene 177 oraciones; se puede observar en la **figura 1.2** cómo en 50% de veces están seleccionadas las oraciones del 1 al 5, mientras que de 25% de veces están seleccionadas de 6 al 10; a partir de la oración 15 se tiene un 20% de frecuencia de selección, que va descendiendo conforme aumenta el número de oraciones. Por ejemplo, las oraciones que van de la 61 a la 177 sólo se seleccionan una vez en los 567 documentos.

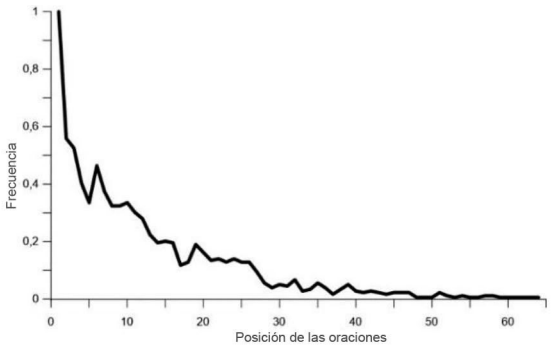


Figura 1.2 Relación posición frecuencia en el *corpus* DUC02

En estos sesenta años de investigación en resúmenes, se ha abordado en mayor medida la investigación en la GART extractivos y en el idioma inglés. Sin embargo, por la naturaleza de los textos, algunos de los métodos creados pueden trabajar con más de un lenguaje. Es por ello que en las pruebas anteriores se presentaron cuatro resúmenes extractivos obtenidos por máquinas, probados en inglés, pero que pueden trabajar en otros lenguajes.

Los resultados de las pruebas del *Test de Turing* muestran claramente cómo la máquina ha logrado confundir al humano en la tarea de GART; incluso sugieren que prefiere los resúmenes generados por la máquina. Con esta validación, se prueba que una máquina puede ser inteligente para hacer un resumen en español e inglés. Sin embargo, es importante reflexionar que se está probando de fondo

que el humano es tan inteligente que puede modelar y reproducir el conocimiento necesario para que una máquina pueda parecerse a él en alguna tarea.

En otras palabras, el *Test de Turing* realizado permite observar que el humano está logrando hacer que la máquina lo emule en la GART. Entonces, si la máquina realmente puede aprender o hacer una búsqueda inteligente de las principales ideas de un texto a través de su estructura, y componer un resumen a partir de ésta, esto puede ser aplicado a diferentes lenguajes.

Para poder hacer experimentos reales a diferentes lenguajes, se requiere hacer uso de *corpus* especializados y herramientas de evaluación. Los *corpus* regularmente están formados por un conjunto de documentos y por uno o dos resúmenes de cada documento hechos por los humanos, llamados *gold standard*,[3] los cuales sirven de referencia para poder evaluar los resúmenes producidos por la máquina (Over *et al.*, 2007). Para el lenguaje inglés, una de las herramientas más utilizadas para la evaluación es *ROUGE* (Lin, 2004), la cual permite obtener una calificación del resumen creado por la máquina en comparación con el del humano. Con un *corpus* y la herramienta de evaluación, se puede determinar más fácilmente en qué estatus se encuentra el estudio de la tarea de GART para cualquier lenguaje.

En los siguientes capítulos se abordan las problemáticas que han enfrentado cuatro de los lenguajes más importantes en la GART. En específico, se escogieron los lenguajes inglés, español, portugués y ruso, porque están dentro los diez más importantes. El inglés se eligió por ser el lenguaje que más investigación ha atraído y por ser el más hablado, ya sea como lengua nativa o como segunda lengua.[4] El español se eligió por ser el segundo lenguaje más hablado de forma nativa, por ser uno de los que menos atención de investigación ha recibido en este tema, y por tener una raíz diferente a la anglosajona (Arévalo, 2017). El lenguaje portugués se eligió por tener una raíz cercana al español y por poseer algunas investigaciones en este tema (Pardo *and* Rino, 2003), (Mihalcea *and* Tarau, 2005). Por último, se estudiará la GART en ruso por tener una raíz diferente a los lenguajes anteriores y por no tener investigaciones referentes en el tema.

[3]Los resúmenes *gold standard* son los hechos por los humanos y se les conoce también como estándar de oro o resúmenes de referencia.

[4]El sitio web https://www.internetworldstats.com/stats7.htm nos proporciona el top de los 10 lenguajes más hablados en el mundo.

Cabe señalar que, en este capítulo, se hace referencia a la máquina como la computadora en la que se puede implementar un método de GART. A partir de los siguientes capítulos nos referiremos a los métodos científicos novedosos y herramientas comerciales como sinónimo del concepto dado a la máquina en el capítulo I.

1.3 Organización del libro

Este libro busca resumir sesenta años de investigación en la tarea de la GART; se sabe que el lenguaje inglés es el más estudiado en ella, por lo que se recuperarán los esfuerzos hechos en ese lenguaje y se aplicarán en español, portugués y ruso. Además, se presentan los resultados de las principales heurísticas (*baseline:random*, *baseline:first* y *Topline*), las herramientas y los métodos científicos novedosos en la GART para cada lenguaje.

El libro está compuesto por ocho capítulos. El capítulo I presenta dos pruebas del *Test de Turing* para la tarea de GART, una para el lenguaje español y la otra para el inglés. El objetivo de las pruebas es que el humano identifique, de entre seis resúmenes, cuáles son los dos hechos por una persona. Además, se muestran los resultados obtenidos y una breve introducción a la problemática de la GART.

En el capítulo II se presentan los elementos principales para resolver la tarea de GART, como son: *corpus*, heurísticas y métricas de evaluación.

En el capítulo III se describen las dos clasificaciones de los resúmenes según su estrategia de condensación; además, se explican algunos métodos científicos novedosos que trabajan en diferentes lenguajes.

En los capítulos IV al VIII, se aborda la tarea de GART para los lenguajes inglés, español, portugués y ruso, respectivamente. Además, se muestran los resultados de evaluaciones realizadas con *corpus* especializados para cada lenguaje. Se prueban los mejores métodos científicos novedosos, las herramientas comerciales y las principales heurísticas para la GART. Finalmente, en el capítulo IX se enlistan las conclusiones y discusiones.

Capítulo II

Corpus, heurísticas y métricas de evaluación

En este capítulo se presenta la descripción de *corpus*, heurísticas y métricas de evaluación que se usan para enriquecer las opciones de GART por la máquina y por el humano para los lenguajes inglés, español, portugués y ruso.

Los primeros estudios sobre la GART se dieron a finales de 1950. Los primeros trabajos fueron realizados por Luhn (1953 y 1958) y Edmundson (1969), los cuales utilizaban sus propios *corpus* y evaluaban de forma manual, por lo que los métodos de GART realizados hasta antes del año 2000 no tenían una referencia de comparación (Luhn, 1953), (Luhn, 1958), (Edmundson *and* Wyllys, 1961), (Edmundson, 1969), (Kupiec *et al*., 1995), (Mani *et al*., 1999). Para poder saber cuál es la calidad de los resúmenes generados por las máquinas en comparación con los creados por los humanos, es necesario contar con un *corpus*, heurísticas de referencia y métricas de evaluación.

2.1 *Corpus*

Según el Diccionario Manual de la Lengua Española (*Corpus*, 2014), un *corpus* es un conjunto extenso de textos de diversas clases, ordenados y clasificados, que sirven como base de una investigación. En el área de la GART los *corpus* pueden estar constituidos por textos de diferentes dominios (por ejemplo, noticias, artículos científicos, textos literarios, de redes sociales, textos de cocina, entre otros) y los resúmenes generados por el humano, llamando a este conjunto *gold standard* (en algunos casos podría tener un solo resumen).

Entre los primeros *corpus* estándar que se crearon, están los de la conferencia *Document Understanding Conferences* (DUC); la cual surgió a partir del año 2000 y está compuesta por siete más: DUC01, DUC02, DUC03, DUC04, DUC05, DUC06 y DUC07. Cada conferencia se compone de varias tareas y para cada una se crearon los *corpus* con sus respectivos *gold standard*. El objetivo de estas conferencias era aumentar los avances de la investigación en la tarea de GART en el lenguaje inglés, a través de experimentos a gran escala y con el ofrecimiento de utilizar los *corpus* y sus *gold standard*.

A partir del año 2008, surgió la conferencia de análisis de textos llamada *Text Analysis Conference* (TAC); organizada por talleres de evaluación realizados para motivar la investigación en el procesamiento del lenguaje natural y sus aplicaciones relacionadas. Uno de los principales objetivos de TAC es construir colecciones de prueba para anticipar las necesidades de evaluación de los sistemas modernos. Fue en los años 2008, 2009, 2010, 2011 y 2014, cuando TAC se enfocó en la tarea de GART, siendo su principal área de estudio los resúmenes para multidocumentos enfocados al usuario. La mayor parte de la investigación en el

área de la GART se ha realizado para el lenguaje inglés, y existen actualmente pocos trabajos para otros lenguajes.

Para el lenguaje español se han realizado investigaciones, pero ninguna de ellas utiliza *corpus* estándar o especializado para GART. En estos casos, hacen uso de *corpus* adaptados de la tarea de extracción de información o simplemente generan los propios (Acero *et al.*, 2001), (Toledo-Báez, 2010), (da Cunha Fanego, 2005), (Villatoro E., 2007), (Plaza, 2011), (Venegas, 2011), (Cabral *et al.*, 2014). Por lo tanto, las investigaciones no pueden ser comparadas y no se puede determinar el avance que se tiene en el área de GART para el lenguaje español.

Para el portugués existe el *corpus TeMário* (Pardo *and* Rino, 2003), el cual ha sido utilizado en la mayoría de las investigaciones (Antiqueira, 2007), (Margarido *et al*., 2008), (Leite *and* Rino, 2009), (Amancio *et al*., 2012), (Cabral *et al.*, 2014), (Cavalieri *et al*., 2015), lo que ha permitido una comparación entre los métodos y herramientas para este lenguaje.

Para el ruso no se tiene conocimiento de algún *corpus* estándar para la tarea de GART. Uno de los trabajos principales realizado por Braslavski *and* Gustelev (2007), toma las noticias del periódico Gazeta.ru. Sin embargo, no está disponible. En el trabajo de Rojas (2016) se construye un *corpus* de las noticias del mismo periódico para la tarea de GART, los resultados se muestran en el capítulo VIII de este libro.

2.2 Heurísticas

La palabra *heurística* se traduce del griego como hallar o inventar. La heurística es la invención de reglas, procedimientos o técnicas para que un humano solucione determinado problema. En la metodología científica, la heurística se aplica para resolver tareas de las que no se tiene un procedimiento algorítmico de solución (Polya *and* Zugazagoitia, 1965).

En este libro, para saber si una máquina es inteligente al momento de construir un resumen, se presentan varias heurísticas intuidas por los dominios de los datos utilizados. A través de la comparación de las heurísticas calculadas se analiza el desempeño de los métodos y herramientas inteligentes.

Las heurísticas calculadas para la tarea de GART consideran la forma en que el humano hace resúmenes. Por ejemplo, para componer una noticia las

personas colocan la información más importante en las primeras oraciones y después describen los detalles de la noticia. Por esta razón, se ha propuesto utilizar la heurística de primeras oraciones llamada *baseline:first* (Ledeneva, 2008); la cual consiste en componer un resumen de las primeras 100, 200 o 400 palabras, donde la cantidad de las mismas depende de la longitud del resumen a generar. Otra de las heurísticas es *baseline:random* que se considera la peor forma de elegir las oraciones para un resumen (*baseline:random*) y, finalmente, está *Topline* que se pondera como la mejor.

2.2.1 *Baseline:first*

Baseline:first consiste en tomar las primeras palabras del texto para conformar el resumen (Ledeneva, 2008). En los experimentos de este libro se toman las cien primeras palabras. Para una máquina inteligente, la meta es superar esta heurística. De manera particular, para el dominio de noticias la meta resulta ser muy elevada ya que este tipo de textos contiene la información más importante al principio del documento. En el *Test de Turing* realizado en el capítulo I se consideró incluir a la heurística *baseline:first*. Los resultados revelan que el humano se confunde cuando ve un resumen generado por *baseline:first*, lo que denota la gran influencia que tienen las primeras oraciones en la construcción de un texto. Para el español, el resumen 3 es el correspondiente a la heurística *baseline:first*; para el inglés, es el 2 el que corresponde a la heurística *baseline:first* (sección 1.2).

2.2.2 *Baseline:random*

La peor forma de realizar un resumen sería elegir al azar las oraciones que lo van a constituir. A esta forma de elección se le denomina *baseline:random,* propuesta y utilizada desde Ledeneva (2008). Cuando una máquina genera un resumen se espera que sea más inteligente y se obtengan mejores textos que sólo al azar. En el capítulo I se presentó la prueba de Turing, en la que se mostraron seis resúmenes, dos de ellos hechos por el humano, dos automáticamente y dos correspondientes a las heurísticas.

Para el lenguaje español, el resumen 5 es el correspondiente a la heurística *baseline:random* mientras que, para el inglés, el resumen 5 de los que conforman el *Test de Turing* es el de la heurística *baseline:random* (sección 1.2).

2.2.3 *Topline*

La heurística *Topline* consiste en obtener la mejor combinación de oraciones de todas las posibles combinaciones. Lo que nos permite saber cuál es el máximo resultado al que podemos llegar al evaluar los resúmenes generados con un *corpus* (Rojas, 2018).

Uno de los principales retos de la GART consiste en hacer resúmenes extractivos mejor parecidos a los resúmenes generados por los humanos (*gold standard*). Sin embargo, para varios dominios, los resúmenes *gold standard* son elaborados de forma abstractiva sustituyendo algunos términos y frases del texto original. De acuerdo con el trabajo de Verma y Lee (Verma *and* Lee, 2017), los *gold standard* de los *corpus* estándar en inglés DUC01 y DUC02 emplean aproximadamente 9% de palabras no utilizadas en los documentos originales. En consecuencia, el nivel de similitud máximo será menor a 100% y, además, si son comparados a partir de varios resúmenes *gold standard*, los límites de desempeño máximo serán menores (debido a la falta de concordancia entre humanos) para cualquier método de la GART.

Para la prueba del *Test de Turing* realizada en el capítulo I, se presentaron para cada lenguaje los dos *gold standard*; para el caso del español, los resúmenes 2 y 4 son los generados por el humano, mientras que, para el inglés, son los números 4 y 6 (sección 1.2).

2.3 Evaluación de resúmenes automáticos

Para poder evaluar los resúmenes generados por una máquina se requiere no solamente la construcción de conjuntos de datos estándares (*corpus*), sino también la utilización de diferentes métodos de evaluación.

Los métodos de evaluación se clasifican en intrínsecos y extrínsecos (Sparck Jones *and* Galliers, 1995). Los primeros se basan en el análisis directo del resumen producido automáticamente. Para juzgar la calidad se pueden utilizar criterios gramaticales, de cohesión y coherencia del texto. Para evaluar el grado de cobertura, generalmente, se recurre a la comparación de los resúmenes hechos automáticamente contra los producidos por los expertos.

Los métodos de evaluación extrínseca estudian el resumen en el contexto de la tarea para la que fue generada, tratando de determinar su efecto en alguna

otra tarea. Estas tareas pueden incluir, por ejemplo, la evaluación de la relevancia (Berker, 2011).

Las métricas de evaluación que se presentan en este capítulo son: similitud de contenido, precisión, recuerdo y *f-measure*; y los métodos de evaluación son *ROUGE* y *Pyramid*.

2.3.1 Similitud de contenido

En Donaway (*et al.*, 2000) se propone una métrica para evaluar la calidad informativa de un resumen; la cual puede aplicarse para resúmenes extractivos y abstractivos. Una de las medidas definidas para calcular dicha similitud es la prueba de vocabulario (*vocabulary test*), donde se emplean métodos tradicionales de recuperación de información para calcular la distancia entre las representaciones vectoriales del resumen automático y el manual, utilizando la métrica del coseno. Esta métrica puede automatizarse y hacer uso de los resúmenes elaborados por los humanos.

2.3.2 Precisión, recuerdo y *f-measure*

Las medidas de precisión y recuerdo son las tradicionales en la recuperación de información (Salton *and* McGill, 1983). También han sido utilizadas para la evaluación de la tarea de GART.

Precisión (P): Refleja la cantidad de oraciones correctas extraídas por la máquina:

$$P = \frac{correctas}{(correctas + incorrectas)} \tag{1}$$

Recuerdo (R): Refleja la cantidad de oraciones correctas que olvidó el sistema:

$$Recuerdo = \frac{correctas}{(correctas + olvidadas)} \tag{2}$$

Se define como *correctas* al número de oraciones extraídas por la máquina y por el humano; *incorrectas* como el número de oraciones extraídas por la máquina, pero no por el humano, y *olvidadas* como el número de oraciones extraídas por el humano, pero no por la máquina.

***F-measure* (F):** Es la medida armónica de aquellas de precisión y recuerdo:

$$F = \frac{2 * (Precisión * Recuerdo)}{(Precisión + Recuerdo)} \tag{3}$$

La medida *f-measure* establece un balance equilibrado entre el recuerdo y la precisión.

2.3.3 *ROUGE*

ROUGE (en español: Evaluación Suplente de Resúmenes Orientada al Recuerdo) (Lin, 2004) fue propuesta por Lin y Hovy (Lin *and* Hovy, 2003), (Lin *and* Och, 2004), (Lin *and* Och, 2004). Este sistema calcula la calidad de un resumen generado de forma automática mediante la comparación con otros creados por humanos. En concreto, se cuenta el número de las diferentes unidades comunes, tales como secuencias de palabras, pares de palabras y *n-gramas*, entre el resumen a evaluarse (el de la computadora) y los resúmenes ideales creados por seres humanos. *ROUGE* incluye varias medidas automáticas de evaluación:

- *ROUGE*-N (co-ocurrencia de *n*-gramas): Expresa la cobertura o recuerdo de *n*-gramas entre un resumen candidato y un conjunto de resúmenes de referencia, y es calculado de la siguiente manera:

$$ROUGE - N = \frac{\sum_{O \in \{ResúmenesDeReferencia\}} \sum_{grama_n \in O} cuenta_{coincidencia}(grama_n)}{\sum_{O \in \{ResúmenesDeReferencia\}} \sum_{grama_n \in O} cuenta(grama_n)} \tag{4}$$

  donde n es la longitud del $n - grama$, y $cuenta_{concordancia}(grama_n)$ es el número máximo de $n - gramas$ que co-ocurren en el resumen candidato y en el conjunto de resúmenes de referencia.

- *ROUGE*-L (Subsecuencia más larga): Una secuencia $S = (s_1, s_2, \ldots, s_n)$ es una subsecuencia de otra secuencia $X = (x_1, x_2, \ldots, x_m)$, si existe una estricta secuencia en aumento $(i_1, i_2, \ldots, i_k)$ de los índices de X tal que para todo $j = 1,2,\ldots,k$, existe $x_{ij} = s_j$. Dadas dos secuencias X e Y, la Subsecuencia Común más Larga (SCL) de X e Y es la subsecuencia común con longitud máxima. Cuando SCL se aplica en la evaluación

de resúmenes, una oración del resumen es vista como una secuencia de palabras. Intuitivamente, la SCL de dos oraciones es la más similar de dos resúmenes *X* e *Y*, donde *X* es de longitud *m* e *Y* de longitud *n*, suponiendo que *X* es una oración del resumen e *Y* es una oración del resumen candidato.

- *ROUGE*-W (subsecuencia ponderada o pesada más larga): Dadas dos secuencias *X* e *Y*, SCL se llama ponderada o pesada si la longitud es calculada usando una función de pesado. Para más detalles sobre la función ponderada ver Lin (2004).
- *ROUGE*-S (co-ocurrencia de bigramas no contiguos): Un bigrama no contiguo es cualquier par de palabras en el orden de la oración que permite un número arbitrario de espacios. La co-ocurrencia de bigramas no contiguos mide estadísticamente la cobertura de los bigramas no contiguos, entre el resumen candidato y el conjunto de resúmenes de referencia.

Lin y Hovy (2003) indicaron que este tipo de medidas se puede aplicar para la evaluación de la calidad de los resúmenes generados automáticamente, ya que lograron 95% de correlación entre juicios humanos.

Para cada métrica del sistema *ROUGE*, se obtienen indicadores de precisión, recuerdo y *f-measure*.

2.3.4 Método de pirámides

El método de evaluación basado en pirámides (en inglés, *Pyramid*) (Nenkova *and* Passonneau, 2004) fue desarrollado por la universidad de Columbia; se basa en la observación de que, los humanos, al realizar un resumen de texto no siempre seleccionan los mismos elementos. Para aplicarlo, los resúmenes generados automáticamente se fragmentan en unidades informativas denominadas *Summarization Content Units* (SCU) y se identifican segmentos similares entre los textos asignando diferentes pesos a cada segmento de información, según el número de resúmenes *gold standard* en el que aparece. Se construye una pirámide de SCU, cuya altura será igual al número de resúmenes de referencia (*gold standard*) considerados. A cada SCU de una capa se le asigna un peso que depende del número de resúmenes en los que aparece, de manera que las SCU de mayor importancia se sitúan en la cúspide de la pirámide. Si es el número de SCU de un resumen que aparezca en el nivel entonces el peso del resumen D se calcula utilizando la ecuación.

$$D = \sum_{i=1}^{n} i \times D_i \tag{5}$$

De este modo, el mejor resumen será aquel que contenga más SCU de los niveles superiores. La **figura 2.1** ilustra un posible ejemplo de evaluación utilizando una pirámide de tres niveles de altura.

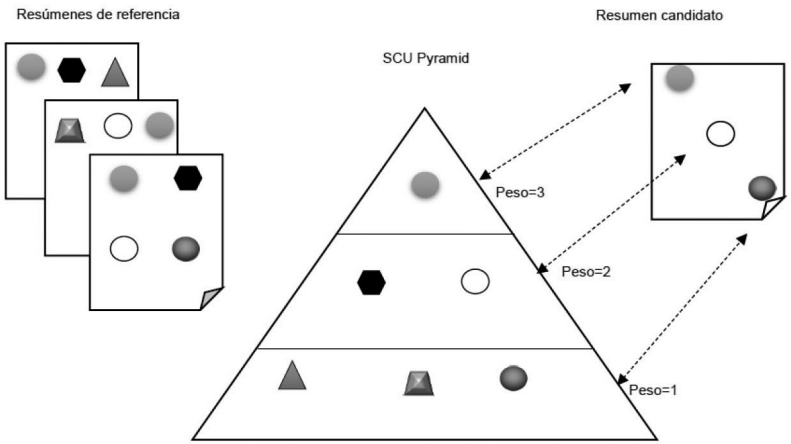


Figura 2.1 Evaluación de resúmenes con método pirámide (Nenkova and Passonneau, 2004)

Capítulo III

Métodos para la generación automática de resúmenes

En este capítulo se presentan los dos principales tipos de resúmenes según su estrategia de condensación, los abstractivos y los extractivos. Se da una descripción de cada uno, así como de algunos métodos científicos novedosos que trabajan de forma abstractiva. Además, se muestra una tabla con las características más importantes que se usan para la tarea de GART, principalmente respecto de los extractivos.

Entre los métodos propuestos para la GART están los abstractivos, que necesitan una gran cantidad de recursos lingüísticos (Miranda, 2013), (Lloret *and* Palomar, 2011), (Mateo *et al.*, 2003), por lo que tienen una alta dependencia del lenguaje o requieren de procesos sofisticados para poder generar un resumen. También están los métodos extractivos, que sólo utilizan la estructura y distribución del texto original, por lo que son menos dependientes del lenguaje (Ledeneva, 2008), (Ledeneva *et al.*, 2011), (Mihalcea *and* Tarau, 2005), (Last *and* Litvak, 2010), (García-Hernández *and* Ledeneva, 2013), (Mendoza *et al.*, 2014), entre otros. Los métodos que trabajan con un solo lenguaje pueden mostrar mejores resultados que los que son aplicables a varios. Sin embargo, la investigación de los métodos científicos novedosos se ha enfocado en la construcción de aquellos que trabajen con diferentes lenguajes, debido a su amplia gama de aplicaciones y por el crecimiento exponencial de la información.

3.1 Métodos abstractivos

Los métodos para la GART por abstracción se basan en la comprensión y reescritura del texto. Para la construcción de resúmenes de tipo abstractivo, según Plaza (2011), se pueden distinguir tres etapas:

1. Construcción de una representación semántica de las oraciones del documento.
2. Realización de operaciones de selección, agregación y generalización sobre estas representaciones.
3. Finalmente, la traducción de la representación al lenguaje natural.

Actualmente, existen muy pocas investigaciones sobre este tipo de resúmenes debido a la complejidad y al alto costo computacional que se requiere para su construcción. A continuación, se describen algunos de los métodos abstractivos.

3.1.1 Sistema Summons

Summons (McKeown *and* Radev, 1995) es un método para la GART empleado en múltiples documentos. Está basado en la creación de plantillas para todos los artículos relacionados, con información del evento de la noticia. Se realiza un *clustering* con las plantillas para identificar los temas principales, y estos *clusters* pasan a la etapa de generación para combinarse. En esta etapa hace uso de dos componentes:

1. Un planeador de contenido que genera la representación conceptual del significado del texto.
2. Un componente lingüístico, donde se seleccionan las palabras adecuadas para referirse a los conceptos contenidos en la información elegida. También utiliza varios recursos lingüísticos, como diccionarios léxicos, gramáticas, ontologías, bases de conocimiento, entre otras. Finalmente, hace uso de todos los recursos para construir el resumen.

3.1.2 Cut and paste

El método Cut and paste (Jing, 2001) es aplicado a diferentes dominios para un solo documento. En la primera parte, se extraen las oraciones más importantes del texto: se utilizan las relaciones léxicas entre las palabras para identificarlas, y se incorporan medidas estadísticas usadas en Recuperación de Información, entre algunas características del documento. En la segunda parte, se basa en dos módulos para reducir las frases extraídas:

1. Reducción de oraciones, donde se eliminan frases confusas de las oraciones extraídas (se hace uso de recursos como WordNet, conocimiento sintáctico, frases creadas por humanos), y
2. Combinación de oraciones extraídas (hace uso de las reglas identificadas en resúmenes generados por humanos).

3.1.3 Grafos conceptuales

Miranda (2013) propone un método para GART abstractivos basado en grafos conceptuales, el cual requiere información semántica (utiliza WordNet), además de patrones verbales (utiliza VerbNet) para mantener la coherencia y estructura del grafo. En general, el método consiste en los procesos de ponderación y poda de los nodos de los grafos conceptuales (síntesis), apoyados con las operaciones de generalización y unión. El proceso de ponderación se basa en la estructura y los flujos semánticos de los grafos conceptuales ponderados (Miranda-Jiménez *et al.*, 2013), así como en el algoritmo HITS[5] de Kleinberg (1999), (Mihalcea, 2004) para

[5]HITS: (Hiperlinked Induced Topic Search) Es un algoritmo iterativo diseñado para la clasificación de páginas web de acuerdo con su grado de "autoridad". Además, es un algoritmo que hace una distinción entre las "authorities" (páginas con un gran número de enlaces entrantes) y "hubs" (páginas con un gran número de enlaces salientes).

determinar la importancia de los nodos. El proceso de poda considera la información de ponderación del algoritmo HITS y utiliza los patrones verbales de VerbNet para mantener la coherencia en las estructuras durante el proceso de eliminación de los nodos. Los grafos resultantes después de aplicar las operaciones, se consideran la representación del resumen a nivel conceptual.

Los métodos abstractivos son difíciles de construir, además de ser empleados para un lenguaje específico debido al uso de plantillas o diccionarios que se están trabajando.

3.2 Métodos extractivos

Los métodos extractivos son actualmente los más investigados por su bajo costo y fácil implementación computacional, además de su utilidad para los humanos cuando están buscando información de manera rápida. Por esto, muestran la información más relevante sin cambiar la del texto original.

Los métodos extractivos para la GART consideran la estructura y distribución de las oraciones para poder seleccionar las más importantes. En la **tabla 3.1** se muestran las características del texto más utilizadas para la generación de resúmenes de tipo extractivos. El objetivo principal de estos métodos es saber cuál de las oraciones tiene mayor peso, lo que indica si puede o no pertenecer al resumen.

Según Ledeneva (2017), los resúmenes de tipo extractivo se caracterizan por estar formados bajo las siguientes etapas:

- **Selección de términos.** Durante esta etapa uno debe decidir qué unidades contarán como términos, por ejemplo, pueden ser palabras, n-gramas u oraciones.
- **Pesado o ponderación de términos.** Se trata de un proceso de ponderación (o estimación) de los términos individuales con respecto al contenido del documento.
- **Pesado o ponderación de oraciones.** Es el proceso de asignación de una medida numérica de utilidad a la oración. Por ejemplo, una de las maneras de estimar la utilidad de una oración es sumar los pesos de utilidad de los términos individuales de los cuales se compone la oración.
- **Selección de oraciones.** Se seleccionan oraciones u otras unidades como partes finales del resumen. Una de las formas más sencillas para lograrlo es asignar a las

oraciones alguna medida numérica que refleje su utilidad dentro del texto original, y sólo elegir las mejores al elaborar el resumen.

Los métodos presentados en este libro para cada uno de los lenguajes son de tipo extractivo y se explican en cada sección del lenguaje en donde se probaron.

3.2.1 Métodos independientes del lenguaje

Actualmente para la tarea de GART el inglés es el más investigado. Esto debido a que es la primera lengua más utilizada en Internet y, por ende, de ésta se han generado los *corpus*, competencias y conferencias.

Sin embargo, regularmente, cuando realizamos una búsqueda de información en Internet, lo hacemos en el lenguaje que dominamos y esperamos que la información obtenida se encuentre en éste, pero no siempre es así, por lo regular debemos de tener más de una opción en los lenguajes que podemos dominar.

Debido a esta necesidad del humano por acceder a información que esté en su lenguaje, los métodos científicos novedosos se han enfocado en el estudio de otros idiomas y han aumentado las investigaciones sobre la GART aplicados a diferentes de ellos.

Un método de GART independiente del lenguaje, según Plaza (2010), consiste en uno que, teniendo como entrada un texto base en cierto lenguaje, genere el resumen en éste y posteriormente se traduzca a diferentes. Sin embargo, otros autores, como Patel (*et al*., 2007), (Mihalcea *and* Tarau, 2005), (Wang *and* Cardie, 2013) y (Last *and* Litvak, 2010) dicen que un método de GART que trabaja con diferentes lenguajes consiste en que, teniendo una colección de documentos multilingües (escritos en varios lenguajes), se genere el resumen mediante una única herramienta. Un requisito importante para cualquier método que trabaje con diferentes lenguajes es que demuestre un funcionamiento igual en diversos de ellos sin adaptaciones especiales, como modificaciones al algoritmo o datos adicionales de cada lengua.

Cuando se habla de hacer resúmenes para varios lenguajes se complica el proceso, ya que las características de cada uno de ellos son diferentes; sin embargo, si se utilizan métodos de tipo estadísticos (extractivos) se pueden simplificar los problemas. A continuación, se describen algunos de los métodos independientes del lenguaje.

A language independent approach to multilingual text summarization

El método propuesto por Patel (2007) es de tipo extractivo, independiente del lenguaje y para un solo documento; está basado en un algoritmo genético, el cual considera los factores estructural y estadístico para la generación de resúmenes en los idiomas: inglés, hindi, gujarati y urdu. Para el pesado de las oraciones, se utiliza el contenido de información de una oración, su índice de referencia y la ubicación de las características. Para el inglés, se probó utilizando el *corpus* DUC02. Para los otros lenguajes se usan noticias propias en cada uno de ellos.

***Essential** Summarizer*

Es un método independiente del lenguaje que puede trabajar en veinte diferentes. Está basado en el análisis estadístico del texto, y para la construcción de los resúmenes hace uso de técnicas como reconocimiento de señales semánticas, especialización por dominio y consideración de expresiones, o conceptos que son importantes para el usuario (Lehmam, 2010).

Using a Keyness Metric for Single and Multi-Document Summarization

Es un método para la GART para uno y múltiples documentos en inglés y en árabe. Está basado en la frecuencia de las palabras y cálculos de probabilidad. El método utiliza dos etapas principales: el cálculo de *Log_Likelohood*[6] y el proceso de elaboración del resumen con respecto a los resultados del *Log_Likelohood* (El-Haj *and* Rayson, 2013).

Existen métodos científicos novedosos que dicen ser independientes del lenguaje, pero solamente prueban con una colección de documentos regularmente en inglés (García-Hernández *and* Ledeneva, 2013), (Ledeneva *et al.*, 2011), (Ledeneva *and* García-Hernández, 2013), (Mendoza *et al.*, 2014), entre otros.

La **tabla 3.2** es el resultado de una investigación sobre los principales métodos científicos novedosos, cuyo objetivo es presentar los parámetros utilizados más importantes, como son: lenguaje que utilizan, la colección con la que prueban, si es para uno o múltiples documentos y si es de tipo abstractivo o extractivo.

[6]Es una función de los parámetros de un modelo estadístico.

Tabla 3.1 Características del texto utilizadas para la GART

No.	Características del texto/ Referencia	(Mendoza *et al.*, 2014)	(Bossard *et al.*, 2008)	(Ouyang *et al.*, 2010)	(Nandhini and Balasundaram, 2014)	(Lin, 1999)	(Hirao *et al.*, 2002)	(Katragadda *et al.*, 2009)	(Uddin *and* Khan, 2007)	(Oräsan, 2003)	(Berker, 2011)	(Alfonseca and Rodríguez, 2003)	(Suanmali *et al.*, 2011)	(Qazvinian *et al.*, 2008)	(Mateo *et al.*, 2003)	(Babar *and* Patil, 2015)	(Kiyoumarsi, 2015)	Total
1	Posición de la oración	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	14
2	Longitud de las oraciones	✓	✓		✓	✓	✓		✓		✓	✓	✓			✓	✓	11
3	Relación de la oración con el título	✓	✓		✓	✓			✓	✓			✓	✓	✓	✓	✓	11
4	Temática (frecuencia) – cobertura	✓				✓			✓		✓		✓		✓	✓		7
5	Nombres propios				✓	✓							✓		✓	✓	✓	6
6	Datos numéricos					✓			✓		✓		✓			✓		5
7	Centralidad		✓		✓						✓		✓				✓	5
8	Similitud con una consulta		✓			✓						✓			✓			4
9	Frases de referencia				✓					✓	✓						✓	4
10	Cohesión/Similitud	✓												✓			✓	3
11	Similitud con los fragmentos		✓			✓								✓				3
12	Palabras de activación -Trigger words				✓		✓								✓			3
13	Nombre de entidades						✓				✓							2
14	Peso del término															✓	✓	2
15	Sentimiento		✓															1
16	Similitud con la primera oración		✓															1

(*continua*)

(*continuación*)

No.	Características del texto/ Referencia	(Mendoza *et al.*, 2014)	(Bossard *et al.*, 2008)	(Ouyang *et al.*, 2010)	(Nandhini and Balasundaram, 2014)	(Lin, 1999)	(Hirao *et al.*, 2002)	(Katragadda *et al.*, 2009)	(Uddin *and* Khan, 2007)	(Orăsan, 2003)	(Berker, 2011)	(Alfonseca and Rodríguez, 2003)	(Suanmali *et al.*, 2011)	(Qazvinian *et al.*, 2008)	(Mateo *et al.*, 2003)	(Babar *and* Patil, 2015)	(Kiyoumarsi, 2015)	Total
17	Longitud de la palabra				✓													1
18	Palabras polisílabas				✓													1
19	Ocurrencia de sustantivos				✓													1
20	Pronombre y adjetivo					✓												1
21	Día de la semana y el mes					✓												1
22	Cita					✓												1
23	Tipografía del texto														✓			1
24	Similaridad de oración con oración															✓		1
25	Indicador de conceptos principales																✓	1
26	Ocurrencia de información no esencial																✓	1

Tabla 3.2 Métodos científicos novedosos para la GART

Método	Colección	Lenguaje	Evaluación	Documentos	Tipo
(Mihalcea, 2004)	DUC02	Inglés	*ROUGE*	Único	Extractivos
(García-Hernández *and* Ledeneva, 2013)	DUC02	Inglés	*ROUGE*	Único	Extractivos
(Mendoza *et al.*, 2014)	DUC02, DUC01	Inglés	*ROUGE*	Único	Extractivos
(Mendoza Becerra, 2015)	DUC01, DUC02, DUC05, DUC06	Inglés	*ROUGE*	Único y múltiples	Extractivos
(Ledeneva *and* García-Hernández, 2017)	DUC02	Inglés	*ROUGE*	Único	Extractivos
(Krishna *and* Reddy, 2016)	DUC02	Inglés	*ROUGE*	Único	Extractivos
(Igave *and* Gaikwad, 2016)	DUC05	Inglés	*ROUGE*	Múltiples	Extractivos
(Wang *et al.*, 2017)	DUC04	Inglés	*ROUGE*	Múltiples	Extractivos
(Al Saied *et al.*, 2017)	DUC07	Inglés	*ROUGE*	Único	Extractivos
(Lynn et al., 2017)	600 noticias de los periódicos CNN, BBC, UK, TechCrunch and New York Times	Inglés	*ROUGE*	Único	Extractivos
(Bhargava *et al.*, 2016)	50 documentos de DUC y 51 de Amazon	Inglés	*ROUGE*	Único	Abstractivos
(Bing *et al.*, 2015)	TAC 2011	Inglés	*ROUGE*	Múltiples	Abstractivos
(Miranda-Jiménez *et al.*, 2013)	DUC03	Inglés	Precisión, recuerdo y *f-measure*	Único	Abstractivos
(Genest *and* Lapalme, 2011)	TAC 2010	Inglés	*Pyramid*	Múltiples	Abstractivos
(Khan *et al.*, 2018)	DUC03	Inglés	*ROUGE*	Múltiples	Abstractivos
(Mihalcea *and* Tarau, 2005)	DUC02, TeMário	Inglés y portugués	*ROUGE*	Único	Extractivos
(Patel *et al.*, 2007)	DUC02, hindi	Inglés, hindú, gujarati y urdu	Evaluación intrínseca	Único	Extractivos
(Villatoro E., 2007)	Desastres, CAST	Español, inglés	*ROUGE*	Múltiples	Extractivos
(Last *and* Litvak, 2010)	DUC02, hebreo	Inglés y hebreo	*ROUGE*	Único	Extractivos
(Saggion, 2011)	TAC multilingüe 2011	Árabe, inglés, francés e hindú	*ROUGE*	Múltiples	Extractivos
(Last *and* Litvak, 2010)	DUC02, hebreo	Inglés, hebreo y árabe	*ROUGE*	Único	Extractivos
(El-Haj *and* Rayson, 2013)	Multiling 2013	Inglés, árabe	*ROUGE, AutoSumm-ENG, MeMoG* y *NPowER*	Único y múltiples	Extractivos
(Mingli *et al.*, 2016)	200 documentos sobre tecnología	Chino	*ROUGE*	Único	Extractivos

Capítulo IV

Herramientas para la generación automática de resúmenes

Este capítulo está dedicado al análisis de las herramientas comerciales para la GART. Se describe el método que utiliza cada una de ellas y los pasos que se deben realizar para su funcionamiento, con el objetivo de tener un panorama de la calidad de estas herramientas respecto de las heurísticas y los métodos científicos novedosos.

Las herramientas comerciales son aquellas que se encuentran disponibles para su uso ya sea en Internet (línea) o disponibles para su instalación en una computadora. Generalmente, el método con el que trabajan no es publicado ya que la herramienta tiene un costo y su funcionamiento interno no es de dominio público. Las herramientas comerciales se clasifican en instalables y en línea; las primeras son aquellas que necesitan ser instaladas en una computadora para su funcionamiento, mientras que a las segundas se puede acceder mediante cualquier computadora que cuente con Internet.

El objetivo principal de este capítulo es presentar las herramientas comerciales instalables y en línea que se han utilizado para la tarea de GART, además de mostrar su funcionamiento.

4.1 Herramientas instalables

4.1.1 *Copernic Summarizer*

Esta herramienta fue desarrollada exclusivamente para la GART, es flexible y adecuada para esta tarea pues ofrece diferentes opciones para la longitud del resumen a generar, entre las que están: 5%, 10%, 25% y 50% del número de palabras del documento original, que puede tener una extensión de 100, 250 y 1000 de ellas.

De acuerdo con Copernic Summarization-Technologies White Paper (2003), *Copernic Summarizer* utiliza los siguientes métodos:

1. Modelo estadístico (S-Model). Se utiliza para encontrar el vocabulario del texto.
2. Procesos intensivos de conocimiento (K-Process). Considera la forma en que los humanos hacen resúmenes de texto, teniendo en cuenta las siguientes etapas.
   a) **Detección del lenguaje.** Detecta el lenguaje (inglés, alemán, francés o español) del documento para aplicar procesos específicos.
   b) Reconocimiento del límite de las oraciones y "tokenización". Implementa varias heurísticas como identificar listas con viñetas y cadenas especiales (correo electrónico y fórmulas científicas), para aislar a las oraciones.
   c) Extracción de conceptos. *Copernic Summarizer* utiliza técnicas de aprendizaje automático para extraer palabras claves.

d) Segmentación del documento. Organiza la información que se puede dividir en segmentos más grandes relacionados.
e) Selección de oraciones. Las oraciones se seleccionan según su importancia (peso), descartando las que disminuyen la legibilidad y la coherencia.

A continuación, se presentan los pasos que se siguen para realizar un resumen con la herramienta *Copernic Summarizer*, los cuales se enumeran en la **figura 4.1**.

1. Se pega el texto a resumir.
2. Posteriormente, se selecciona la opción de palabras a resumir.
3. Automáticamente, la herramienta genera el resumen. Se puede imprimir, guardar y enviar la información.

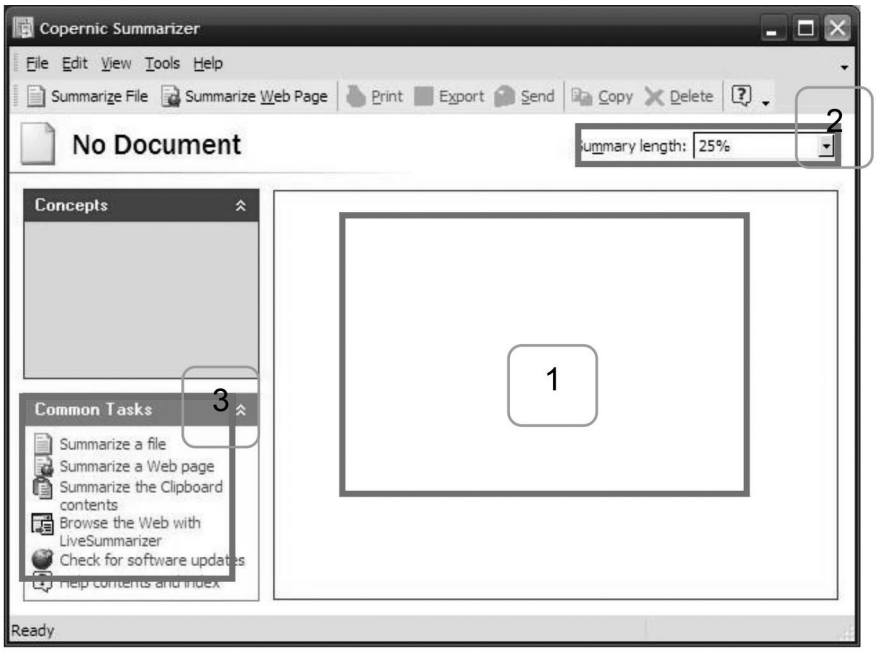


Figura 4.1 Interfaz de la herramienta *Copernic* Summarizer

4.1.2 *Microsoft Office Word Summarizer*

Microsoft Office Word es un procesador de textos que permite crear documentos a los que se le pueden añadir imágenes, gráficos, tablas y un sinfín de objetos que hacen más atractivos los trabajos (Villar, 2005); tiene la opción de GART en las versiones 2003 y 2007. El resumen creado por *Microsoft Office Word* es el resultado de un análisis de palabras clave y la selección de las más frecuentes en el documento. Las oraciones que contienen estas palabras son incluidas en el resumen; asimismo, este programa permite generar resúmenes de 10 o 20 oraciones; 100 o 500 palabras (o menos); o bien, en porcentajes de 10%, 25%, 50% y 75% de palabras del documento original. Si algunos de los porcentajes no son adecuados, el usuario lo puede cambiar según sus necesidades. A continuación, se describe la forma en que se realizan los resúmenes en las versiones *Microsoft Office Word* 2003 y 2007.

- ***Microsoft Office Word* 2003**

A continuación, en la **figura 4.2** se describen los pasos a seguir para la GART con la herramienta *Microsoft Office Word* 2003.

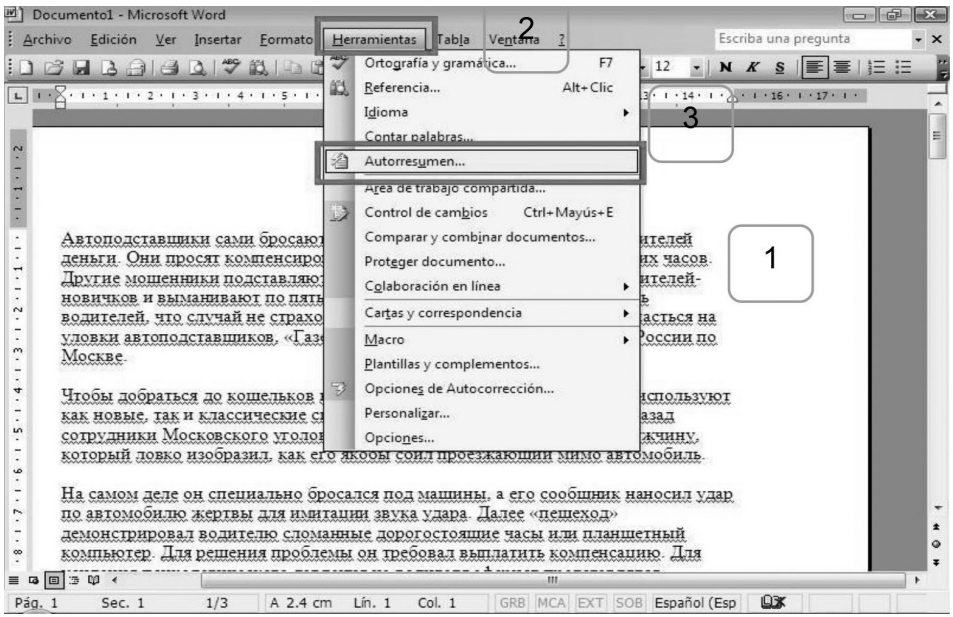


Figura 4.2 Interfaz para activar la opción Autorresumen

1. Se coloca el texto a resumir.
2. Se da clic en la pestaña Herramientas.
3. Se selecciona la opción Autorresumen…

Automáticamente, se muestra un recuadro con los parámetros de Microsoft Word en los que el usuario puede seleccionar la opción adecuada de representación del resumen de acuerdo con sus requerimientos.

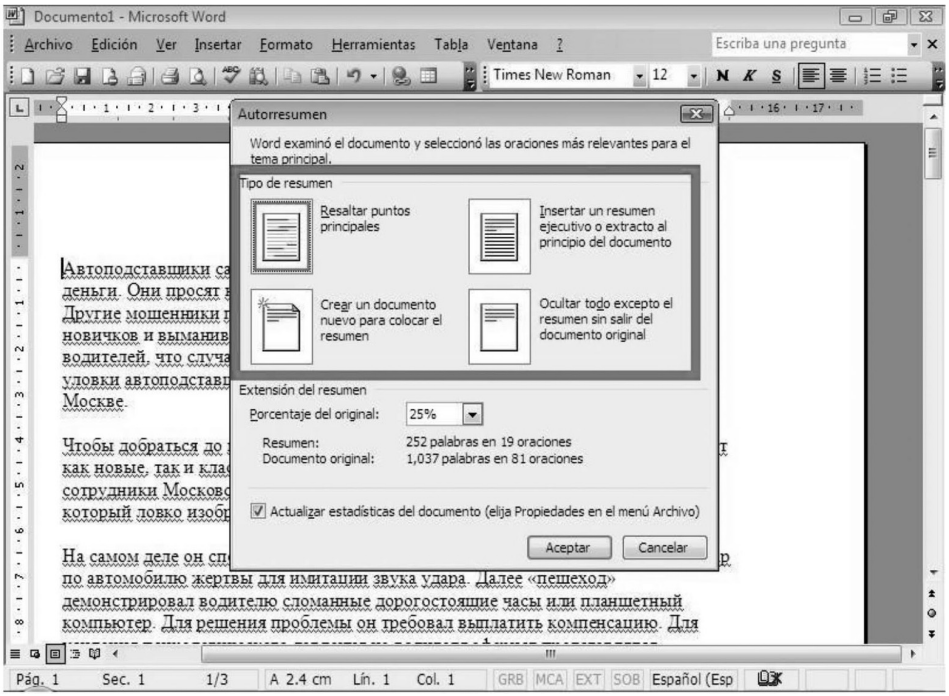


Figura 4.3 Interfaz para seleccionar parámetro de tipo de resumen

Como se ve en la **figura 4.3**, el primer parámetro que se puede modificar es la forma de salida del resumen, las opciones son: Resaltar puntos principales, Insertar un resumen ejecutivo o extracto al principio del documento, Crear un documento nuevo para colocar el resumen y Ocultar todo excepto el resumen sin salir del documento.

El segundo parámetro por modificar es la extensión del resumen. Las opciones que el usuario puede seleccionar son: el tamaño del texto que requiera,

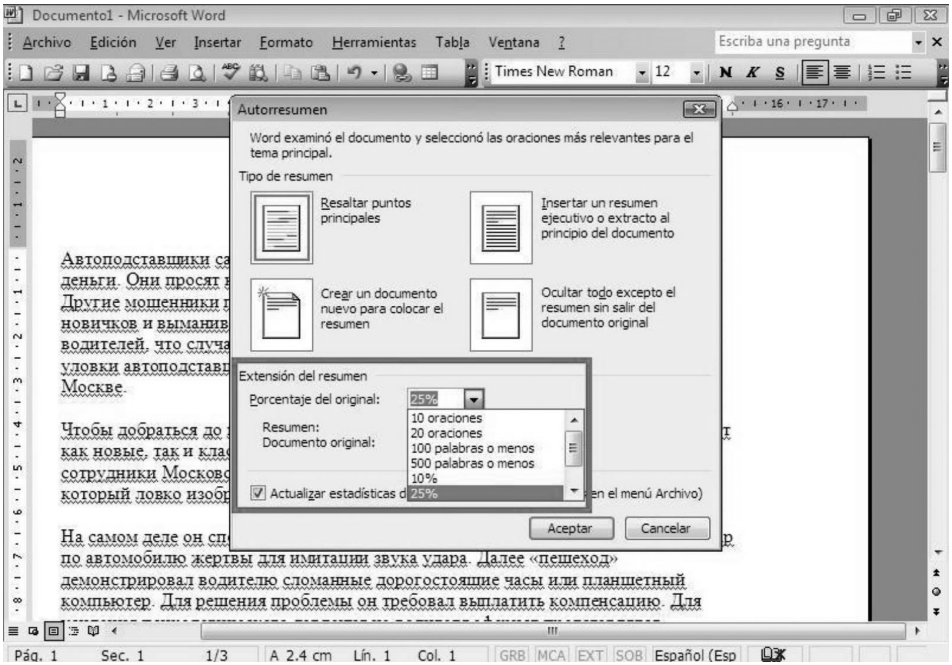


Figura 4.4 Interfaz para seleccionar parámetro de la extensión del resumen

el cual puede ser: por oraciones (10 o 20), por número de palabras (100 o menos, 500 palabras o menos) o porcentaje (10%, 25%) (**Figura 4.4**). Finalmente, se da clic en el botón Aceptar para generar el resumen automático.

- ***Microsoft Office Word* 2007**

La GART en *Microsoft Office Word* 2007 es similar a la de 2003. Sin embargo, se requiere activar la opción Autorresumen para generar los resúmenes. A continuación, se muestran los pasos a seguir para activar esta opción.

1. Se selecciona el botón Office (parte superior izquierda).
2. Se selecciona el botón Opciones de Word (**figura 4.5**).
3. Se selecciona la opción Personalizar (parte derecha superior-intermedia, **figura 4.6**).
4. Se ubica la opción Comandos disponibles en:
5. Se selecciona la opción Todos los comandos
6. Se selecciona la opción Autorresumen...
7. Se selecciona el botón Agregar >>

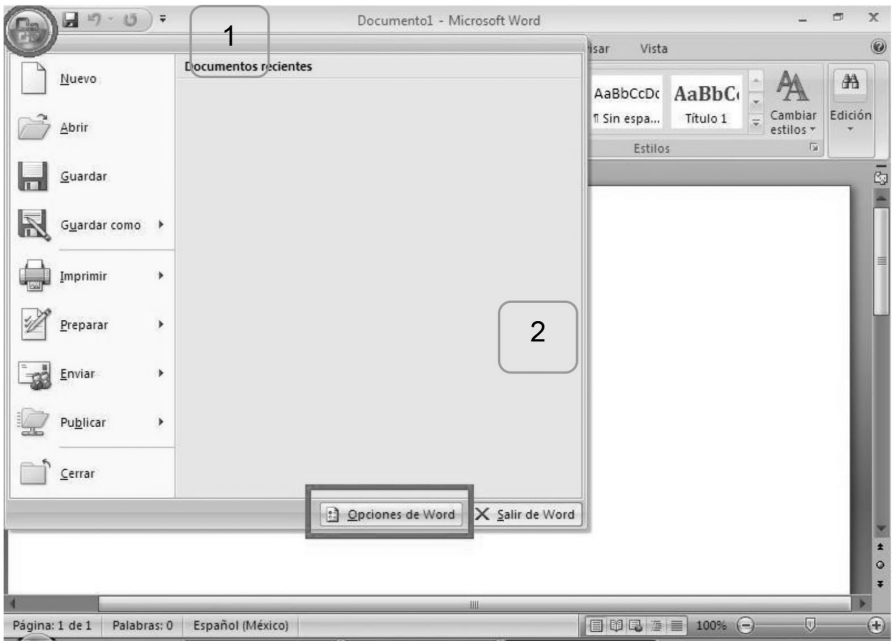


Figura 4.5 Interfaz para seleccionar la opción Autorresumen

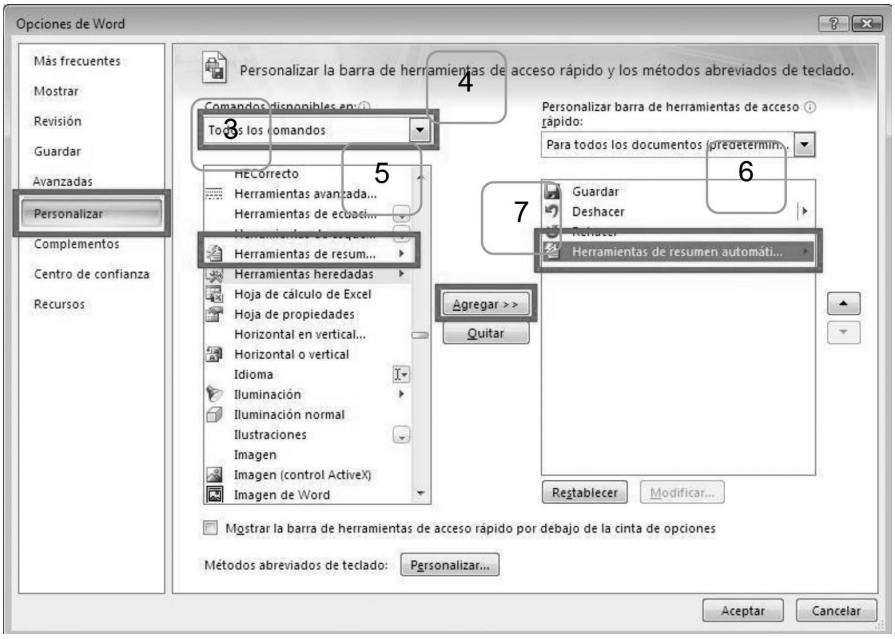


Figura 4.6 Interfaz para activar opción Autorresumen

Una vez agregada la opción de Autorresumen, ésta aparece en la barra de herramientas (**figura 4.7**), con la cual se puede acceder a la ventana con las opciones para generar un resumen.

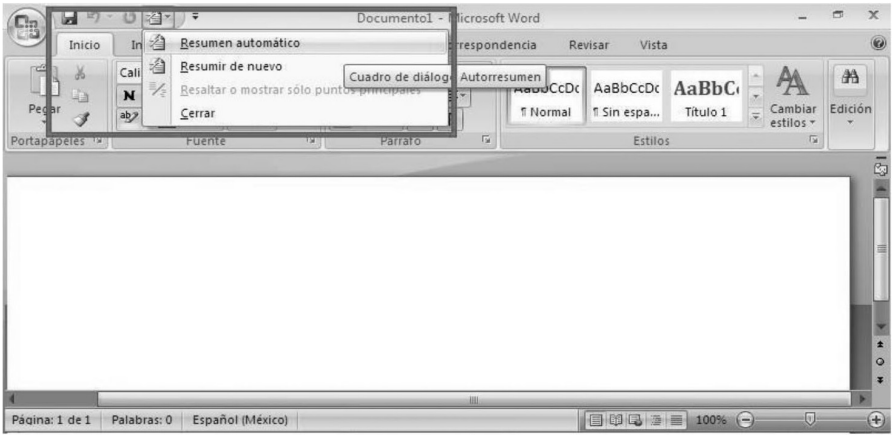


Figura 4.7 Botón de activación de la opción Autorresumen

4.2 Herramientas en línea

4.2.1 *SweSum*

SweSum (Hassel *and* Dalianis, 2003) fue el primer programa para resumir textos para sueco. Sin embargo, actualmente trabaja con los lenguajes: inglés, español, francés, noruego, italiano, danés, griego, farsi (persa) y alemán. Los aspectos que utiliza para valorar las oraciones son su posición y sus valores numéricos.

El proceso de generación de resúmenes de *SweSum* consta de tres etapas:

1. Se realiza la "tokenización", fragmentación del texto en oraciones y la extracción de palabras claves.
2. Se elabora un *ranking* con las oraciones que aparecen con mayor frecuencia.
3. Se resume el texto.

A continuación, se muestra cómo realizar un resumen con la herramienta *SweSum* (**figura 4.8**).

1. Se escribe o se pega el texto a resumir en el recuadro, o bien, se selecciona un archivo desde la propia computadora.
2. Se elige el tipo de texto (académico o periódico).
3. Se elige el lenguaje.
4. Se asigna el porcentaje o palabras deseadas del texto a resumir.
5. Por último, se da clic en el botón *Summarize* para generar el resumen.

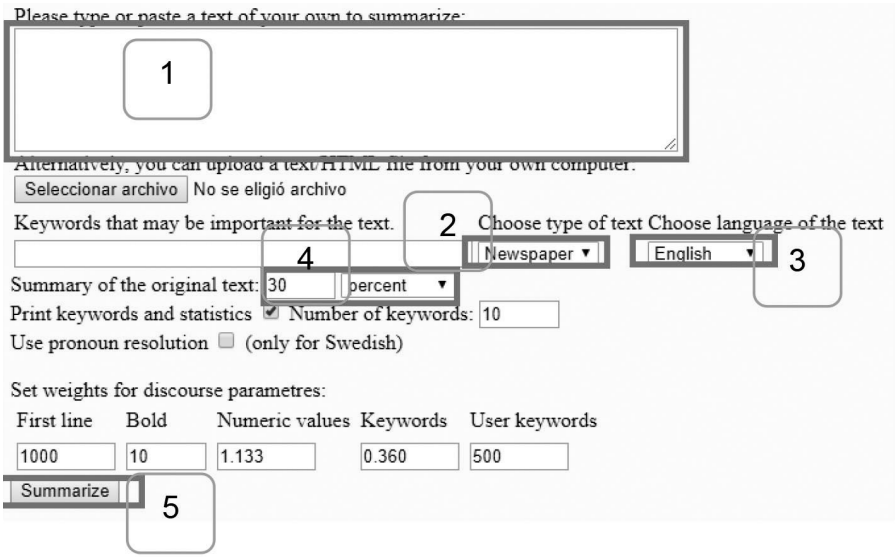


Figura 4.8 Interfaz de la herramienta *SweSum*[7]

4.2.2 *T-Conspectus*

Es una aplicación web para resumir los artículos en inglés, alemán y ruso, dentro del dominio de noticias.

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

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

[7]Proyecto sueco de resumen on-line, disponible en: http://swesum.nada.k7th.se/index-eng.html

1. La primera es el preprocesamiento, donde se realizan cuatro procedimientos principales:
   a) Título: si el texto a resumir contiene un título, éste será utilizado para la asignación de pesos adicionales a las palabras claves (es recomendable introducir textos con título).
   b) Divide el texto en párrafos: el generador de resúmenes necesita saber los límites del párrafo para encontrar su primera y última frase, y poner puntuaciones basadas en la posición.
   c) Divide los párrafos en oraciones: este proceso se divide en dos subetapas; la primera es la descomposición inicial del párrafo y luego la corrección posterior a la división del párrafo en oraciones.
   d) "Tokenización" en cada oración: la oración es dividida en palabras.
2. La segunda etapa es la de puntuación de resúmenes mediante el pesado de términos y el pesado de la oración. Se crea una lista ordenada en una tabla que contiene las oraciones con sus respectivos pesos.
3. La tercera etapa es la generación del resumen; selecciona un número "n" de las primeras frases de la lista del paso anterior. El número de oraciones que se seleccionará en el resumen final se calcula en función del usuario.

Una vez explicada la forma en que *T-Conspectus* genera los resúmenes, se describirán los pasos para hacer un resumen con esta herramienta (**figura 4.9**):

1. Se pega o se escribe el texto a resumir. Sin embargo, también se puede resumir un texto desde una URL o al seleccionar el archivo desde la computadora.
2. Se especifica por medio de valor porcentual el tamaño del texto resultante (va desde 5 hasta 70. El aumento va en una escala de 5 en 5).
3. Mediante estas opciones, se muestran las palabras claves y estadísticas del resumen.
4. El botón para generar el resumen es *Summarizer*, el cual sólo se activa si hay texto en el recuadro.
5. El botón Remove Text elimina el texto o URL contenido; este botón sólo se activa si los recuadros correspondientes tienen texto.

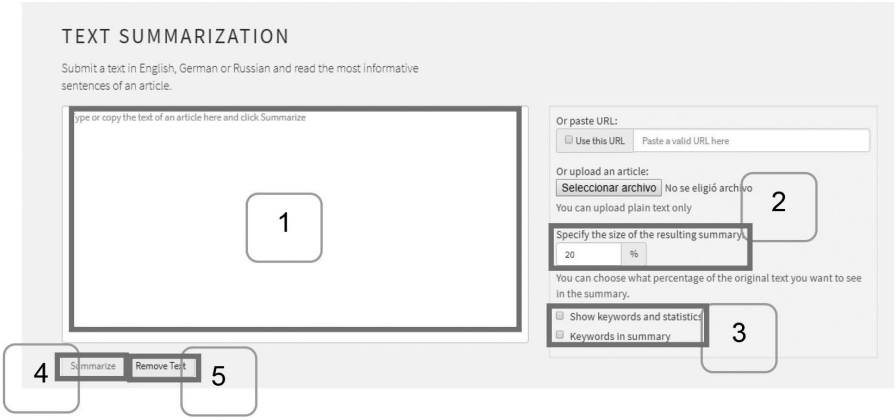


Figura 4.9 Interfaz de la herramienta comercial en línea *T-Conspectus*[8]

4.2.3 *Open Text Summarizer (OTS)*

Open Text Summarizer[9] es una aplicación de código abierto para resumir textos. Puede ser descargada de Internet de forma gratuita. Sin embargo, también puede encontrarse la interfaz en línea. OTS analiza automáticamente los textos y trata de identificar las partes más importantes; soporta varios lenguajes: inglés, alemán, español, ruso, hebreo y otros veinticinco más.

A continuación, se describen los pasos para generar un resumen en la herramienta *Open Text Summarizer* (**figura 4.10**):

1. Se pega o escribe el texto a resumir. Sin embargo, también se puede resumir un texto desde una URL.
2. Se elige el tipo de salida de la información, puede ser de dos: de forma general o para palabras claves.
3. Se elige el tamaño de salida para el resumen.
4. Se selecciona el lenguaje en el que se realizará el resumen.
5. Se da clic en el botón Enviar para generar el resumen.

[8]Es una aplicación web para resumir artículos periodísticos en inglés, alemán y ruso, disponible en: http://tconspectus.pythonanywhere.com/

[9]Esta es una interfaz web para generar resúmenes. La herramienta analiza automáticamente los textos en varios idiomas e intenta identificar las partes más importantes del texto, se encuentra disponible en: https://www.splitbrain.org/services/ots

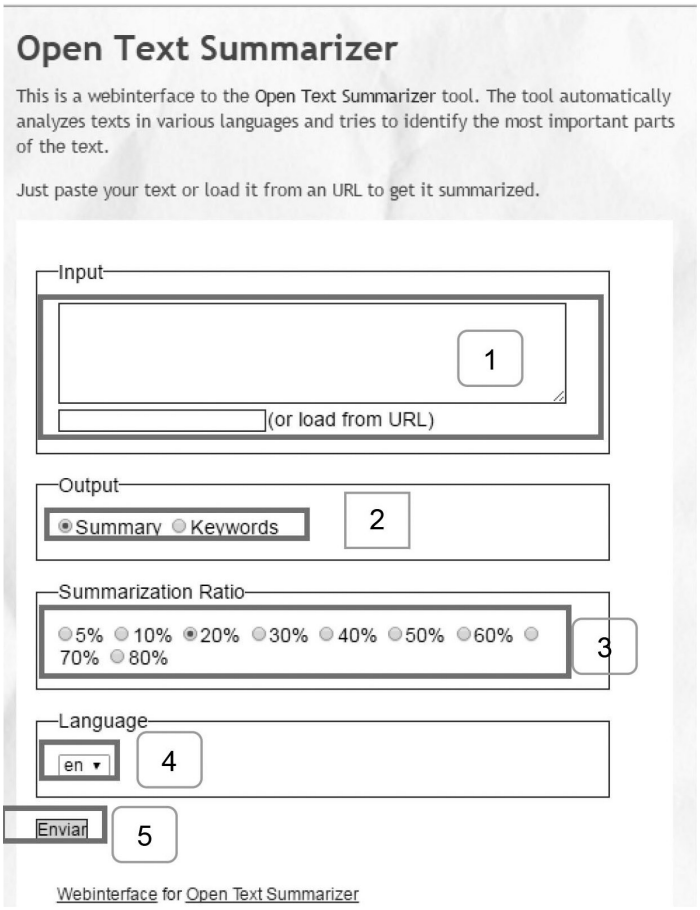


Figura 4.10 Interfaz de la herramienta comercial en línea *Open Text Summarizer*

4.2.4 *Text Compactor*

Esta herramienta para la GART es gratuita y está disponible en línea. Fue creada para ayudar a estudiantes, maestros o profesiones con dificultades para procesar cantidades grandes de información. Las etapas de la herramienta *Text Compactor* para generar el resumen son las siguientes:

1. Se calcula la frecuencia de cada palabra en el texto.
2. La puntuación se calcula para cada frase, basándose en la frecuencia asociada con las palabras que contiene.
3. Para la frase más importante se considera la de mayor frecuencia.

Cabe mencionar que esta herramienta funciona mejor para generar resúmenes en libros de texto y material de referencia. Sin embargo, no tiene el mismo funcionamiento con texto de ficción (es decir, en historias sobre personajes, lugares o eventos imaginarios).

A continuación, se describe cada uno de los pasos para generar un resumen con *Text Compactor* (**figura 4.11**):

1. Se pega o se escribe el texto a resumir.
2. Se define el tamaño del texto a resumir. El tamaño va de 0 a 100%.
3. Se presenta el resumen.

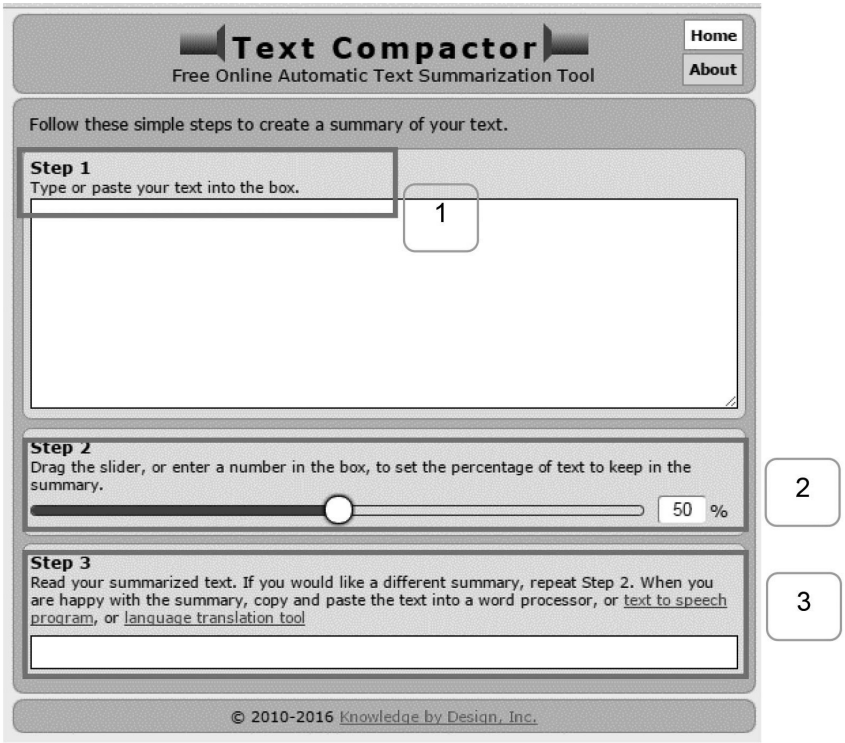


Figura 4.11 Interfaz de la herramienta comercial en línea *Text Compactor*

4.2.5 *Summarizing*

Summarizing[10] es una herramienta en línea para la GART de artículos. Las etapas que utiliza se basan en la detección de las ideas principales del texto, en obtener una descripción de las ideas, lo cual refleja el estilo de escritura del autor para, finalmente, reescribir el texto en el resumen. Esta herramienta tiene los parámetros para generar resúmenes de 100, 150, 200 y 300 palabras. La versión que se presenta es la de prueba.

A continuación, se presentan los pasos para realizar un resumen con esta herramienta (**figura 4.12**).

1. Se pega o escribe el texto a resumir.
2. Se selecciona el número de palabras que tendrá la longitud del resumen.
3. Se da clic en *Summarize* para generar el resumen.

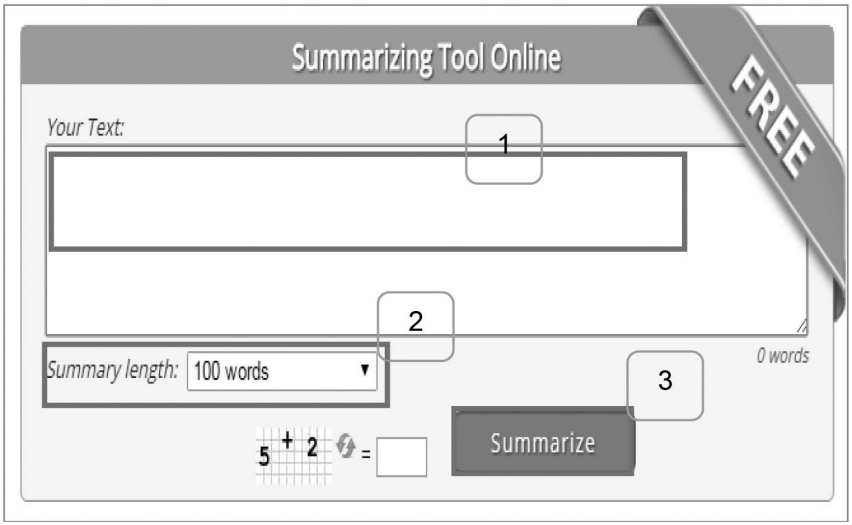


Figura 4.12 Interfaz de la herramienta *Article Summarizing Online*

[10]Es una herramienta en línea para la generación de resúmenes, disponible en su versión de prueba en: https://www.summarizing.biz/best-summarizing-strategies/article-summarizer-online/. También se puede adquirir la versión completa con un costo monetario.

4.2.6 *Summarizer*

Summarizer[11] es una herramienta que nos permite generar resúmenes automáticos. Está disponible como componente de *Intellexer API* y como aplicación de escritorio; recibe un documento de origen, extrae texto sin formato, brinda el procesamiento sintáctico y semántico, extrae la información para la generación del resumen y, finalmente, asigna un valor determinado por oraciones. Este valor define la importancia de la oración en lo que respecta a la idea del texto. *Summarizer* genera resúmenes en diferentes porcentajes (1% a 99%).

A continuación, se presentan los pasos para realizar un resumen con la herramienta en línea *Summarizer* (**figura 4.13**).

1. Se inserta la URL o se pega el texto a resumir.
2. Posteriormente, se selecciona una opción: porcentaje o párrafo. De acuerdo con la opción seleccionada, se activa la casilla.
3. Se coloca el porcentaje que se desea para resumir el texto (1-99%).
4. Por último, se da clic en el botón *Summarize*.

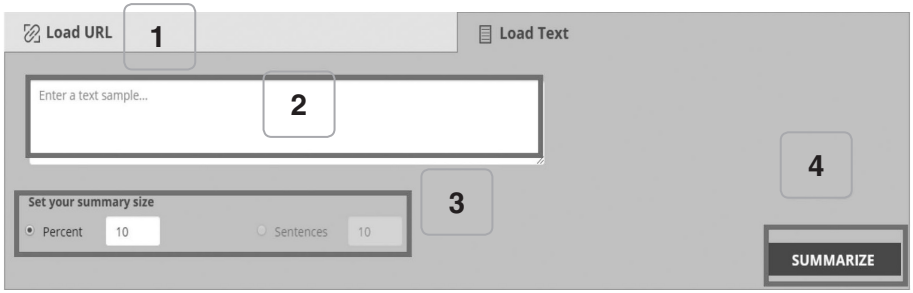


Figura 4.13 Interfaz de la herramienta *Summarizer* en línea

A continuación, se muestra la interfaz instalable de la herramienta *Summarizer* (**figura 4.14**).

[11]Es una herramienta en línea para la generación de resúmenes incluida en el conjunto de herramientas de procesamiento del lenguaje natural de Intellexer, desarrolladas por EffectiveSoft. Se puede acceder a ella desde la página web: http://esapi.intellexer.com/Summarizer

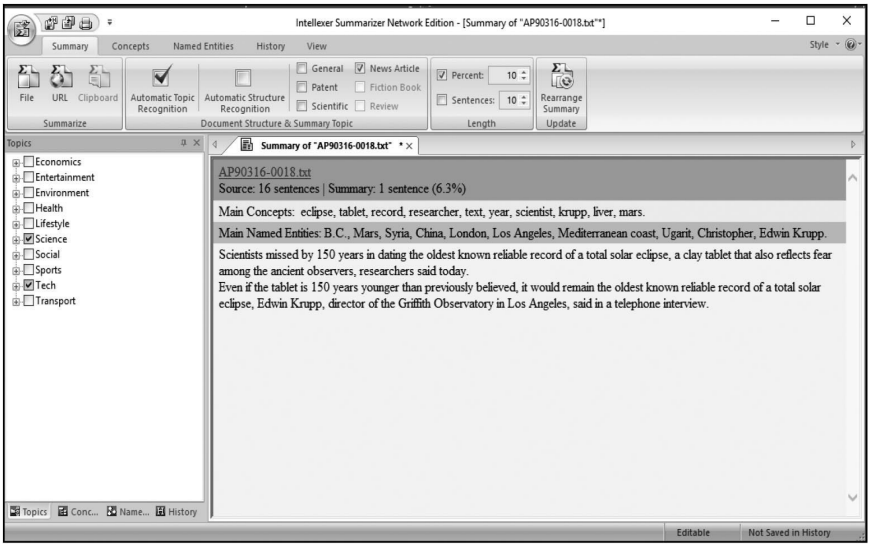


Figura 4.14 Interfaz de la herramienta *Summarizer* instalable

4.2.7 *Tools4noobs*

Tools4noobs[12] es una herramienta en línea que crea automáticamente un resumen de un texto (por lo general de grandes tamaños); ya sea que pegue el texto o la URL de la página web a resumir, para que la herramienta dé un breve resumen (**figura 4.15**).

El proceso para la GART que realiza esta herramienta consta de tres fases:

a) Extracción de frases del texto dado.
b) Identificación de las palabras clave en el texto y conteo de la relevancia de cada una de ellas.
c) Identificación de las frases con la mayoría de las palabras claves.

A continuación, se presentan los pasos para realizar un resumen con esta herramienta en línea *Tools4noobs* (**figura 4.15**).

1. Se pega o escribe el texto a resumir. También tiene la opción de colocar un URL.

[12]Es una herramienta en línea para la creación de resúmenes, que se encuentra disponible de forma gratuita en el sitio web: https://www.tools4noobs.com/summarize/

2. Se define el umbral, es decir, el tamaño que tendrá el resumen. Además, se tiene la opción de definir el número de líneas que se desea tenga el resumen, el número de caracteres y el mínimo de palabras de una oración.
3. *Tools4noobs* permite seleccionar una serie de opciones para visualizar el resumen. Entre ellas están: relevancia de la oración, remarcar las palabras clave más relevantes, número de palabras claves, resaltado de las palabras claves, mostrar las frases más destacadas en el texto.
4. El botón Summarize it!: se utiliza para generar el resumen.

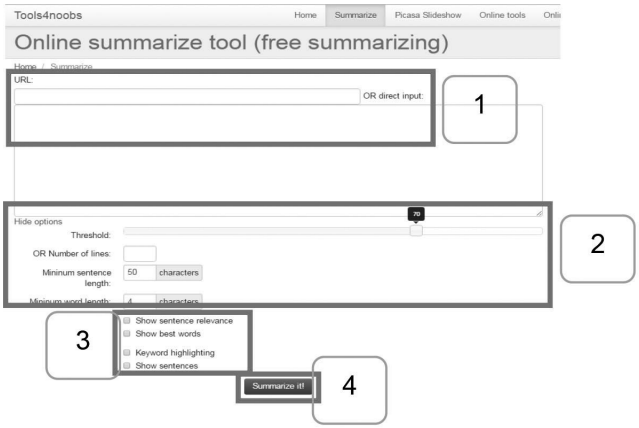


Figura 4.15 Interfaz de la herramienta comercial en línea *Tools4noobs*

4.2.8 *Pertinence Summarizer*

Pertinence Summarizer[13] pertenece a la gama de productos desarrollados con tecnología denominada KENiA© (basada en la extracción de conocimiento y arquitectura de notificación), generada por la empresa francesa *Pertinence Mining*. *Pertinence Summarizer* es una herramienta en línea que permite generar resúmenes en doce lenguajes (alemán, inglés, árabe, chino, coreano, español, francés, italiano, japonés, portugués, ruso y neerlandés) de los documentos de texto en diversos formatos (html, pdf, doc, rtf y txt).

[13]Es una herramienta en línea para la generación automática de resúmenes que actualmente ya no se encuentra disponible. Sin embargo, se deja la liga de la página a la que se accedió para probar la herramienta. http://pertinence.net/index_en.html

Los porcentajes son colocados de forma automática, así como el número de palabras [1% (34 *words*), 5% (171 *words*),…y n% (n *words*)].

Hay tres formas de introducir el texto a resumir, la primera es copiando y pegando el texto en la página; la segunda es abriendo el documento desde su origen, y la tercera es introduciendo la dirección de la página web a traducir (**figura 4.16**).

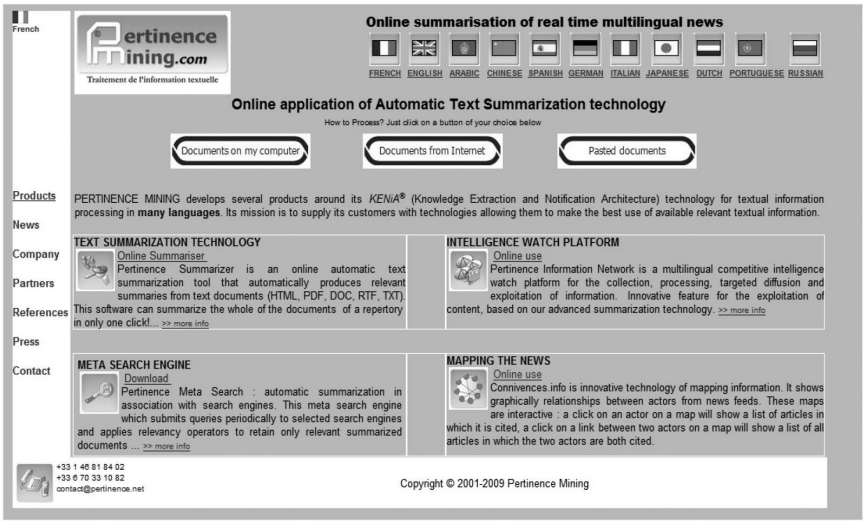


Figura 4.16 Interfaz de la herramienta *Pertinence Summarizer*

4.2.9 *Shvoong*

Shvoong[14] fue fundado en 2005 por Avi Shaked y Avner Avrahami. *Shvoong* es una herramienta que permite generar resúmenes automáticos en veintiún lenguajes diferentes. A diferencia de otras herramientas, *Shvoong* no devuelve el resumen como tal, sino que subraya el texto que considera más importante del documento original.

A continuación, se presentan los pasos para realizar un resumen con la herramienta en línea *Shvoong* (**figura 4.17**).

[14]Es una herramienta en línea para la generación de resúmenes que se encuentra disponible en la página web: http://es.shvoong.com/summarizer/

1. Se pega o se escribe el texto.
2. Se elige el lenguaje.
3. Se asigna el porcentaje que debe tener el resumen.
4. Dar clic en el botón Summarize! para generar el resumen.

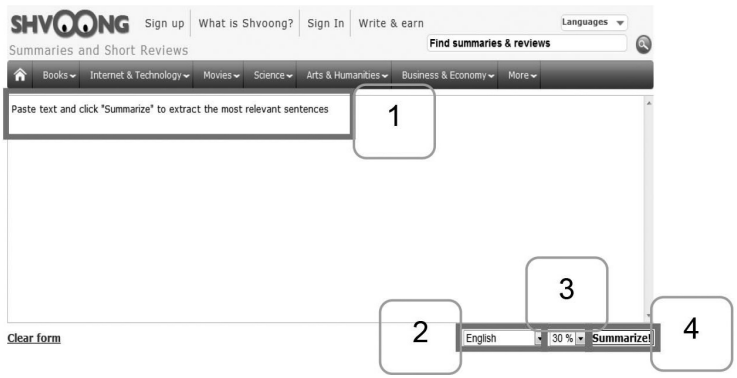


Figura 4.17 Interfaz de la herramienta *Shvoong*

4.2.10 *Resumo*

Resumo[15] es un generador de resúmenes de textos multilenguaje, la interfaz está en portugués, en la **figura 4.18** se puede apreciar.

A continuación, se muestran los pasos para generar un resumen con esta herramienta:

1. Se inserta texto para resumirlo.
2. Se elige si se desea un resumen por número de líneas o por porcentaje de texto.
3. Se selecciona el lenguaje del texto a resumir.
4. Se da clic en el botón Fazer Resumo para generar el resumen.

[15]Es una herramienta en línea para la generación de resúmenes, está disponible en su versión de prueba, sin embargo, si se desea resumir textos grandes se debe adquirir la versión de paga. Se encuentra disponible en la página web: https://www.turbinetext.com/Resumo/

Figura 4.18 Interfaz de la herramienta *Resumo*

4.2.11 *BigdataSummarizer*

BigdataSummarizer[16] es una herramienta que realiza resúmenes de textos en veintiún lenguajes, como: chino, inglés, francés, alemán, italiano, ruso, español, etc. Trabaja con 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% y 100% del porcentaje del texto original (**figura 4.19**).

A continuación, se muestran los pasos para generar un resumen con esta herramienta:

1. Se pega o se escribe el texto a resumir. También tiene la opción de colocar URL.
2. Se selecciona el lenguaje del texto origen.
3. Se selecciona el dominio del texto.
4. Se da clic en el botón Summarize para generar el resumen.

4.3 Resumen de herramientas probadas en diferentes lenguajes

En la tabla 4.1 se muestra un listado de las herramientas comerciales instalables y en línea para la GART ya descritas en las secciones anteriores. Sin embargo, se listan mencionando cuál es el lenguaje en el que fueron probadas. Los resultados se presentan en los capítulos posteriores.

[16]Es una herramienta en línea para la generación de resúmenes especialmente para el lenguaje ruso, se encuentra disponible en la página web: https://bigdatasummarizer.com/summarizer/online/advanced.jsp?ui.lang=es

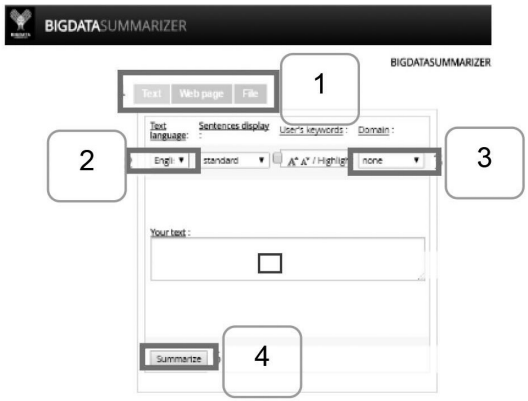


Figura 4.19 Interfaz de la herramienta *BigdataSummarizer*

Tabla 4.1 Herramientas probadas en diferentes lenguajes

Herramienta	Tipo	Inglés	Español	Portugués	Ruso
Copernic Summarizer	Instalable	✓	✓		
Microsoft Office Word 2003/2007	Instalable	✓	✓		✓
SweSum	Línea	✓			
T-Conspectus	Línea	✓			✓
OTS	Línea	✓	✓	✓	✓
Text Compactor	Línea	✓	✓		✓
Summarizing	Línea	✓	✓		
Summarizer	Línea	✓			
Tools4noobs	Línea	✓			✓
Pertinence Summarizer	Línea	✓			
Shvoong	Línea	✓		✓	
Resumo	Línea				✓
BigdataSummarizer	Línea				✓
Total		11	5	2	7

Capítulo V

Generación automática de resúmenes para el lenguaje inglés

Este capítulo está dedicado a la presentación puntualizada del estudio de la tarea de GART para el lenguaje inglés. Se describen los *corpus* DUC01 y DUC02, los cuales son utilizados para realizar estas pruebas. También se muestran los resultados de las principales heurísticas, se describen y dan los resultados de las herramientas comerciales y los métodos científicos novedosos probados para este idioma. Finalmente, se muestra una comparación general de las heurísticas, los métodos científicos novedosos y las herramientas comerciales probadas con los *corpus* DUC01 y DUC02.

El lenguaje inglés es el tercero hablado a nivel mundial, con 339 millones de personas que lo tienen como lengua nativa (Arévalo, 2017) (**tabla 5.1)**.

Tabla 5.1 Principales lenguas habladas en el mundo

No.	Lenguaje	Países	Hablantes
1	Chino	35	1302
2	Español	21	427
3	Inglés	106	339
4	Árabe	58	267
5	Hindi	4	260
6	Portugués	12	202
7	Bengalí	4	189
8	Ruso	17	171
9	Japonés	2	128
10	Lahnda	8	117
11	Javanés	3	84.3
12	Coreano	7	77.3
13	Alemán	26	76.9
14	Francés	53	75.9
15	Télugu	2	74.2
16	Marathi	1	71.4
17	Turco	8	71.4
18	Urdu	6	68.6
19	Vietnamita	3	68
20	Tamil	7	67.8
21	Italiano	13	63.4
22	Persa	30	61

Sin embargo, a pesar de estar en la tercera posición, es la lengua que más países utilizan (106 países). Por lo que a nivel mundial el inglés ocupa el primer lugar de los más utilizados en Internet (**tabla 5.2**).

Tabla 5.2 Las 10 lenguas más usadas en Internet[17]

No.	Lenguaje	Usuarios de Internet
1	Inglés	1052
2	Chino	804
3	Español	337
4	Árabe	219
5	Portugués	169
6	Hindi	168
7	Francés	134
8	Japonés	118
9	Ruso	109
10	Alemán	92
	Otra	950

La GART tiene sesenta años de investigación, su estudio comenzó en la década de 1950 con el trabajo de Luhn en 1958 (Luhn, 1958), quien fue el primero en realizar resúmenes automáticos de tipo extractivo. Posteriormente, el estudio de esa tarea continuó con los trabajos de Edmundson (1969), Kupiec (*et al*., 1995), Paice (1990), Jing (*et al*., 1998), Minel (*et al*., 1997), Barzilay, Elhadad (1999), Benbrahim *and* Ahmad (1995), Carbonell *and* Goldstein (1998), Marcu (1997), McKeown and Radev (1995), Mani (*et al*., 1999) y otros. La investigación de GART hasta antes del año 2000 estaba centrada en el lenguaje inglés debido a los recursos (*corpus* y medidas de evaluación estándar) que hay en esta tarea para este idioma. Durante todos estos años se ha dado un gran avance en la investigación con respecto a otros lenguajes. Es por ello que se pueden tomar como soporte las investigaciones en el inglés para otros lenguajes.

A lo largo de las investigaciones en el inglés para la generación de resúmenes, han surgido las diferentes heurísticas que sirven como referencia para la evaluación de los métodos y las herramientas generadoras de resúmenes. Sin embargo, no se

[17]Según un estudio que revela las lenguas más usadas en internet https://www.internetworldstats.com/stats7.htm

sabía cuánto influían estas heurísticas tanto en la construcción del resumen, como en las preferencias del humano en cuanto a la elección de un resumen. Para esto, en el Capítulo I se realizó un *Test de Turing*. En esta prueba se colocaron dos resúmenes hechos por humanos, dos de manera automática y se incluyeron las dos heurísticas: *baseline:first* y *baseline:random.*

Como se mencionó anteriormente, la heurística *baseline:random* propone como resumen la selección aleatoria de oraciones, es decir, se hace sin inteligencia. No obstante, se consideró este resumen con el objetivo de saber si el humano es capaz de distinguir este tipo de textos. Aunque esto podría representar una desventaja respecto de los resúmenes hechos por sistemas. En la **tabla 1.2** se muestran los resultados de los pares de resúmenes elegidos con respecto al humano-máquina, sin hacer distinción de las heurísticas. Sin embargo, es necesario hacer dicha distinción para determinar la influencia que presentan para el humano.

En la **tabla 5.3** se presentan los resultados del *Test de Turing* para el lenguaje inglés, pero con la subdivisión de los pares de resúmenes considerando humano-máquina (*baseline-first*), humano-máquina (*baseline-random*), máquina (método)-máquina (método), máquina (método)-máquina (*baseline-first*), máquina (método)-máquina (*baseline-random*) y *baseline:random-baseline:first*.

Tabla 5.3 Resultados del *Test de Turing* respecto de *baseline* para inglés

Pares de resúmenes con respecto a las heurísticas *baseline*		Porcentaje de confusión entre los resúmenes seleccionados (%)
Humano	— *Baseline:first*	24
Humano	— *Baseline:random*	15
Máquina	— *Máquina*	10
Máquina	— *Baseline:first*	12
Máquina	— *Baseline:random*	12
Baseline:random	— *Baseline:first*	27

La **tabla 5.3** da un panorama de la confusión que causan las heurísticas *baseline.* En la primera fila, se muestra cómo 24% de las ocasiones el humano eligió como resumen hecho por el humano a la heurística *baseline:first*, mientras que para la elección de *baseline:random* con humano sólo 15% eligió esta combinación. Sin embargo, de manera interesante, 27% de las personas pensó que los resúmenes

del humano eran la combinación de las heurísticas. Cabe mencionar que para el lenguaje inglés estas heurísticas resultaron ser de confusión para el humano, algo que no sucedió en el leguaje español (**tabla 5.3**).

5.1 Conferencias, talleres y *corpus*

A principios del año 2000 se creó el programa de evaluación *Document Understanding Conferences* (DUC)[18] cuyo objetivo era permitir a los investigadores de la tarea de GART en inglés realizar experimentos a gran escala. Para continuar con el trabajo hecho por DUC, surgió *Text Analysis Conference* (TAC),[19] que comprende una serie de talleres de evaluación organizados para alentar la investigación en el procesamiento del lenguaje natural y sus aplicaciones relacionadas. Uno de los principales objetivos de TAC es construir colecciones de prueba que evolucionan para anticipar las necesidades de evaluación de los sistemas modernos. Fue en 2008, 2009, 2010, 2011 y 2014 cuando TAC se enfocó en la tarea de GART, siendo su principal área de estudio los resúmenes para múltiples documentos enfocados al usuario.

5.1.1 *Document Understanding Conferences* (DUC)

En el año 2000 se inició un nuevo programa de evaluación de resúmenes, inicialmente patrocinado por DARPA; se trataba de un grupo de investigadores de resúmenes, expertos que proponían un plan de trabajo para la construcción de un *corpus*, de tal forma que se fomentara el campo de la evolución de sistemas de resúmenes de texto (Baldwin *et al.*, 2000). Lo anterior fue lo que proporcionó la guía para la creación de DUC, con su primera evaluación piloto en el 2001. El plan exigía la evaluación de resúmenes de únicos y múltiples documentos a nivel específico de comprensión de texto.

DUC fue creado principalmente para una evaluación intrínseca; para DUC01 y DUC02 se utilizó también una evaluación intrínseca, mientras que para DUC03–

[18] *Document Understanding Conferences* (DUC) su objetivo es permitir a los investigadores de la tarea de automatic text summarization en el lenguaje inglés realizar experimentos a gran escala.
https://www-nlpir.nist.gov/projects/duc/index.html

[19] *Text Analysis Conference* (TAC), está organizada por una serie de talleres de evaluación creados para mejorar la evaluación de los sistemas. https://tac.nist.gov/

DUC07 se utilizó una evaluación extrínseca. Dentro de DUC, la evaluación intrínseca ha consistido en juicios directos tanto de la información lingüística bien formada como del grado en que un resumen automático expresa el mismo contenido que uno creado manualmente (a partir del mismo conjunto de documentos a resumir).

La idea principal cuando se creó DUC fue que se hiciera con textos genéricos. Los resúmenes deberían ser de diferentes tipos, estimado a un lector de periódicos adulto con un nivel considerable de educación para hacer los *gold standard*.

Inicialmente para DUC01 y DUC02 la longitud que se consideró fue de 50, 100, 200 y 400 para múltiples documentos y 100 para un solo documento, después de años de investigación se observó que la calidad de los resúmenes no dependía del tamaño. Sin embargo, se observó que para resúmenes cortos una medida de referencia sería <= 75 bytes ya que para hacer estos resúmenes generalmente no se necesita aplicar ningún tipo de método gramatical o lingüístico. Además de que, tener resúmenes de esta longitud o menos permite a los usuarios de búsquedas web poder elegir entre un gran número de ellos de manera más rápida.

Cuando se creó DUC, se buscaba que pudiera enfocarse en la creación de resúmenes extractivos y se esperaba que aquellos pasaran rápidamente de generar resúmenes extractivos a abstractivos. Sin embargo no ha sido así, sobre todo en la creación de resúmenes muy cortos.

En DUC02 se tiene resúmenes hechos por los humanos que no son completamente extractivos, lo que hace que no se pueda llegar nunca a tener un resumen completamente igual a los de los humanos cuando se aplica un método que genera automáticamente resúmenes extractivos.

Los artículos de periódicos forman parte de la amplia literatura disponible de interés para muchas personas en los diferentes países. Las noticias han formado parte de la base para la investigación en las distintas competencias, como son: TREC[20] (recuperación de información), MUC[21] (extracción de información), TDT[22] (detección y seguimiento de temas) y SUMMAC[23] (resumen).

[20]Text REtrieval Conference (TREC). Disponible para fomentar la investigación en la recuperación de información de grandes colecciones de textos. https://trec.nist.gov/

[21]Message Understanding Conferences (MUC). Su propósito es presentar la tarea de extracción de información y su utilidad para los usuarios de internet y los investigadores del procesamiento del lenguaje natural. https://www-nlpir.nist.gov/related_projects/muc/

[22]Topic Detection and Tracking (TDT). Es una iniciativa para investigar el estado del arte en la búsqueda y seguimiento de nuevos eventos en un flujo de noticias transmitidas. https://ciir.cs.umass.edu/tdt

[23]TIPSTER Text Summarization Evaluation Conference (SUMMAC). Evaluación a gran escala, independiente del desarrollador, de los sistemas automáticos de resumen de texto. https://www-nlpir.nist.gov/related_projects/tipster_summac/

En gran parte, la elección de los artículos de los periódicos en DUC se debió a su disponibilidad y al hecho de que los grupos de investigación ya habían trabajado en este género. La estructura piramidal de los artículos de periódicos significaba que los sistemas simples *baseline*, que creaban resúmenes de las primeras oraciones en un artículo o incluso un conjunto de artículos, eran difíciles de superar. A pesar del avance en las investigaciones y del pasar de los años, las pesquisas no han cambiado de dirección y siguen trabajando con este género. No obstante, actualmente se trabaja con el género de resúmenes cortos aplicados a redes sociales.

En la **tabla 5.4** se hace una descripción breve de los *corpus* proporcionados por DUC en los años 2001- 2007.

Tabla 5.4 Descripción de los *corpus* DUC

Corpus	**DUC01**	**DUC02**	**DUC03**	**DUC04**	**DUC05**	**DUC06**	**DUC07**
Carpetas	28	59	30	114	50	50	10
Archivos	309	567	624	1000	1600	1250	250
Evaluación automática				*ROUGE*	*ROUGE/BE*	*ROUGE/BE*	*ROUGE/BE*
Evaluación manual	SEE	SEE	SEE	SEE	*Pyramid*	*Pyramid*	*Pyramid*
Un documento	X	X	X	X			
Múltiples documentos	X	X	X	X	X	X	X
Tamaño un documento	100 palabras	100 palabras	10 palabras	10 palabras			
Tamaño múltiples documentos	50, 100, 200, 400	10, 50, 100, 200, 400	*Viewpoint* (100) *Question /topic* (100) *Event* (100)	*Event* (100) *Who is question* (100)	*Complex question* (250)	*Complex question* (250)	*Complex question* (250)
Enfoque del resumen para múltiples documentos							

5.1.2 *Text Analysis Conference* (TAC)

Text Analysis Conference (TAC) conforma una serie de talleres de evaluación organizados para alentar la investigación en el procesamiento del lenguaje natural y aplicaciones relacionadas, contiene una gran colección de pruebas, es decir, procedimientos de evaluación comunes. TAC posee conjuntos de tareas conocidas como "pistas", cada una de las cuales se centra en un subproblema particular de procesamiento del lenguaje natural. Las pistas de TAC se aplican en las tareas del usuario final, pero también incluyen evaluaciones de componentes situadas en el contexto de dichas tareas.

Los objetivos de TAC son:

- Promover la investigación en PNL basada en grandes colecciones de pruebas comunes.
- Mejorar las metodologías y medidas de evaluación para PNL.
- Construir una serie de colecciones de prueba que evolucionan para anticipar las necesidades de evaluación de los sistemas modernos de PNL.
- Aumentar la comunicación entre la industria, el mundo académico y el gobierno mediante la creación de un foro abierto para el intercambio de ideas de investigación.
- Para acelerar la transferencia de tecnología de los laboratorios de investigación a productos comerciales, demostrando mejoras sustanciales en las metodologías de PNL sobre problemas del mundo real.

TAC surgió de DUC NIST para la síntesis de texto en la tarea de respuestas a preguntas de la conferencia TREC. En la **tabla 5.5** se muestra una descripción de los *corpus* TAC2008 – TAC 2011 y TAC2014.

5.1.3 *Corpus* utilizados para la evaluación y la comparación

DUC01[24] es un *corpus* de noticias en inglés sobre desastres naturales, el cual fue diseñado para la generación de resúmenes de múltiples y de un solo documento.

[24]Para el acceso a los datos del ***corpus*** DUC01 dirigirse a https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html

Tabla 5.5 Descripción de los *corpus* TAC

Corpus	2008	2009	2010	2011	2014
Múltiples documentos	Resumen de actualización	Resúmenes A de información y B de actualización	Análisis semántico	Resúmenes A de información y B de actualización	Resumen basado en citas
Lenguajes	Inglés	Inglés	Inglés	Inglés	Inglés
Tamaño	100 palabras	100 palabras	100 palabras	100 palabras	250 palabras
Dominio	Noticias	Noticias	Noticias	Noticias	Artículos médicos
Composición	48 temas con 20 documentos relevantes divididos en dos conjuntos A y B. Con 4 resúmenes de referencia para cada conjunto de artículos.	44 temas, cada tema con 1 título y la narrativa con 20 documentos relevantes divididos en dos conjuntos A y B.	46 temas, con 20 documentos para cada tema.	46 grupos de documentos. Cada grupo de documentos consta de 10 artículos de noticias y se clasifica en un solo tema.	20 documentos biomédicos. Cada uno tiene un conjunto de 10 documentos de citas con referencias del documento original.

DUC01 está compuesto por treinta conjuntos de referencia y treinta de prueba, estos últimos comprenden trescientos nueve documentos. Cada conjunto contiene los documentos originales, así como los resúmenes para uno solo y múltiples documentos generados manualmente. Este *corpus* está etiquetado, lo cual permite tener una separación clara de las oraciones y, por ende, un mejor manejo de la información con la que está constituido. Además, cuenta con los resultados de diferentes medidas de *baseline*. Para este *corpus* se ha calculado el desempeño de las herramientas comerciales y el *Topline.*

Para este libro se considera el *corpus* DUC01 para la GART de un solo documento, por lo que el número de palabas requeridas para los resúmenes es de cien.

DUC02[25] es un *corpus* de noticias en inglés sobre diferentes temas de tecnología, alimentación, política, finanzas, entre otros. Fue diseñado para la GART en función de dos tareas: múltiples y un solo documento. Está compuesta por quinientos sesenta y siete documentos. Para cada uno de ellos se crearon dos resúmenes con una longitud mínima de 100 palabras; dos humanos expertos los elaboraron. Además, cuenta con los resultados de diferentes medidas de *baseline*. DUC02 es uno de los más utilizados por los investigadores en el área de GART. Además está etiquetado, lo que permite tener una separación clara de las oraciones. Para este *corpus* se ha calculado el desempeño de las herramientas comerciales y el *Topline*.

Para este libro se considera el *corpus* DUC02 para la GART de un solo documento, por lo que el número de palabas requeridas para los resúmenes es de cien.

5.2 Heurísticas

Para realizar el cálculo y los experimentos para las heurísticas, las herramientas comerciales y los métodos científicos novedosos, se utilizan los *corpus* DUC02 y DUC01, debido a que son dos de las colecciones más utilizadas para la tarea de GART para un solo documento en el lenguaje inglés.

[25]Para el acceso a los datos del *corpus* DUC02:
https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

5.2.1 *Baseline:random*

Ledeneva (2008) propuso calcular la heurística *baseline:random* con una evaluación promedio al seleccionar diez veces de manera aleatoria las oraciones para un documento. Siguiendo esta metodología, García (2008) reporta un *f-measure* de 0.38981 como medida de la heurística *baseline:random* para el *corpus* DUC02 evaluado con *ROUGE*. A partir de la misma metodología de Ledeneva (2008) para calcular el *baseline:random* en DUC01, Alvarado (2017) reporta un valor de 0.36587.

5.2.2 *Baseline:first*

En García (2008) se reporta el valor de *baseline:first* para la colección DUC02 evaluada con *ROUGE* como: 0.4729. Alvarado (2017) reporta el valor de 0.44862 para *baseline:first* en la colección DUC01. Esta heurística sirve como referencia a los métodos y herramientas que trabajan con el dominio de noticias, pero no se tiene la evidencia de que se puedan obtener buenos resultados con algún otro dominio. Cabe mencionar que para el lenguaje inglés esta heurística fue superada hace diez años, es decir, cincuenta años después de que la investigación en el área de GART diera inicio.

5.2.3 *Topline*

Para calcular el *Topline* del *corpus* DUC01, Rojas (2018) utilizó trescientos ocho documentos individuales para generar diferentes combinaciones de oraciones mediante un algoritmo genético. Los parámetros utilizados en dicho algoritmo para calcular el *Topline* de la colección DUC01 se presentan en la **tabla 5.6**.

Tabla 5.6 Parámetros del algoritmo genético para Topline en DUC01

Experimento	Elite	Generaciones	Individuos	Selección		Cruza	Mutación	
				Tipo	P		Tipo	P
1	Si	20	120	Torneo	3	CX	Inserción	8

Los resultados obtenidos de la heurística *Topline* para el *corpus* DUC01 en la evaluación, utilizando la herramienta *ROUGE*, se presentan en la **tabla 5.7**.

Tabla 5.7 Resultados de Topline para DUC01

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.59796	0.59046	0.59408
ROUGE-2	0.33622	0.33234	0.33422
ROUGE-SU4	0.34619	0.34202	0.34404

Para el cálculo de la heurística *Topline* para el *corpus* DUC02, Rojas (2018) utilizó quinientos sesenta y siete documentos individuales al generar los resúmenes, considerando combinaciones de oraciones mediante un algoritmo genético. Los parámetros utilizados en dicho algoritmo para calcular el *Topline* de la colección DUC02 se presentan en la **tabla 5.8**.

Tabla 5.8 Parámetros del algoritmo genético para Topline en DUC02

Experimento	Elite	Generaciones	Individuos	Selección		Cruza	Mutación	
				Tipo	P		Tipo	P
1	Si	30	150	Torneo	3	CX	Inserción	8

Los resultados obtenidos de la heurística *Topline* para el *corpus* DUC02 en la evaluación, utilizando la herramienta *ROUGE*, se presentan en la **tabla 5.9**.

Tabla 5.9 Parámetros del algoritmo genético para Topline en DUC02

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.62601	0.62164	0.62367
ROUGE-2	0.35877	0.35624	0.35742
ROUGE-SU4	0.36107	0.35851	0.35970

5.3 Herramientas comerciales

Para el lenguaje inglés se presentan evaluaciones con dos de los principales *corpus* DUC01 y DUC02 evaluados con la herramienta *ROUGE*. A continuación, en la **tabla 5.10** se presentan las herramientas comerciales probadas en cada *corpus*.

Tabla 5.10 Herramientas evaluadas con los *corpus* DUC01 y DUC02

Herramienta	Tipo	Inglés	
		DUC01	DUC02
Copernic Summarizer	Instalable	✓	✓
Microsoft Office Word 2003/2007	Instalable	✓	✓
SweSum	Línea	✓	
T-Conspectus	Línea	✓	
OTS	Línea	✓	✓
Text Compactor	Línea	✓	
Article *Summarizing* Online	Línea	✓	
Summarizer	Línea	✓	
Tools4noobs	Línea		✓
Pertinence Summarizer	Línea		✓
Shvoong	Línea		✓

Cabe mencionar que todas las herramientas de la **tabla 5.10** pueden ser aplicadas para los *corpus* DUC01 y DUC02, ya que trabajan para el lenguaje inglés. Sin embargo, no se presentan los resultados para las dos colecciones debido a que la generación de los resúmenes en esta herramienta se hace documento por documento, por lo que el tiempo para poder realizarlos es considerable; además de que algunas de ellas ya no se encuentran disponibles.

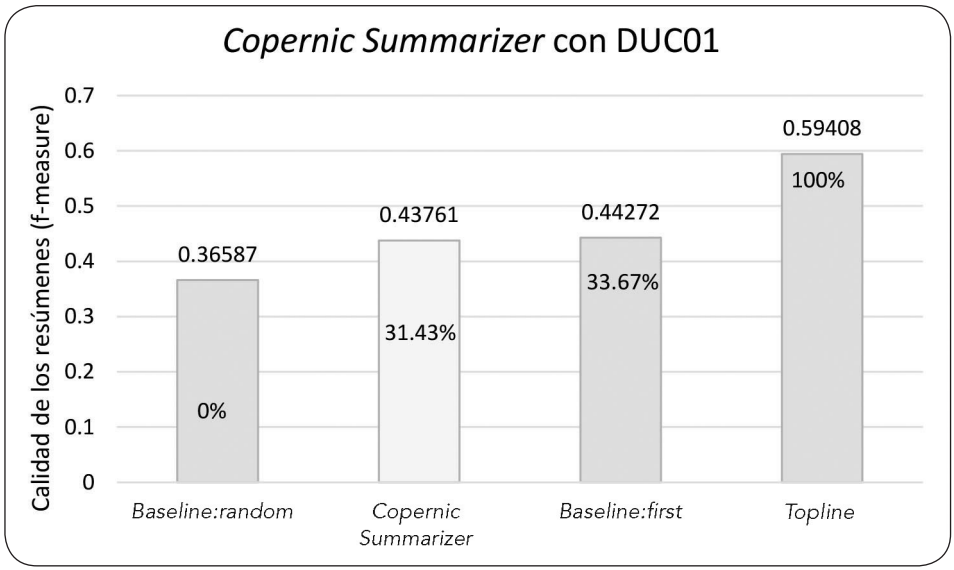
Hay herramientas que no tienen la opción de generar resúmenes a cien palabras, por lo que se aplicó la fórmula 6 para calcular el porcentaje que permite tener más de cien para cada documento.

$$\frac{\textit{número de palabras deseadas}}{\textit{número de palabras totales en el documento}} * 100 \qquad (6)$$

5.3.1 *Copernic Summarizer*

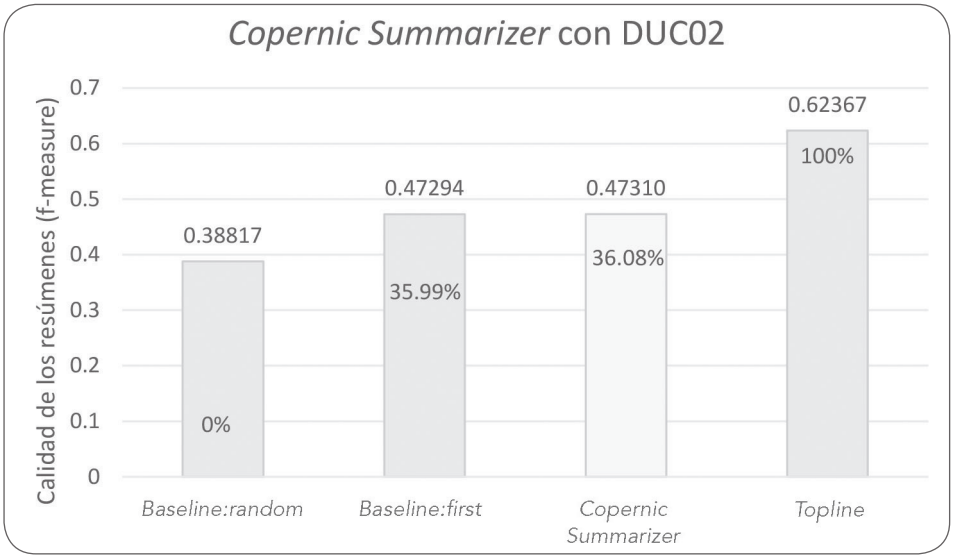
Para la herramienta *Copernic Summarizer* se evaluaron las colecciones DUC01 y DUC02; tiene la opción de generar resúmenes a cien palabras (longitud requerida para los dos *corpus*), por lo que se seleccionó esta opción. A continuación, se muestran los resultados obtenidos por esta herramienta en las colecciones DUC01 y DUC02.

Gráfica 5.1 Resultados de *Copernic Summarizer* usando DUC01 en comparación con las diferentes heurísticas



En las **gráficas 5.1** y **5.2** se muestran los resultados obtenidos con la herramienta *Copernic Summarizer* para los *corpus* DUC01 y DUC02, respectivamente. Como se puede observar, esta herramienta no supera a *baseline:first* en la colección DUC01. Sin embargo, para DUC02 sí se supera esta heurística. Considerando que *baseline:random* es la peor forma de hacer un resumen, se le asigna un valor de cero; mientras que a la heurística *Topline* se le asigna uno máximo de cien, ya que se considera que es la mejor forma. Entonces, tomando como referencia a *baseline:random* y a *Topline*, *Copernic Summarizer* obtiene 31.43% de avance en la tarea de GART para el *corpus* DUC01, mientras que para DUC02 obtiene 36.08%.

Gráfica 5.2 Resultados de *Copernic Summarizer* usando DUC02 en comparación con las diferentes heurísticas



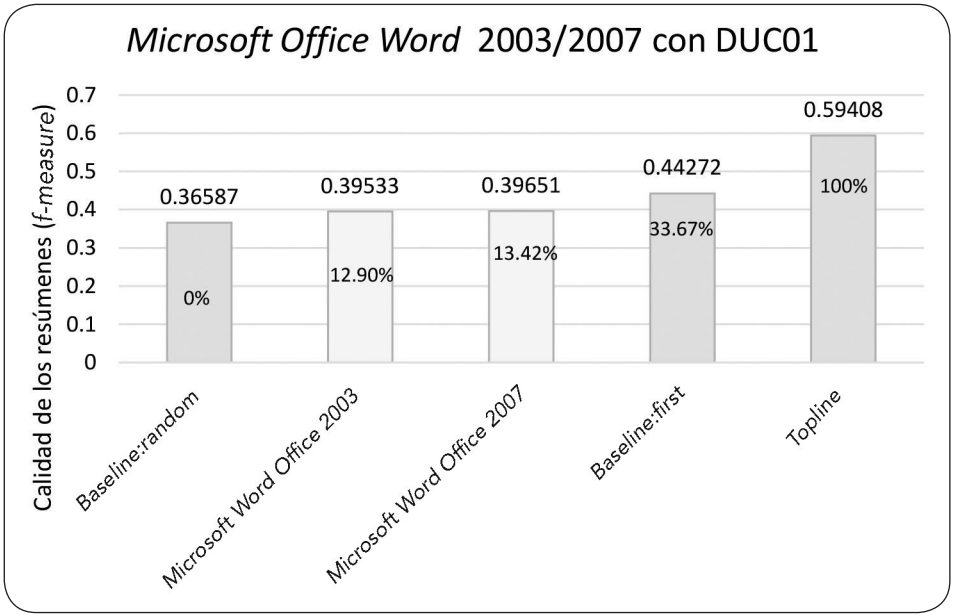
5.3.2 *Microsoft Office Word*

Para la herramienta *Microsoft Office Word* se evaluaron las colecciones DUC01 y DUC02; tiene disponible la opción de generar resumen en las versiones 2003 y 2007. Para ello, es posible hacer resúmenes a cien palabras, una opción requerida por los *corpus* DUC01 y DUC02. Sin embargo, a pesar de que la herramienta cuenta con esta opción los resúmenes hechos por *Microsoft Office Word* son de menor longitud, por lo que se hizo uso de la fórmula 6 para el cálculo del porcentaje que se debía utilizar. A continuación, se muestran los resultados obtenidos por *Microsoft Office Word* en las colecciones DUC01 y DUC02.

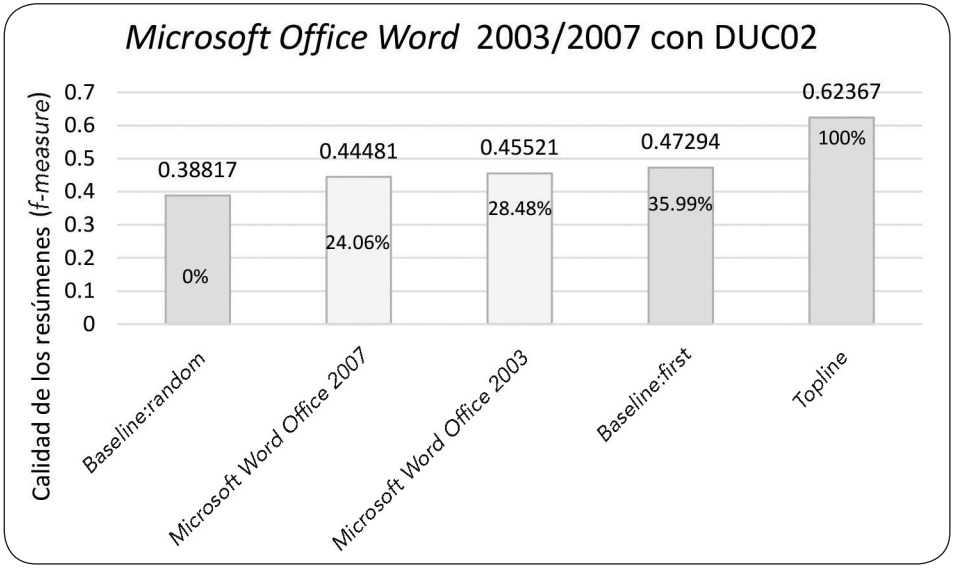
En la **gráfica 5.3** se muestran los resultados para la herramienta *Microsoft Office Word* en sus versiones 2003 y 2007, con el *corpus* DUC01. Como se puede observar, para la colección DUC01 los resultados obtenidos con esta herramienta no superan la heurística *baseline:first*. Sin embargo, se obtienen mejores resultados que con *baseline:random*.

En la **gráfica 5.4** se muestran los resultados para la herramienta *Microsoft Office Word* en sus versiones 2003 y 2007, con el *corpus* DUC02. Para esta

Gráfica 5.3 Resultados de *Microsoft Office Word* 2003/2007 usando DUC01 en comparación con las diferentes heurísticas



Gráfica 5.4 Resultados de *Microsoft Office Word* 2003/2007 usando DUC02 en comparación con las diferentes heurísticas

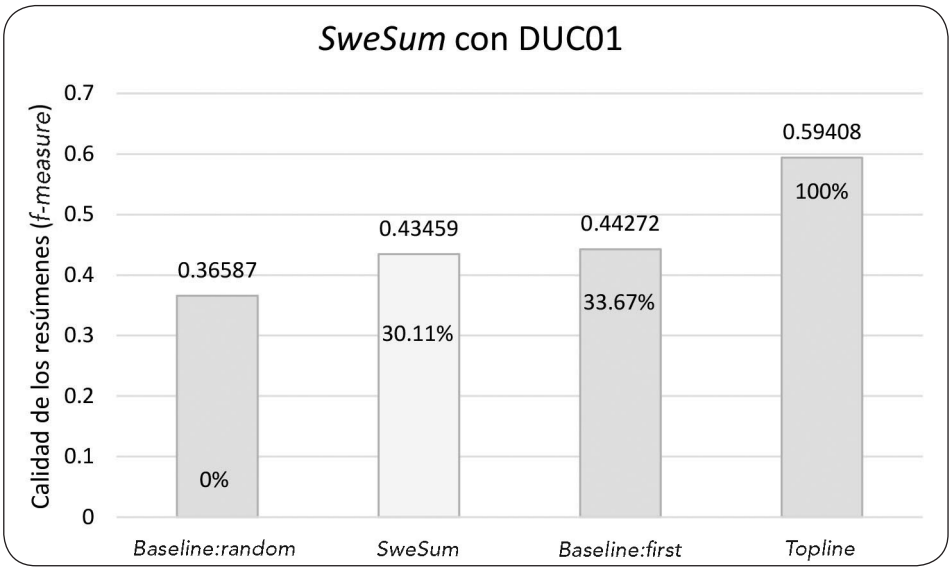


colección de documentos, al igual que para DUC01, la herramienta no supera a la heurística *baseline:first.* Sin embargo, no se esperaba que la versión de *Microsoft Office Word* 2003 obtuviera mejores resultados que la más actual, *Microsoft Office Word* 2007.

5.3.3 *SweSum*

Para la herramienta *SweSum* se evaluó la colección DUC01; es preciso calcular el porcentaje correspondiente a cada documento con la fórmula 6, de tal modo que para cada uno de ellos se tengan cien palabras, como lo indican las especificaciones propias de la colección. A continuación se muestran los resultados obtenidos por esta herramienta en DUC01, evaluados con *ROUGE*.

Gráfica 5.5 Resultados de *SweSum* usando DUC01 en comparación con las diferentes heurísticas

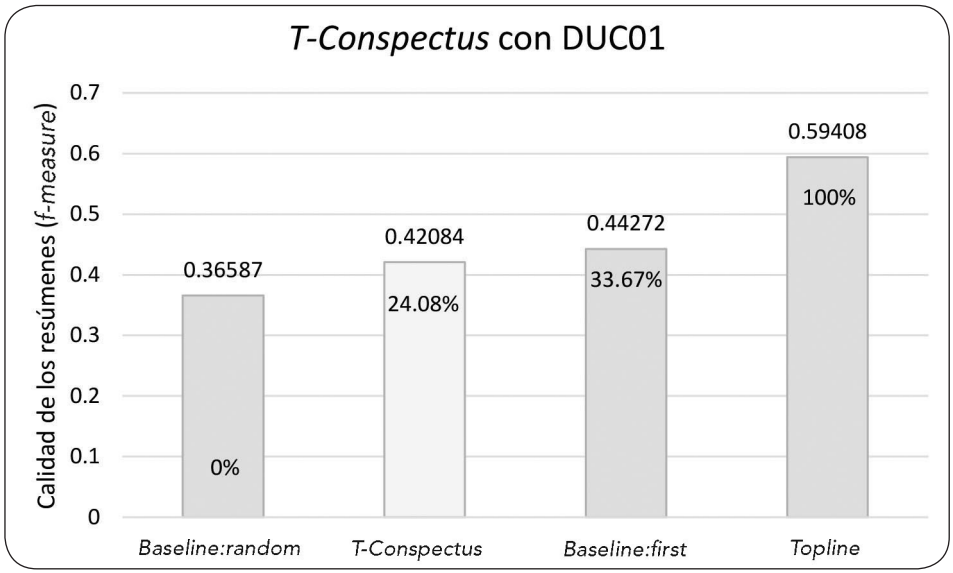


En la **gráfica 5.5** se muestran los resultados para la herramienta *SweSum* con el *corpus* DUC01; se puede observar que no superan la heurística *baseline:first*. Sin embargo, la herramienta presenta 30.11% de avance en la tarea de GART con respecto a *baseline:random* y *Topline*.

5.3.4 *T-Conspectus*

Para la herramienta *T-Conspectus* se evaluó la colección DUC01. Es preciso calcular el porcentaje correspondiente a cada documento con la fórmula 6, de modo que para cada uno de ellos se tengan cien palabras, como lo indican las especificaciones propias de la colección. A continuación se muestran los resultados obtenidos por esta herramienta en DUC01, evaluados con *ROUGE*.

Gráfica 5.6 Resultados de *T-Conspectus* usando DUC01 en comparación con las diferentes heurísticas



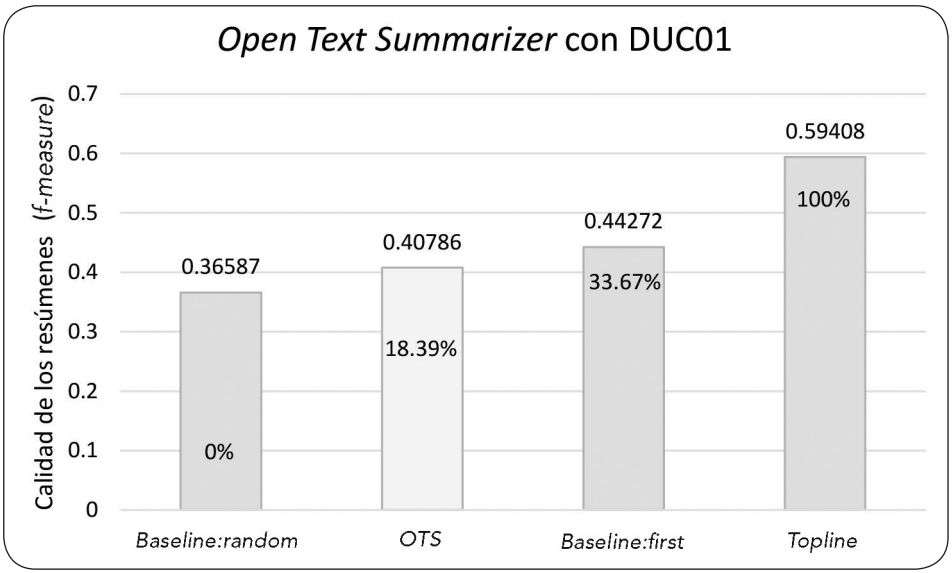
En la **gráfica 5.6** se muestran los resultados para la herramienta *T-Conspectus* con el *corpus* DUC01. Como se puede observar, los resultados obtenidos no superan la heurística *baseline:first*. Sin embargo, se tiene 24.08% de avance respecto de *baseline:random* y *Topline*.

5.3.5 *Open Text Summarizer (OTS)*

Para la herramienta *Open Text Summarizer (OTS)* se evaluaron las colecciones DUC01 y DUC02. Es preciso calcular el porcentaje correspondiente a cada documento usando la fórmula 6, de tal manera que para cada uno de ellos se tengan

cien palabras como lo indican las especificaciones propias de las colecciones. A continuación se muestran los resultados obtenidos por esta herramienta en DUC01, evaluados con *ROUGE*.

Gráfica 5.7 Resultados de *Open Text Summarizer* usando DUC01 en comparación con las diferentes heurísticas



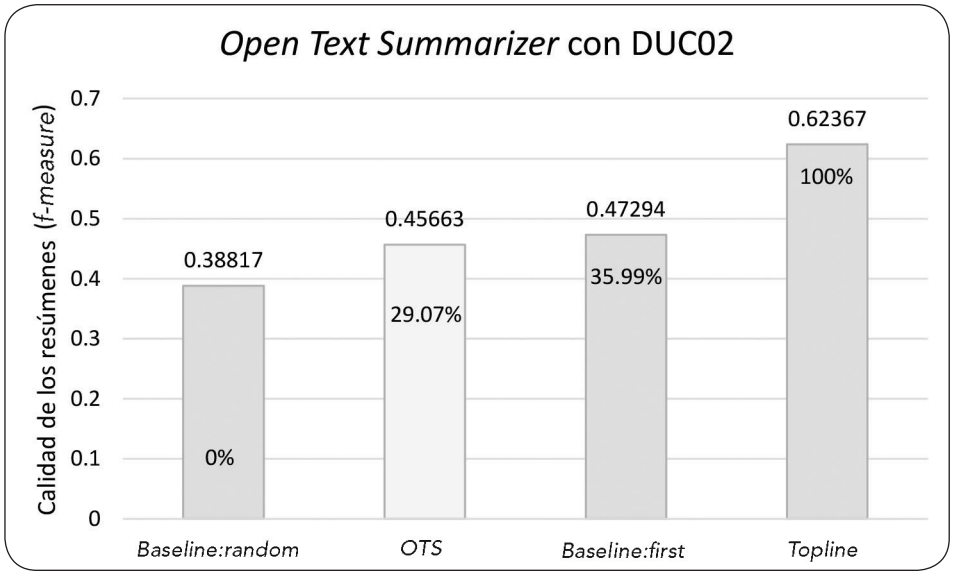
En la **gráfica 5.7** se muestran los resultados obtenidos con *Open Text Summarizer* usando el *corpus* DUC01. Como se puede observar, esta herramienta no supera a *baseline:first*. Si se considera a *baseline:random* como la peor forma de hacer un resumen y a *Topline* como la mejor, entonces *Open Text Summarizer* obtiene 18.39%; hasta ahora es el porcentaje más bajo.

Para el *corpus* DUC02 la herramienta OTS obtiene 29.07% de avance respecto de las heurísticas *baseline:random* y *Topline* (**gráfica 5.8**).

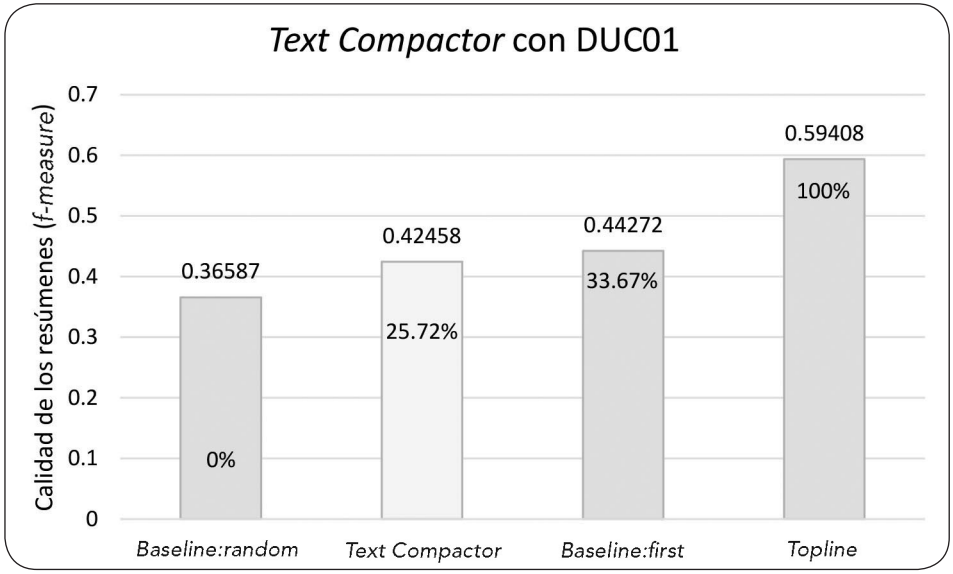
5.3.6 *Text Compactor*

Para *Text Compactor* se evaluó la colección DUC01. Es preciso calcular el porcentaje correspondiente a cada documento con la fórmula 6, de tal manera que para cada resumen de los documentos de la colección se tengan cien palabras.

Gráfica 5.8 Resultados de *Open Text Summarizer* usando DUC02 en comparación con las diferentes heurísticas



Gráfica 5.9 Resultados de *Text Compactor* usando DUC01 en comparación con las diferentes heurísticas

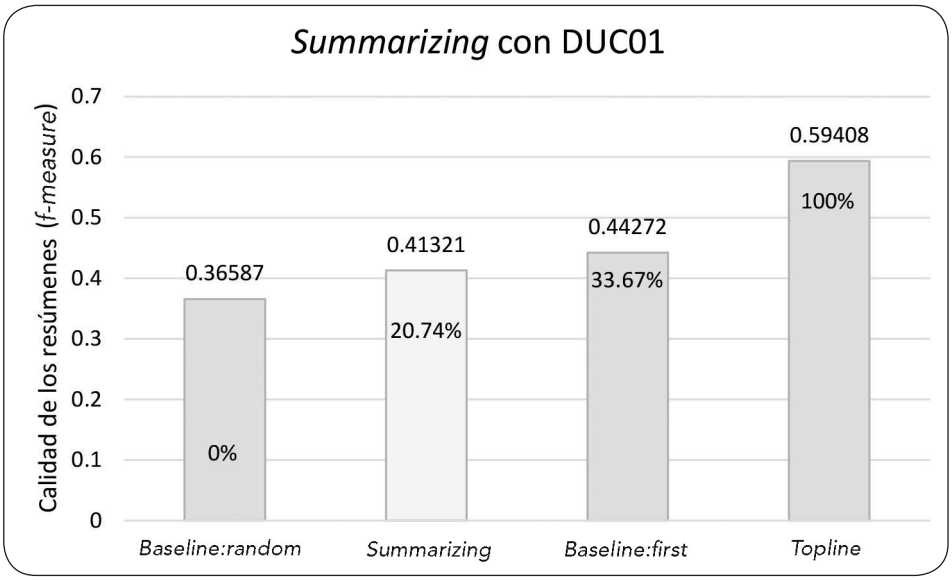


Los resultados obtenidos por esta herramienta muestran que T-Compactor tiene 25.72% de avance respecto de *baseline:random* y *Topline* para el *corpus* DUC01 evaluado con *ROUGE*.

5.3.7 *Summarizing*

Para *Summarizing* se evaluó la colección DUC01. La herramienta Article *Summarizing* Online tiene la opción de generar resúmenes a cien palabras (longitud requerida por el *corpus*). A continuación se muestran los resultados obtenidos por esta herramienta en DUC01 evaluada con *ROUGE*.

Gráfica 5.10 Resultados de *Summarizing* usando DUC01 en comparación con las diferentes heurísticas

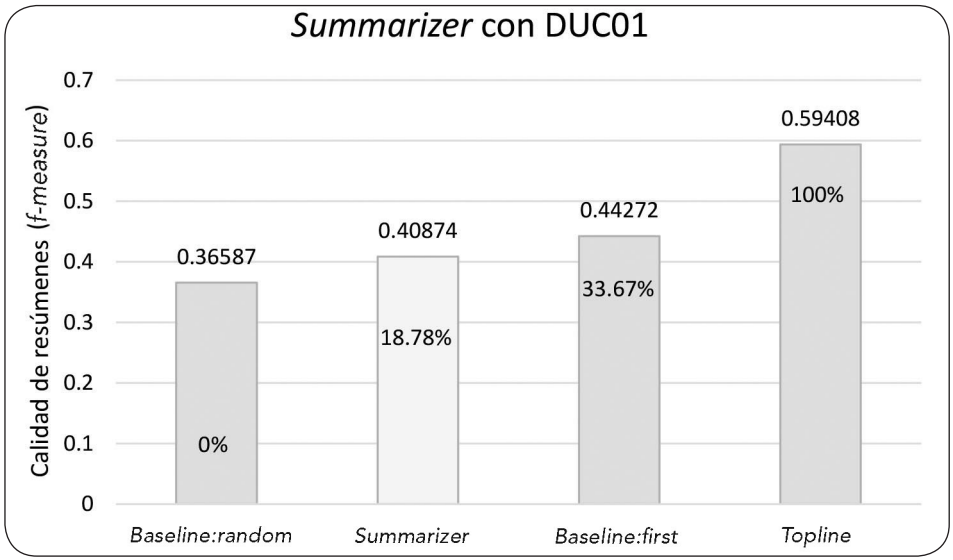


En la **gráfica 5.10** se muestran los resultados para la herramienta *Summarizing* con el *corpus* DUC01. Como se puede observar, los resultados obtenidos no superan la heurística *baseline:first*. Si se considera a *baseline:random* como la base para saber si un resumen es bueno (0%) y a *Topline* como el máximo que se puede obtener (100%), entonces *Summarizing* obtiene 20.74%.

5.3.8 *Summarizer*

Para *Summarizer* se evaluó la colección DUC01. Para esta herramienta se tiene que calcular el porcentaje correspondiente a cada documento, de tal forma que para cada uno de ellos se tengan cien palabras. El avance con esta herramienta es de 33.67% (ver **gráfica 5.11**) respecto de la peor heurística (*baseline:random*) y de la mejor (*Topline*).

Gráfica 5.11 Resultados de *Summarizer* usando DUC01 en comparación con las diferentes heurísticas

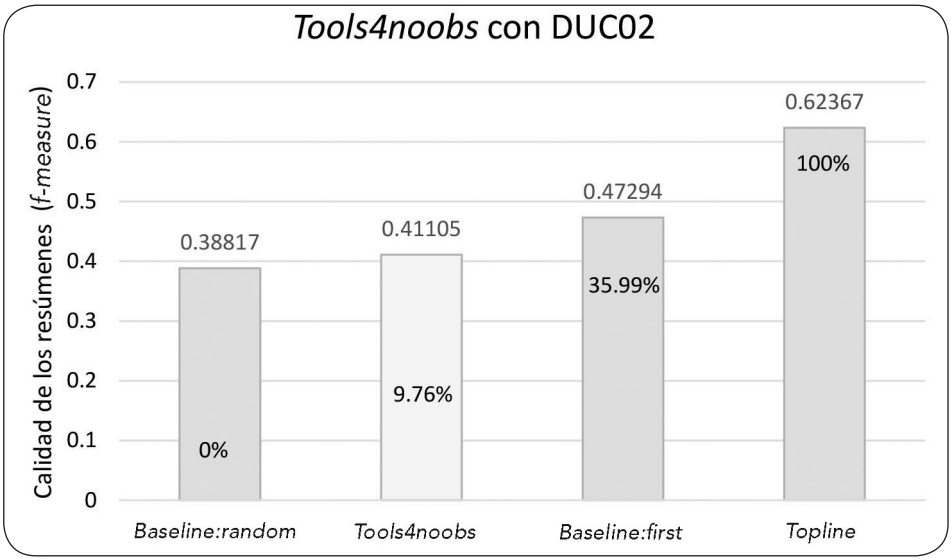


5.3.9 *Tools4noobs*

Para *Tools4noobs* se evaluó la colección DUC02. Es necesario calcular el porcentaje correspondiente a cada documento, de tal manera que para cada uno de ellos se tengan cien palabras; el cálculo se hace con la fórmula 6. A continuación se muestran los resultados obtenidos por esta herramienta en la colección DUC02, evaluados con *ROUGE*.

En la **gráfica 5.12** se muestran los resultados obtenidos con *Tools4noobs* usando el *corpus* DUC02. El nivel de avance que se tiene para esta herramienta

Gráfica 5.12 Resultados de *Tools4noobs* con DUC02 en comparación con las diferentes heurísticas



es el más bajo respecto de las probadas para el lenguaje inglés, ya que obtiene solamente 10% en comparación con las heurísticas *baseline:random* y *Topline*.

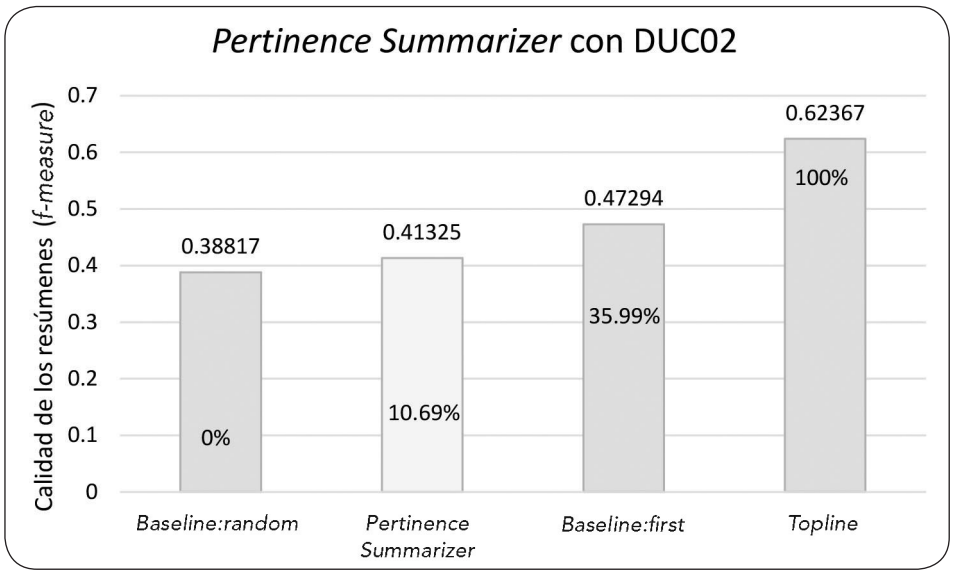
5.3.10 *Pertinence Summarizer*

Para *Pertinence Summarizer* se probó la colección DUC02. Para esta herramienta se tiene que calcular el porcentaje para que todos los resúmenes contengan cien palabras con el uso de la fórmula 6. El nivel de avance que presenta es de 10.69%, lo que la coloca en el segundo lugar más bajo de las herramientas comerciales.

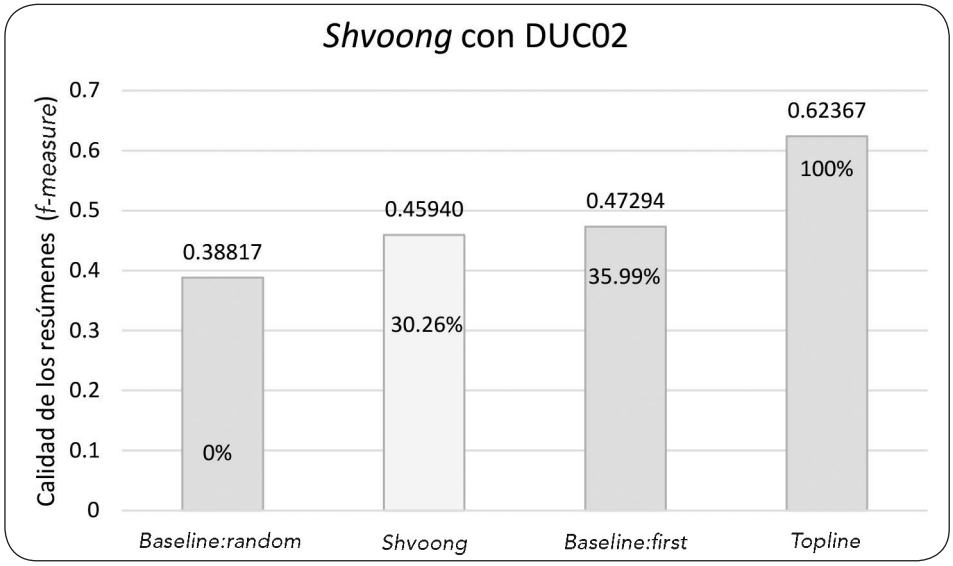
5.3.11 *Shvoong*

Para *Shvoong* se evaluó la colección DUC02. Se calcula el porcentaje correspondiente para cada documento, de tal modo que para cada uno de ellos se tengan cien palabras; se hace uso de la fórmula 6. A continuación se muestran los resultados obtenidos por esta herramienta en DUC02, evaluados con *ROUGE*.

Gráfica 5.13 Resultados de *Pertinence Summarizer* usando DUC02 en comparación con las diferentes heurísticas



Gráfica 5.14 Resultados de *Shvoong* usando DUC02 en comparación con las diferentes heurísticas



En la **gráfica 5.14** se muestran los resultados obtenidos con *Shvoong* usando el *corpus* DUC02. Como se puede observar, esta herramienta no supera a *baseline:first*. Sin embargo, tiene 30.26% de avance respecto de las heurísticas *baseline:random* y *Topline*.

5.4 Métodos científicos novedosos

En esta sección se presentan los métodos científicos novedosos que se han probado en el lenguaje inglés. En la **tabla 5.11** se listan los diferentes métodos que se explicarán en esta sección, y se menciona el *corpus* que utilizan.

Tabla 5.11 Métodos científicos novedosos evaluados con los *corpus* DUC01 y DUC02

Método	DUC01	DUC02
AG-4feature	✓	✓
Ma-SingleDocSum	✓	✓
UnifiedRank		✓
AG-Bag-words		✓
AG-Bigramas		✓
AG-Multi		✓
TextRank		✓
SFMs k-best		✓
SFMs (1 *best+first*)		✓
Agrupamiento con *SFMs*		✓

5.4.1 *Ma-SingleDocSum*

El método *Ma-SingleDocSum* propuesto por Mendoza (Mendoza *et al*., 2014) está basado en un algoritmo memético, enfocado en la generación de resúmenes para un solo documento. Además de utilizar operadores genéticos para hacer resúmenes, utiliza la búsqueda local. Los parámetros que considera para la

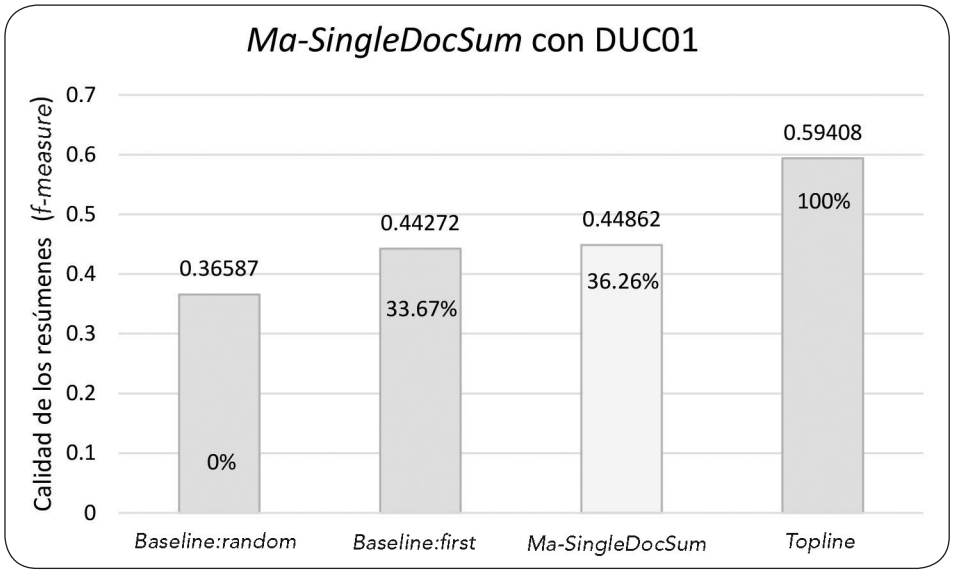
función de aptitud son: posición de las oraciones, relación de la oración con el título, longitud de la misma, cohesión y la convergencia (conocida como temática del texto).

Se sabe que, según la tabla 3.1 mostrada en la sección 3.2.1, la posición de las oraciones, la relación de la oración con el título y su longitud son de las características más utilizadas para la tarea de GART.

A continuación se muestra una comparativa del método *Ma-SingleDocSum* con las heurísticas para las colecciones DUC01 y DUC02 evaluadas con la herramienta *ROUGE*.

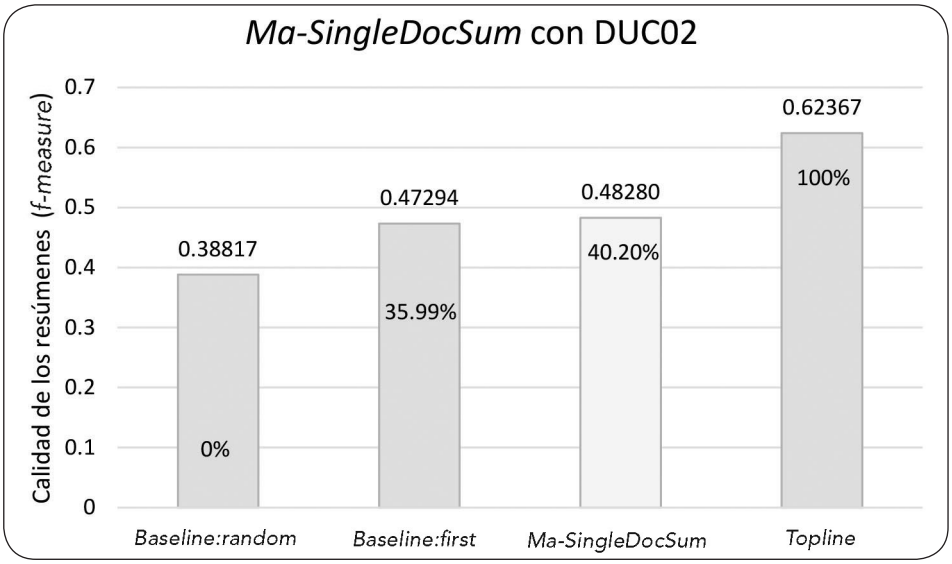
Para la colección DUC01, el método supera las dos heurísticas de referencia (**gráfica 5.15**), al igual que para DUC02 (**gráfica 5.16**).

Gráfica 5.15 Resultados del método *Ma-SingleDocSum* usando DUC01 en comparación con las diferentes heurísticas



Cabe mencionar que este es uno de los mejores métodos para la GART, el cual por sus características puede ser aplicado a lenguajes como el inglés y el español.

Gráfica 5.16 Resultados del método *Ma-SingleDocSum* usando DUC02 en comparación con las diferentes heurísticas



5.4.2 *UnifiedRank*

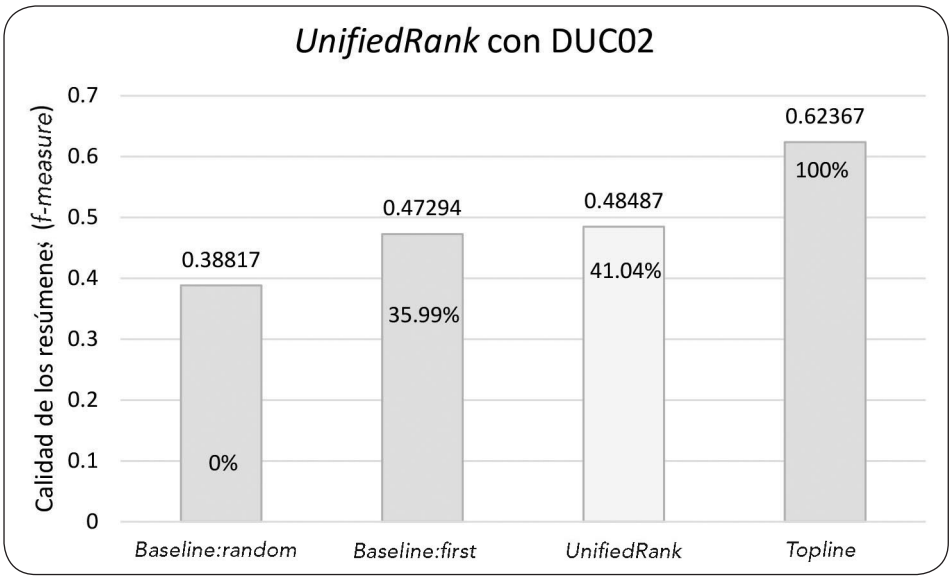
El método *UnifiedRank* (Wan, 2010) se basa en grafos enfocados a generar resúmenes para un solo documento y para multidocumentos. Este trabajo examina la influencia que existe entre la generación de resúmenes para los mismos. El *corpus* con el que opera para un solo documento es DUC02 y la herramienta con la que evalúa es *ROUGE*.

Hasta el momento, el método *UnifiedRan*k es el que mejores resultados presenta para el *corpus* DUC02 (**gráfica 5.17**).

5.4.3 *AG-Bag-words*

El método propuesto por García-Hernández y Ledeneva (2013) es uno de los que han obtenido los mejores resultados. Está realizado con un algoritmo genético y utiliza el modelo de texto "bolsa de palabras". La función de aptitud usada en este trabajo toma dos características principales:

Gráfica 5.17 Resultados de *UnifiedRank* usando DUC02 en comparación con las diferentes heurísticas



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

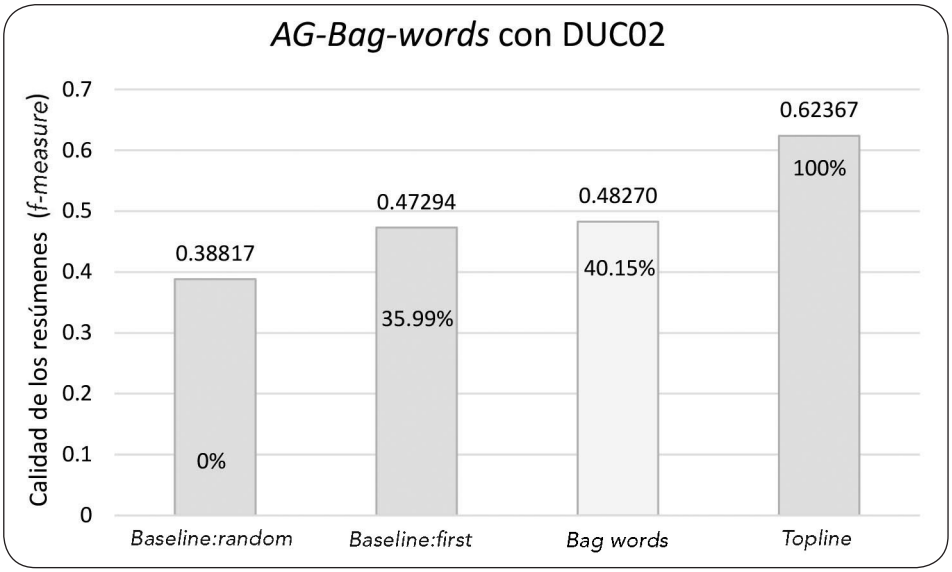
En la **gráfica 5.18** se muestra la comparación de método con las diferentes heurísticas; se puede observar que Bag words supera a la heurística *baseline:first.*

5.4.4 *AG-Bigramas*

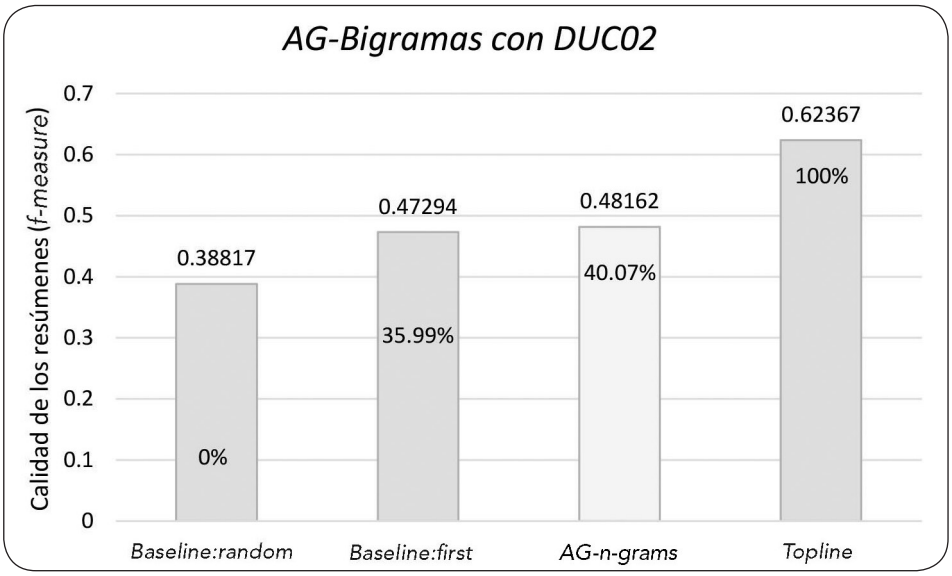
El método *AG-Bigramas* fue propuesto por Matias (2013). Está basado en un algoritmo genético, aplicando específicamente el modelo de texto bigramas. Se emplea para un solo documento en el lenguaje inglés. En la función de aptitud, se considera la posición de las oraciones y la frecuencia de los términos. Matias (2013) señala que los bigramas permiten tener mejor manejo de la información y menos pérdida.

El método *AG-Bigramas* es uno de los que han presentado mejores resultados, superando a las dos heurísticas: *baseline:random* y principalmente a *baseline:first* (**gráfica 5.19**).

Gráfica 5.18 Resultados de *Bag words* usando DUC02 en comparación con las diferentes heurísticas



Gráfica 5.19 Resultados de *AG-Bigramas* usando DUC02 en comparación con las diferentes heurísticas



5.4.5 *AG-Multi*

El método *AG-Multi* propuesto por Matias (2016) es uno de los más utilizados para la GART para diferentes lenguajes; está basado en un algoritmo genético y utiliza como modelo de texto n-gramas. Para la función de aptitud, considera dos de las características más utilizadas en el estado del arte (García-Hernández *and* Ledeneva, 2013), las cuales son: frecuencia de los términos y posición de las oraciones.

Frecuencia de los términos

Para la generación de un resumen (S), el límite máximo de palabras (m) debe ser considerado. En consecuencia, el número de unidades de recuperación siempre está limitado por el umbral de máximo de palabras. Por lo tanto, el resumen de referencia (el hecho por el humano) debe tener, por un lado, las palabras más relevantes del texto original (T) y, por el otro, expresividad; es decir, no debe ser redundante.

La relevancia de w está representada por la frecuencia de aparición de la palabra en el texto original ($frequency(w,T)$, y la expresividad es representada si sólo se consideran diferentes palabras que el resumen puede tener $(\{word \in S\})$. En este sentido, el mejor resumen contendría las palabras más frecuentes con respecto al texto original, y cada una deberá ser diferente. Para tener una medida normalizada, García-Hernández (2013) propone que la suma de las frecuencias de las diferentes palabras en el resumen se debe dividir por la suma de las frecuencias de las palabras más frecuentes en el texto original.

$$\beta = \frac{\sum_{p=\{word \in S\}}^{m} frecuency(p,T)}{\sum_{q=\{word \in T\}}^{m} frecuency(q,T)} \tag{7}$$

Posición de las oraciones

Por lo general, la tarea de GART considera importante esta característica, como lo muestra la tabla 3.1; de dieciséis trabajos analizados, catorce la utilizan para resolver la tarea. Se parte de la heurística que ha demostrado que las primeras oraciones son muy importantes para la GART. Una de las ideas para darle más

importancia a las primeras oraciones sería considerar la primera con importancia *xn*. La segunda con importancia *xn* − 1 hasta la última, que tendría importancia de 1; pero esto pudiera ser muy drástico porque de un texto de treinta oraciones, se diría que la primera es treinta veces más importante que la última, pero pudiera corresponder a las conclusiones y ésta no tendría posibilidad de aparecer en el resumen.

En el trabajo de García-Hernández y Ledeneva se propone suavizar la importancia de las oraciones. Para ello, es posible utilizar la ecuación general de una recta con pendiente *m*. La pendiente indica la importancia que se le da a las primeras oraciones o a las últimas; si es negativa, las primeras oraciones tienen más importancia, (−1) significa que baja hacia la derecha en ángulo de 45 grados, 0 quiere decir que todas las oraciones tienen la misma importancia y positiva que las últimas oraciones tienen más importancia; (1) significa que sube hacia la derecha en ángulo de 45 grados.

Para un texto con *n* oraciones, si la oración es seleccionada para el resumen (este es el cromosoma $|c_i| = 1$) entonces su relevancia se define como m($i - x$) + x, donde $x = 1 + (n - 1/2$ y t es la pendiente por descubrir. Con el fin de normalizar la medida de la posición de la oración (δ), se calcula la importancia de las primeras *k* oraciones, donde *k* es el número de oraciones elegidas.

Entonces, la fórmula para calcular la importancia de las primeras oraciones quedaría de la siguiente manera:

$$\delta = \frac{\sum_{|C_i|=1}^{n} m(i-x)+\mathrm{x}}{\sum_{j=1}^{k} m(j-x)+1}, x = 1+\frac{(n-1)}{2}. \tag{8}$$

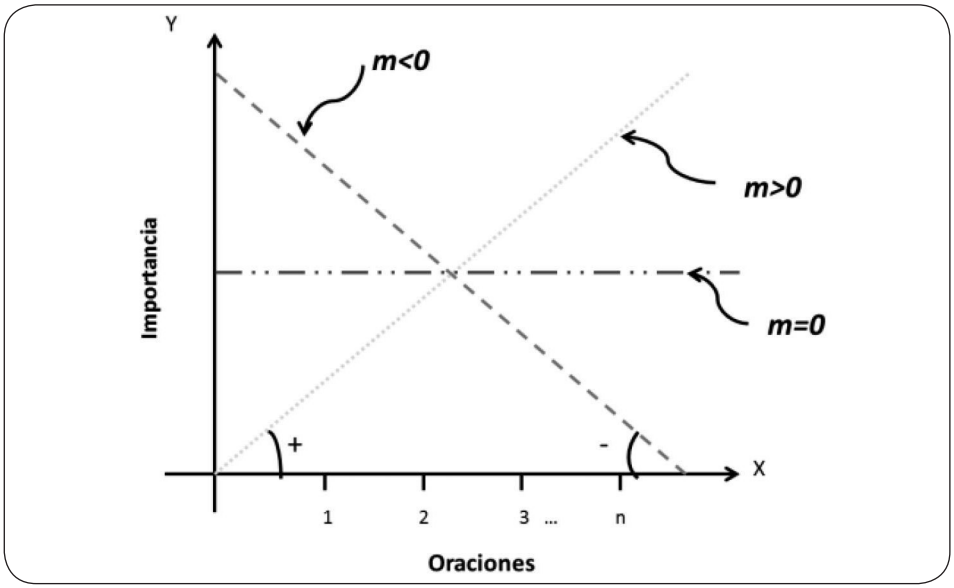
Finalmente, para obtener el valor de la función de aptitud se aplica la siguiente fórmula:

$$fitness = \beta * \delta \tag{9}$$

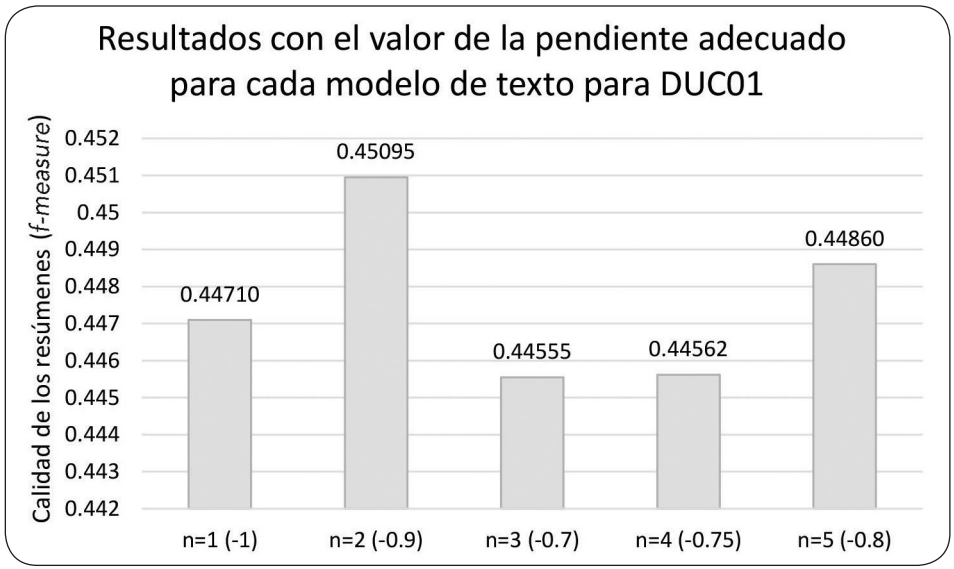
Como se ha mencionado, la pendiente de la recta (*m*) nos puede ayudar a determinar la importancia de las oraciones. El valor de *m* puede variar para poder suavizar dicha importancia.

En la **gráfica 5.20**, se observa una representación de la pendiente de una recta. Para la evaluación del método se consideran los siguientes valores de la

Gráfica 5.20 Representación gráfica del valor de la pendiente de una recta



Gráfica 5.21 Resultados de la pendiente adecuada para cada modelo de texto para DUC01



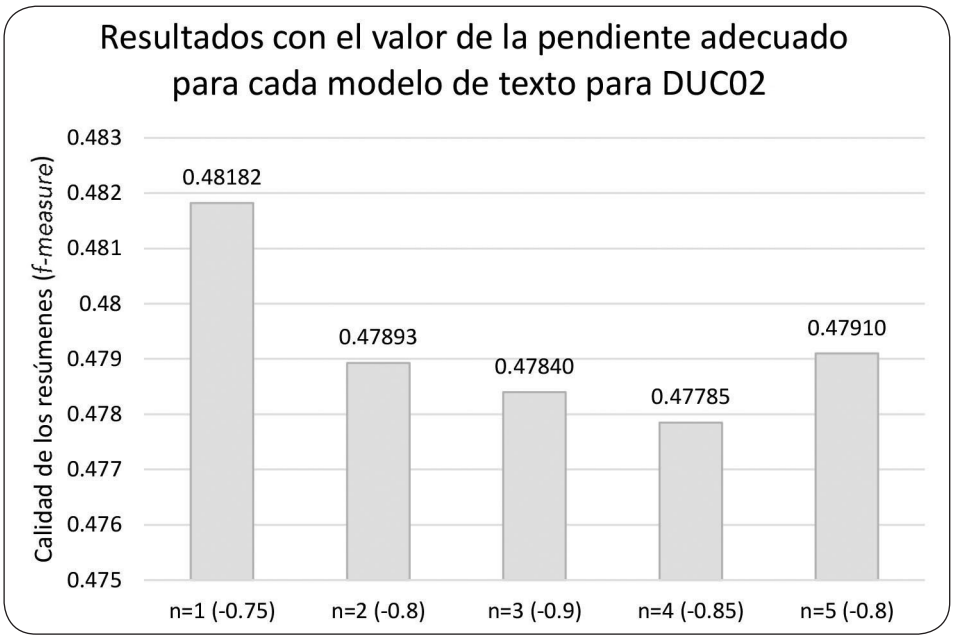
pendiente: $m = -0.25$, $m = -0.3$, $m = -0.375$, $m = -0.45$, $m = -0.5$, $m = -0.55$, $m = -0.6$, $m = -0.625$, $m = -0.65$, $m = -0.7$, $m = -0.75$, $m = -0.8$, $m = -0.85$, y $m = -0.9$. Estos valores se tomaron de forma aleatoria.

Como se mencionó, el modelo de texto utilizado en Matias (2016) es n-gramas. Se hizo un análisis por modelo de texto ($n = 1$ hasta $n = 1$) para determinar el mejor valor para la pendiente (posición de las oraciones) y el mejor modelo de texto para las colecciones DUC01 y DUC02. A continuación, en la **gráfica 5.21** se muestra el mejor valor obtenido para cada modelo de la colección DUC01.

El mejor modelo de texto para la colección DUC01 es bigramas ($n = 2$) con una pendiente de $m = -0.9$.

Para la colección DUC02 también se probaron los modelos de texto n-gramas (**gráfica 5.22**).

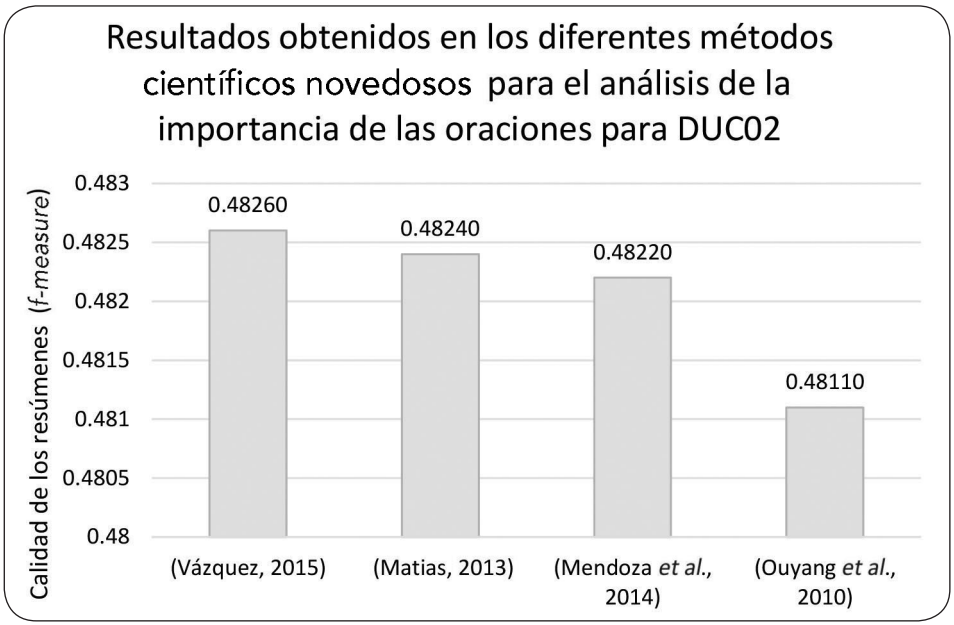
Gráfica 5.22 Resultados con el valor de la pendiente adecuada para cada modelo de texto



El mejor modelo de texto para la colección DUC02 fue *n-gramas* con (modelo "bolsa de palabras").

Una de las características más utilizadas en la GART es la posición de las oraciones. Por esto, Matias (2016) hace un estudio de las diferentes formas de calcular dicha posición. En la **gráfica 5.23** se muestran los resultados obtenidos aplicando los diferentes métodos al estado del arte para la colección DUC02.

Gráfica 5.23 Resultados obtenidos en los diferentes métodos científicos novedosos para el análisis de la importancia de las oraciones



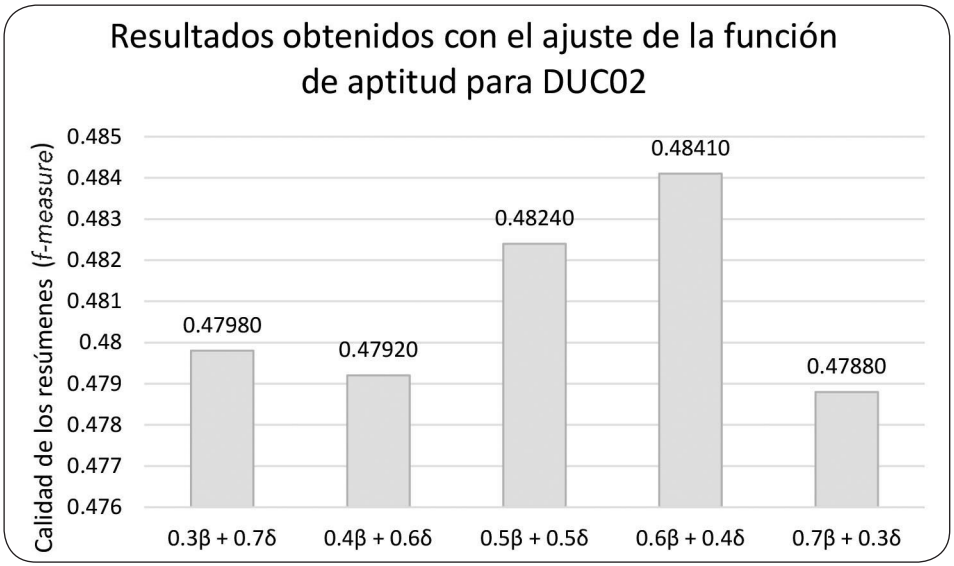
Como se puede observar, la fórmula propuesta en Vázquez (2015) es la que obtiene mejores resultados para la colección DUC02.

Para la colección DUC01 se hicieron pruebas con la forma propuesta por Vázquez (2015). Sin embargo, los resultados no fueron favorables (Alvarado B., 2017).

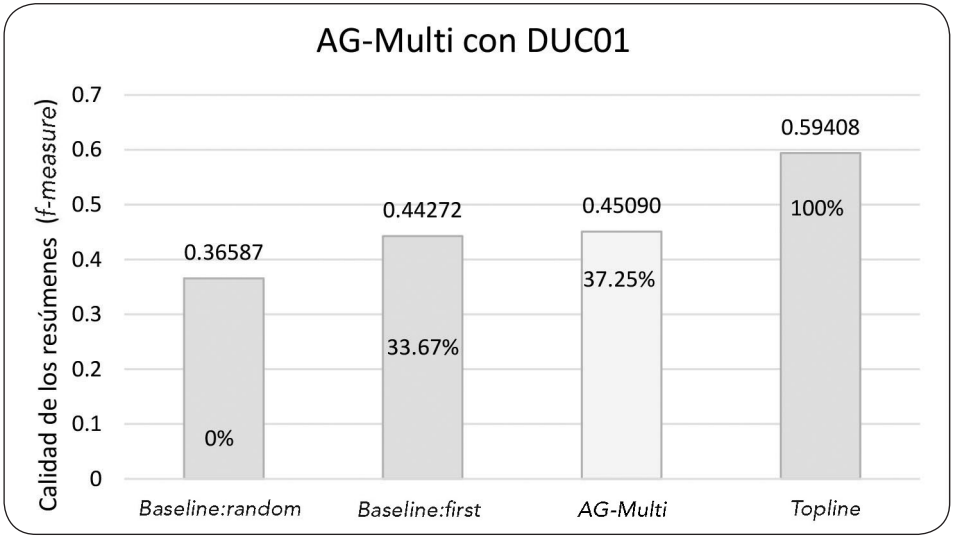
Además de la posición de las oraciones, el método de Matias (2016) considera la frecuencia de términos por lo que se realizó un ajuste de pesos en los parámetros β y δ la colección DUC02.

Los mejores resultados para la colección DUC01 usando el método de Matias (2016) se obtienen con el modelo de texto bigramas y el valor de pendiente $m = -0.9$.

Gráfica 5.24 Resultados obtenidos con el ajuste de la función de aptitud

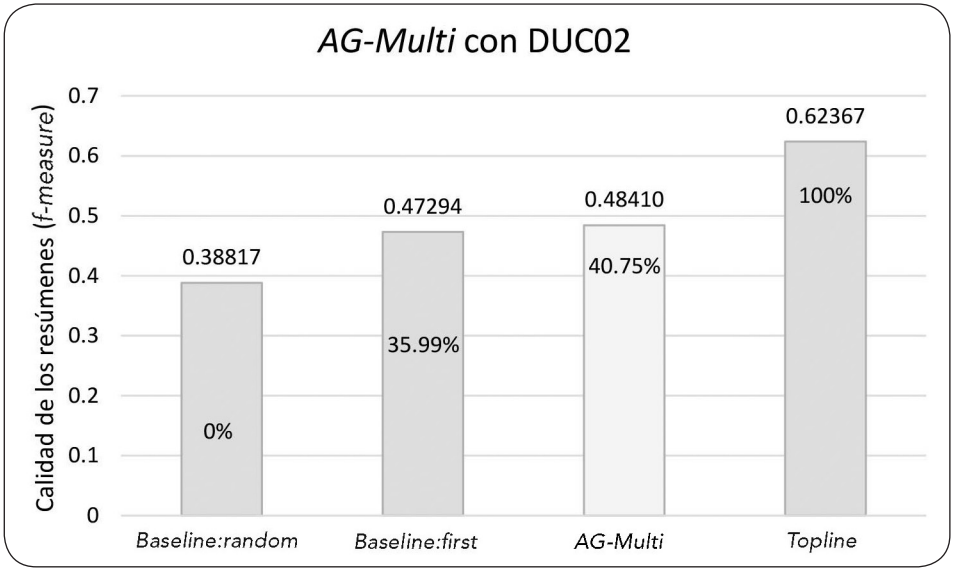


Gráfica 5.25 Resultados de *AG-Multi* usando DUC01 en comparación con las diferentes heurísticas



Para la colección DUC02 el mejor resultado se obtiene con el modelo de texto "bolsa de palabras", al aplicar la fórmula propuesta en Vázquez (2015) y

Gráfica 5.26 Resultados del método *AG-Multi* usando DUC02 en comparación con las diferentes heurísticas



con una combinación de 0.6 en frecuencia de términos y 0.4 en posición de las oraciones.

Este método ha sido probado para los leguajes inglés, español, portugués y ruso; puede funcionar con o sin preprocesamiento, dando resultados aceptables, ya que para cada lenguaje probado se ha superado el *baseline:first*.

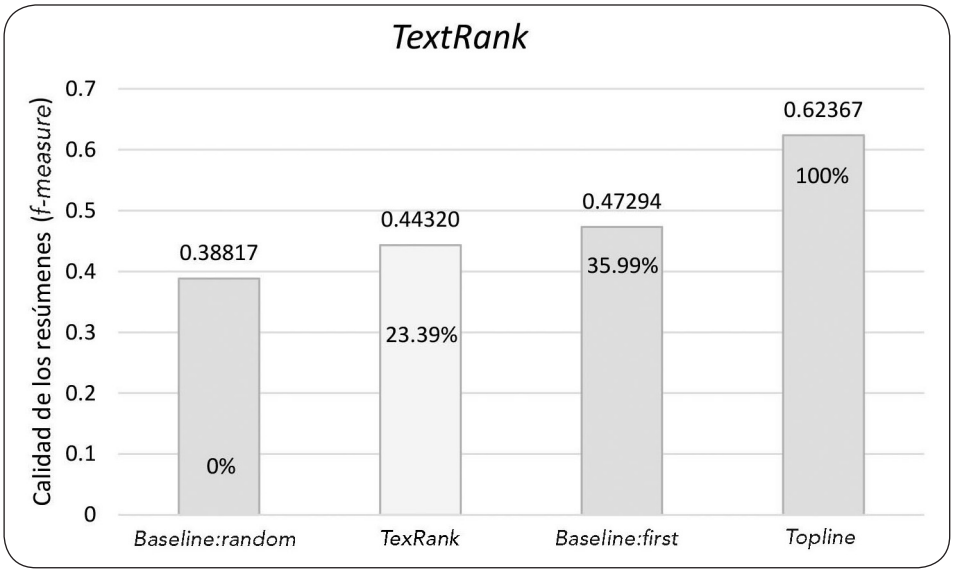
5.4.6 *TextRank*

Este método consiste en un algoritmo de ponderación basado en grafos. De acuerdo con Rada Mihalcea (Mihalcea, 2004) se construye un grafo para representar el texto, de manera que los nodos son palabras (u otras entidades de texto) interconectadas mediante arcos con relaciones significativas. Para la tarea de extracción de oraciones, el objetivo es calificar oraciones enteras y ordenarlas de mayor a menor importancia. Por lo tanto, se agrega un arco al grafo por cada oración en el texto.

Para establecer las conexiones entre oraciones se define una relación de similitud, donde la relación entre dos oraciones puede ser vista como un proceso

de "recomendación": una oración que señala a cierto concepto en el texto brinda al lector una "recomendación" para referirse a otras que señalan los mismos conceptos y, por tanto, un vínculo puede establecerse entre dos oraciones cualesquiera que compartan un contenido común. Para probar este método, se utiliza el *corpus* DUC02.

Gráfica 5.27 Resultados del método *TextRank* usando DUC02 en comparación con las diferentes heurísticas



A pesar de ser uno de los métodos más reconocidos y usados en el estado del arte, para la colección DUC02 TexrRank no supera a la heurística *baseline:first* (**gráfica 5.27**).

5.4.7 Secuencias Frecuentes Maximales (SFM *k-best*)

Este trabajo presenta un método basado en estadística, que es independiente del dominio y del lenguaje, para generar el resumen extractivo de un solo documento. En su trabajo, Ledeneva (2008), (*et al.*, 2008) muestra experimentalmente que las palabras que son partes de bigramas (secuencias de dos palabras) repetidas más de una vez en el texto, son buenos términos para describir el contenido del mismo, al

igual que las llamadas Secuencias Frecuentes Maximales (secuencias de palabras que se repiten cierto número de veces y que además no están contenidas en otras frecuentes). También se muestra que la frecuencia del término, como pesado de términos, brinda buenos resultados (mientras sólo se cuenten las ocurrencias de uno de ellos en bigramas repetitivos).

Ledeneva aplica una técnica de cuatro pasos para generar el resumen. Dichos pasos son la selección de términos, pesado de términos, pesado y selección de oraciones. En la selección de términos se extraen: las SFM, los bigramas repetitivos (deben aparecer por lo menos dos veces en el texto), y las palabras simples o unigramas. En el pesado de términos se usa la frecuencia del término, que consiste en el número de veces que aquél ocurre en el texto dentro de una SFM. También se utiliza como pesado la máxima longitud de una SFM que contenga al término, así como el asignar un mismo peso para todos. En el pesado de oraciones sólo se suma el peso de todos los términos contenidos en esa oración.

Finalmente, la selección de las oraciones que conformarán el resumen se lleva a cabo mediante dos criterios: primero, se eligen las mejores oraciones, es decir, las que obtuvieron mayor peso; esto se lleva a cabo hasta alcanzar la longitud deseada (cien palabras) del resumen. En el segundo criterio se seleccionan las k oraciones mejores, además de las primeras que aparecen en el documento (kbest+first). Esto se realiza hasta alcanzar la longitud deseada del resumen.

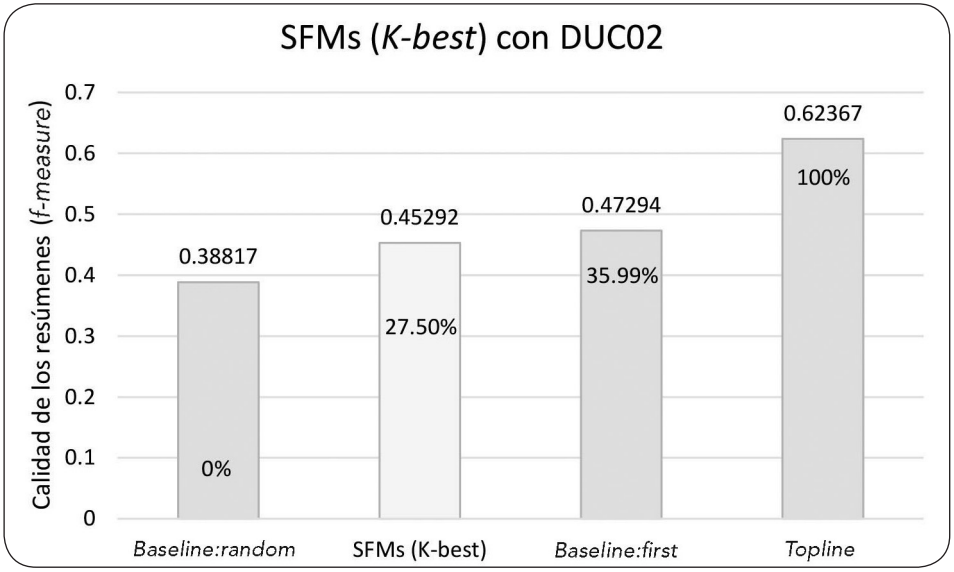
El método SFM (K-best) no supera a la heurística *baseline:first* (**gráfica 5.28**).

5.4.8 SFM (1BEST + FIRST)

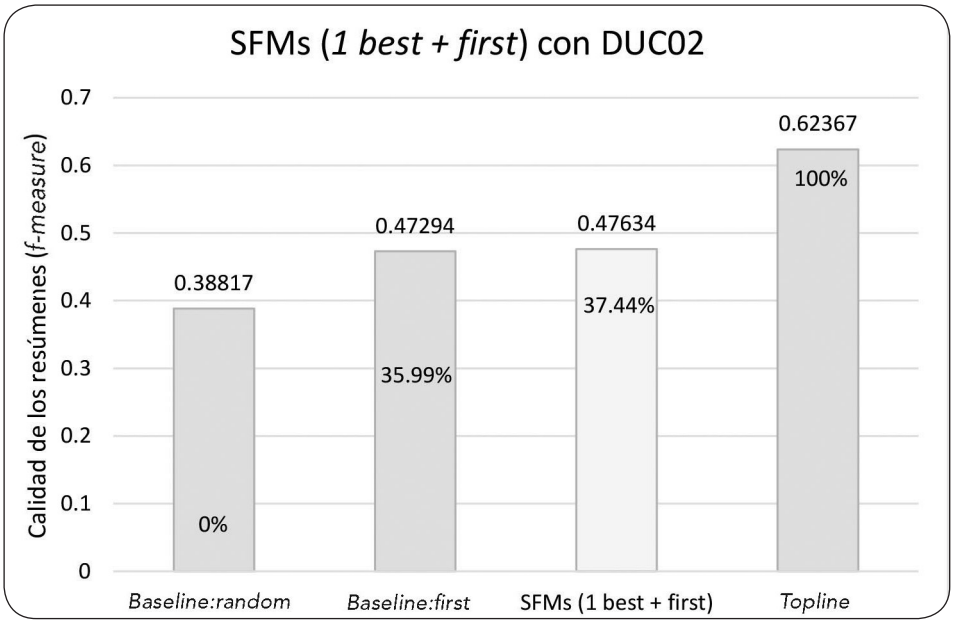
En este trabajo se presenta un método basado en estadística, que es independiente del dominio y del lenguaje, para generar el resumen extractivo de un solo documento. En su trabajo, Ledeneva (2008) muestra experimentalmente que las palabras que son partes de bigramas y se repiten más de una vez en el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

Fue hasta el año 2008 que los métodos científicos novedosos empezaron a superar a la heurística *baseline:first.* Como referencia, está SFM (1best + first) (**gráfica 5.29**).

Gráfica 5.28 Resultados del método SFM (K-best) usando DUC02 en comparación con las diferentes heurísticas



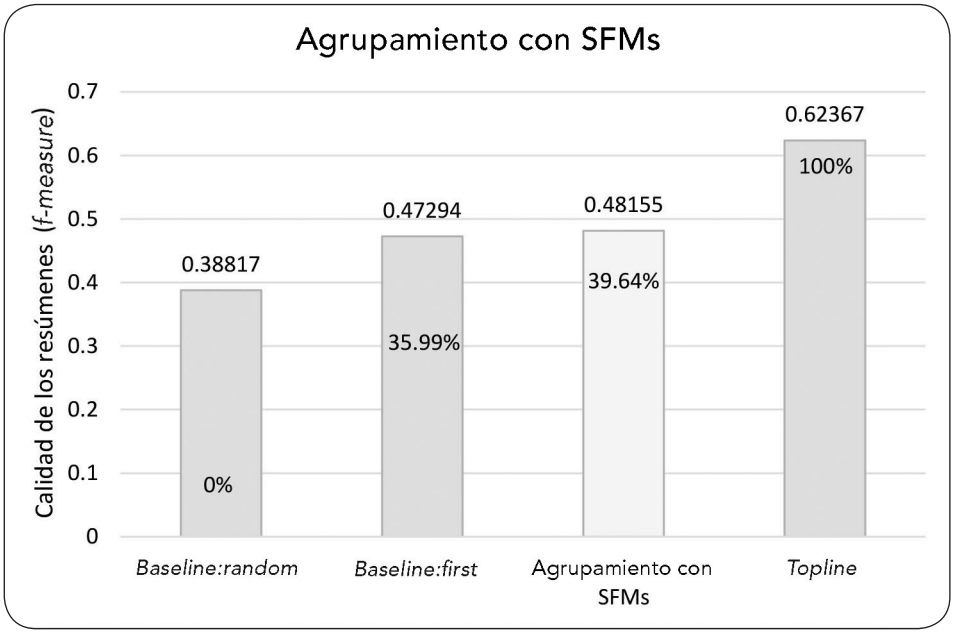
Gráfica 5.29 Resultados de SFM (1best + first) usando DUC02 en comparación con las diferentes heurísticas



5.4.9 Agrupamiento con SFM

En el método anterior de SFM, las sentencias que tienen mayor peso son seleccionadas para componer el resumen. Sin embargo, si existen sentencias muy similares y se eligen, no proporcionan nueva información al resumen. En este trabajo de agrupamiento (clustering) de oraciones con SFM y K-means (García-Hernández *et al*., 2008) se realizan grupos de oraciones, de las cuales se elige la frase más repetitiva de cada grupo y ésta compone el texto. Este trabajo también se hizo con agrupamiento de EM en el trabajo (Ledeneva *et al.*, 2011). Para probar este método se utiliza el *corpus* DUC02.

Gráfica 5.30 Resultados de Agrupamiento con SFM usando DUC02 en comparación con las diferentes heurísticas

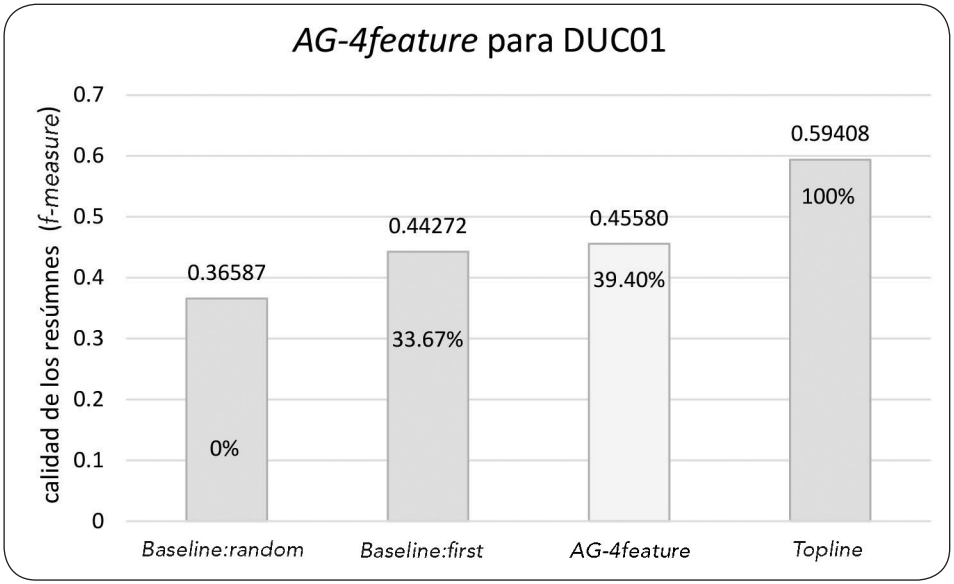


Dentro de los métodos para el lenguaje inglés que logran superar a la heurística *baseline:fisrt*, está Agrupamiento con SFM (**gráfica 5.30**).

5.4.10 *AG-4FEATURE*

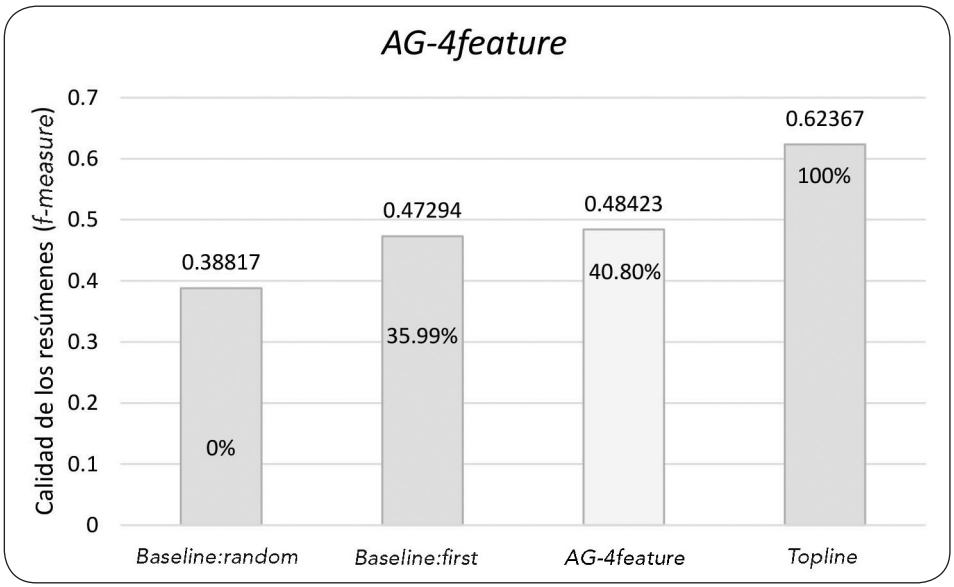
Vázquez (2018) presenta un método para optimizar la combinación de las características: similitud con el título(δ), posición de las oraciones (β), longitud de la oración(γ) y cobertura (α), basado en un algoritmo genético para cada etapa. En su trabajo, concluye que la característica más importante para el lenguaje inglés es: $\alpha = 0.59$, $\beta = 0.36$, $\gamma = 0.02$, $\delta = 0.03$ (Vázquez, Arnulfo García-Hernández, *and* Ledeneva, 2018). A continuación, se presentan los resultados para las colecciones DUC01 y DUC02.

Gráfica 5.31 Resultados de *AG-4feature* usando DUC01 en comparación con las diferentes heurísticas



En las **gráficas 5.31** y **5.32**, se muestran los resultados de la evaluación de los *corpus* DUC01 y DUC02 aplicados al método *AG-4feature*; se observa que el método es uno de los mejores: muestra un avance de 39.40% para el *corpus* DUC01 y de 40.80% para DUC02.

Gráfica 5.32 Resultados de *AG-4feature* usando DUC02 en comparación con las diferentes heurísticas



5.5 Resultados y análisis

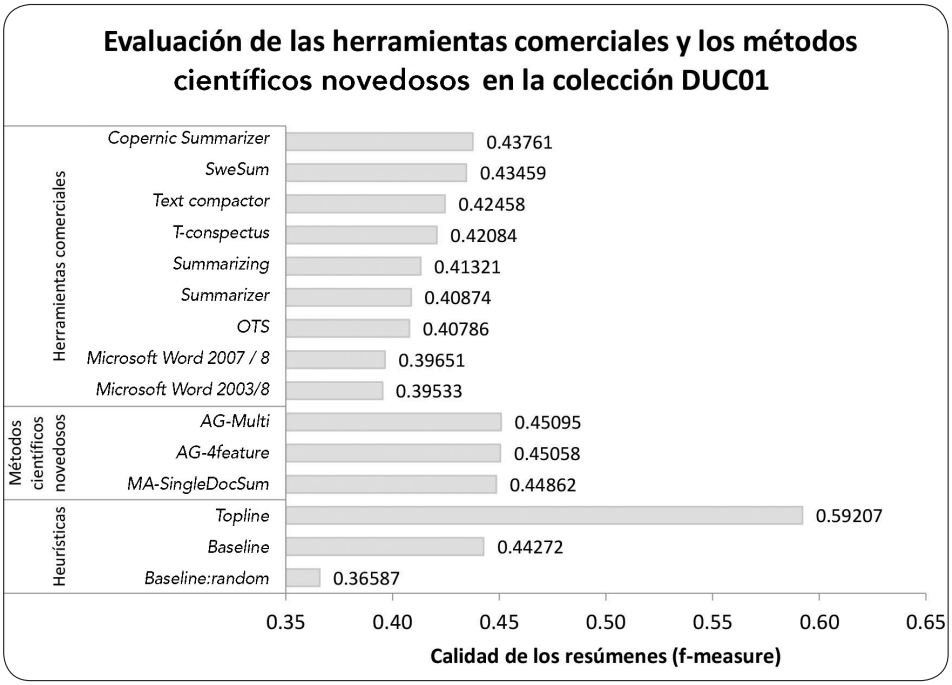
Para realizar la evaluación y comparación de las herramientas comerciales de los métodos científicos novedosos para la GART en el lenguaje inglés, se evaluaron las colecciones DUC01 y DUC02 con *ROUGE*.

La **gráfica 5.33** muestra una comparación entre los métodos científicos novedosos, las herramientas comerciales y las heurísticas para el *corpus* en inglés DUC01.

Para los *corpus* DUC01 y DUC02 en todos los métodos, las herramientas comerciales superan a la heurística *baseline:random* a la cual consideramos la peor forma de realizar un resumen. La segunda heurística a rebasar es *baseline:first*. Cabe mencionar que la GART en inglés, se superó la heurística *baseline: first* hace diez años.

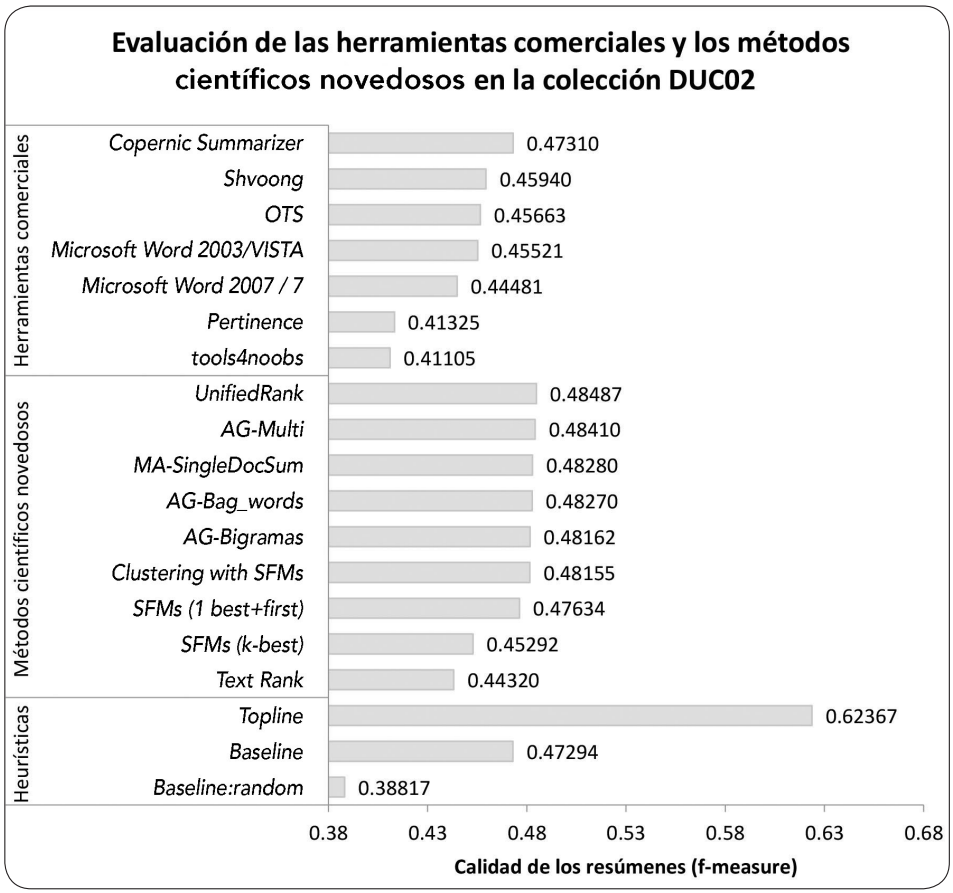
En la **gráfica 5.34** se muestra una comparación entre los métodos científicos novedosos, las herramientas comerciales y las heurísticas para el *corpus* en inglés DUC02. Los resultados están agrupados por herramientas comerciales, métodos científicos novedosos y, finalmente, las heurísticas.

Gráfica 5.33 Evaluación de las herramientas comerciales y los métodos científicos novedosos en la colección DUC01



Para la prueba del *Test de Turing* realizada para el lenguaje inglés, se consideró a la herramienta comercial *Copernic Summarizer* y al método de Matias (2016). El resumen 1 corresponde a la herramienta *Copernic Summarizer* y el resumen 3, al método de Matias (2016) (ver sección 1.2).

Gráfica 5.34 Evaluación de las herramientas comerciales y los métodos científicos novedosos en la colección DUC02



Capítulo VI

Generación automática de resúmenes para el lenguaje español

Este capítulo está dedicado a la presentación puntualizada del estudio de la tarea de GART para el lenguaje español. Se describen las diferentes conferencias, talleres y *corpus* que hay para dicho lenguaje. En especial, se describe el *corpus* TER, el cual es utilizado para hacer pruebas en esta lengua. También se muestran los resultados de las principales heurísticas, las herramientas comerciales y los métodos científicos novedosos probados al respecto. Finalmente, se observa una comparación general de dichas heurísticas y métodos, además de las herramientas comerciales probadas con el *corpus* TER.

El español es la segunda lengua nativa más hablada en el mundo, cuenta con aproximadamente 477 millones de personas. Actualmente, la hablan 572 millones de personas, ya sea como nativa, segunda o extranjera. El español es hablado principalmente en países como México, Colombia, España, Argentina, Perú, Venezuela, Chile, Ecuador, Guatemala, Cuba, Bolivia, República Dominicana, Honduras, Paraguay, El Salvador, Nicaragua, Costa Rica, Panamá, Puerto Rico, Uruguay y Guinea Ecuatorial. Además de ser usada como lengua alternativa en países como Argelia, Australia, Brasil, Canadá, Estados Unidos, China, India, Israel, Japón, Noruega, Rusia, Suiza, Turquía, entre otros, por lo que se ha convertido en la segunda lengua de comunicación a nivel internacional (Arévalo, 2017).

Tabla 6.1 Principales lenguas habladas en el mundo

No.	Lenguaje	Países	Hablantes
1	Chino	35	1302
2	Español	21	427
3	Inglés	106	339
4	Árabe	58	267
5	Hindi	4	260
6	Portugués	12	202
7	Bengalí	4	189
8	Ruso	17	171
9	Japonés	2	128
10	Lahnda	8	117
11	Javanés	3	84.3
12	Coreano	7	77.3
13	Alemán	26	76.9
14	Francés	53	75.9
15	Télugu	2	74.2
16	Marathi	1	71.4
17	Turco	8	71.4
18	Urdu	6	68.6
19	Vietnamita	3	68

20	Tamil	7	67.8
21	Italiano	13	63.4
22	Persa	30	61

El español también tiene una importante participación en Internet y las redes sociales; en la actualidad es el tercer más empleado en Internet por número de internautas.

Tabla 6.2 Lenguas más usadas en Internet[26]

No.	Lenguaje	Usuarios de Internet
1	Inglés	1052
2	Chino	804
3	Español	337
4	Árabe	219
5	Portugués	169
6	Hindi	168
7	Francés	134
8	Japonés	118
9	Ruso	109
10	Alemán	92
	Otra	950

Como lo muestra la **tabla 6.2**, el inglés y el chino superan en número al español. Sin embargo, si se tiene en cuenta que el chino es una lengua que, en general, sólo la hablan sus nativos, el español se colocaría como la segunda de comunicación en Internet (Arévalo, 2017).

A sesenta años de la investigación de la GART y siendo el español uno de los principales lenguajes a nivel mundial, han sido pocas las investigaciones que se han realizado para la tarea. De manera particular, surge el interés por saber en qué nivel se encuentra el estudio de la GART en nuestra lengua materna.

[26]Según un estudio que revela las lenguas más usadas en Internet: https://www.internetworldstats.com/stats7.htm

En el capítulo I se ha brindado una prueba del *Test de Turing* para el lenguaje español (**tabla 1.1**). Los resultados mostraron que sólo 8% de las ocasiones las personas acertaron en indicar los resúmenes hechos por el humano; mientras que 56% se confunde y selecciona un resumen producido por una máquina y uno por un humano y, sorprendentemente, 36% selecciona los dos hechos por la máquina como generados por los humanos. Sin embargo, en la prueba del capítulo I solamente se dividen los resúmenes como elaborados por el humano o por la máquina, mientras que dentro de los resúmenes clasificados como hechos por la máquina están las heurísticas. Para conocer de manera más específica los resultados para el español, es necesario realizar una clasificación completa y correcta, ya que las heurísticas son utilizadas como referencia a la evaluación de la tarea de GART.

En la **tabla 6.3** se muestran los porcentajes de confusión que tiene el humano al seleccionar alguna de las heurísticas. En la primera fila se observa que en 18% de las ocasiones las personas confundieron a la heurística *baseline:first* como la generada por un humano, y 26% eligió a la heurística *baseline:first* y a un resumen hecho por la máquina como los generados por el humano. La heurística *baseline:random* también fue seleccionada en 15%, las personas eligieron esta heurística como el resumen hecho por un humano. Pero, además, 23% se confundió al seleccionar la heurística *baseline:random* y el resumen hecho por la máquina. Los resultados obtenidos de estas pruebas muestran que *baseline:first* tiene considerable correlación sobre los humanos.

Tabla 6.3 Resultados del *Test de Turing* respecto de *baseline* para español

Pares de resúmenes con respecto a las heurísticas *baseline*		Porcentaje de confusión entre los resúmenes seleccionados (%)
Humano	– *Baseline:first*	18
Humano	– *Baseline:random*	15
Máquina	– Máquina	13
Máquina	– *Baseline:first*	26
Máquina	– *Baseline:random*	23
Baseline:random	– *Baseline:first*	5

6.1 Conferencias, talleres y *corpus*

Cabe mencionar que para el lenguaje español se tienen muy pocos recursos (conferencias, talleres y *corpus*), debido a que a pesar de ser uno de los más hablados, no se ha profundizado en la creación de los mismos para la tarea de la GART en español. A continuación se mencionan algunos de los *corpus* que han sido utilizados para generar resúmenes en español, aunque no han sido creados específicamente para esta tarea, sino que se han ajustado para su uso.

6.1.1 *Corpus* Desastres

El conjunto de datos Desastres consiste en trescientas noticias recolectadas de diferentes periódicos publicados en México. Cada una de las oraciones fue marcado utilizando dos etiquetas básicas: relevante y no-relevante (Téllez *et al.*, 2009). Este *corpus* fue creado principalmente para trabajar con sistemas de extracción de información. Sin embargo, se ocupó para un sistema de GART en Villatoro (2006) donde se produjeron resúmenes de cien palabras.

6.1.2 *Corpus* Concisus

Saggion y Szasz (2012) presentan un *corpus* bilingüe, comparable en español e inglés, de pares de resúmenes de tres tipos de eventos: accidentes aéreos, accidentes ferroviarios y terremotos. Fue creado manualmente con información semántica sobre cada evento, y resulta apropiado para la experimentación en extracción de información monolingüe y bilingüe. Cabe mencionar que no está enfocado en la GART como tal, sino en la extracción de información a través de resúmenes de eventos cortos. Este *corpus* no está etiquetado y no cuenta con alguna medida de *baseline* o *Topline*.

6.1.3 *Corpus* utilizado para la evaluación y la comparación

El *corpus* que se utiliza en este libro para la comparación de las herramientas comerciales y los métodos científicos novedosos fue creado especialmente para la tarea de GART en español (Matias, 2016).

Textos en Español para Resúmenes (TER) es una colección de documentos compuesta por doscientas cuarenta noticias en el lenguaje español. El *corpus* TER se integra de noticias periodísticas recabadas del periódico mexicano Crónica, sobre doce diferentes categorías. Para cada documento de la colección, se crearon dos resúmenes por dos humanos expertos.

Algunos de los criterios que se consideraron para la construcción del *corpus* son:

- TER es creado a partir de noticias.
- Es para el lenguaje español.
- Tiene el fin de ser utilizado para GART extractivos.
- Los resúmenes son para un solo documento.
- Las noticias están en formato digital.
- La longitud de los resúmenes debe ser igual o mayor a cien palabras.

Para la construcción de *corpus* se seleccionaron veinte noticias de las siguientes categorías: academia, bienestar, ciudad, cultura, deportes, espectáculos, estados, mundo, nacional, negocios, opinión y sociedad; con un total de doscientos cuarenta textos. Una de las consideraciones más importantes para la selección de las noticias fue que tuvieran diferentes longitudes, pero siempre más de cien palabras.

En la **tabla 6.4** se muestran las categorías en las que está dividido el *corpus* TER: el número de documentos por cada categoría, el total de documentos que componen al *corpus*, el número de oraciones por cada categoría y el promedio de palabras para cada texto. En promedio cada texto tiene 442 palabras y 14 oraciones, entre una a dos páginas por noticia.

6.2 Heurísticas

Para realizar el cálculo de las heurísticas, las herramientas comerciales y los métodos científicos novedosos se utiliza el *corpus* TER, debido a que es el único disponible y especial para la tarea de GART en español.

Tabla 6.4 Parámetros de los textos completos del *corpus* TER

Periódico	Categoría	Número de textos	Número de palabras	Promedio de palabras	Número de oraciones	Promedio de oraciones
Crónica	Academia	20	10966	548.3	382	19.1
	Bienestar	20	11801	590.05	405	20.25
	Ciudad	20	7568	378.4	219	10.95
	Cultura	20	8631	431.55	297	14.85
	Deportes	20	9519	475.95	363	18.15
	Espectáculos	20	8869	443.45	311	15.55
	Estados	20	7471	373.55	185	9.25
	Mundo	20	7108	355.4	247	12.35
	Nacional	20	7533	376.65	186	9.3
	Negocios	20	7523	376.15	229	11.45
	Opinión	20	12716	635.8	443	22.15
	Sociedad	20	6507	325.35	228	11.4
	Total	240	106212		3495	
	Promedio			442.55		14.5625

6.2.1 *Baseline:random*

En la **tabla 6.5** se presentan los resultados obtenidos por la heurística *baseline:random* para el *corpus* TER; cabe mencionar que se hicieron diez corridas para garantizar los resultados mostrados.

Tabla 6.5 Resultados de *baseline:random* para TER

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.4969	0.4980	0.4973
ROUGE-2	0.2930	0.2936	0.2933
ROUGE-SU4	0.3204	0.3208	0.3201

Para *baseline:random* los resultados tienden a ser bajos debido a que las oraciones son seleccionadas de forma aleatoria. Para el estado del arte *baseline:random* sirve como referencia del peor resultado obtenido.

6.2.2 *Baseline:first*

La **tabla 6.6** muestra los resultados obtenidos con la heurística *baseline:first* para el *corpus* TER.

Tabla 6.6 Resultados de *baseline:first* para TER

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.7233	0.7221	0.7226
ROUGE-2	0.6235	0.6224	0.6229
ROUGE-SU4	0.6332	0.6321	0.6326

Como se puede observar, los resultados de *baseline:first* son muy altos lo que muestra que las primeras oraciones para este *corpus* de noticias son muy importantes. Para el lenguaje español, esta heurística es un reto por superar pues su valor es muy alto. Los resultados mostrados en la sección 6.3 señalan que ninguna herramienta comercial la supera; en la sección 6.4 se ve que sólo un método del estado del arte lo puede hacer.

6.2.3 *Topline*

Para el *corpus* TER, el *Topline* alcanzado con *ROUGE-1* es de: 0.8344 en *f-measure*. Si se establece una comparación con el valor de la heurística *baseline:first*, se puede observar que el rango entre esta base a superar es muy corto. Sin embargo, al ser el español un lenguaje poco estudiado en la tarea de GART, se trata de un reto por alcanzar. El resultado de *Topline* se obtuvo mediante un algoritmo genético (Rojas J., 2017). En la **tabla 6.7** se muestran los datos obtenidos en las diferentes configuraciones de *ROUGE*.

Tabla 6.7 Resultados de *Topline* para TER

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.8369	0.8320	0.8344
ROUGE-2	0.7687	0.7642	0.7664
ROUGE-SU4	0.7672	0.7627	0.7649

6.3 Herramientas comerciales

Para el lenguaje español también se utilizaron herramientas comerciales probadas con el *corpus* TER. La longitud requerida para los resúmenes hechos por este *corpus* es de cien palabras, por lo que aquellos generados por los métodos y las herramientas deben ser igual a cien palabras. Para calcular el porcentaje que corresponde al mínimo en número de palabras se hace uso de la siguiente fórmula.

$$\frac{\textit{número de palabras deseadas}}{\textit{número de palabras totales en el documento}} * 100 \qquad (10)$$

Para la GART en español se utilizan las siguientes herramientas comerciales (**tabla 6.8**).

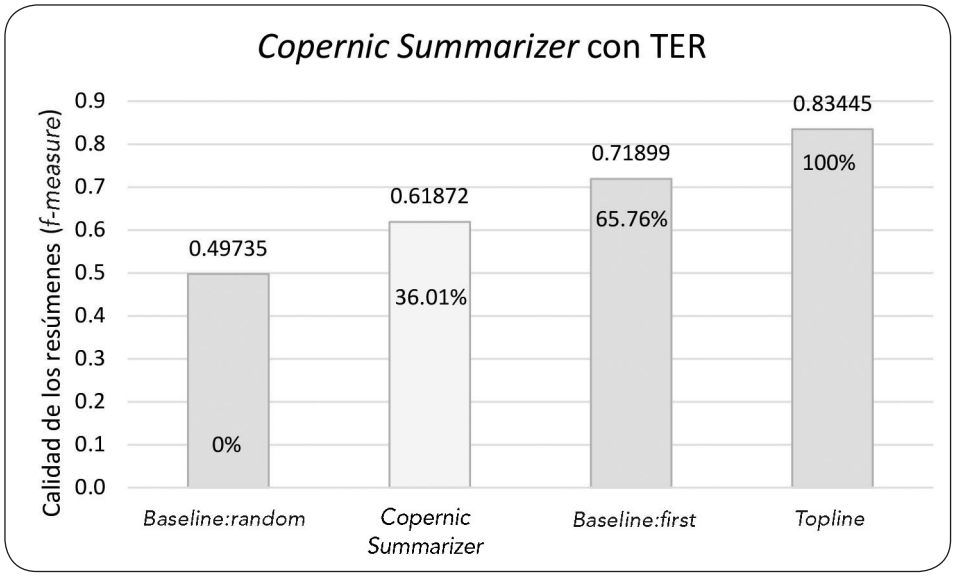
Tabla 6.8 Listado de herramientas probadas en español

Herramienta	Tipo	Español
Copernic Summarizer	Instalable	✓
Microsoft Office Word 2003/2007	Instalable	✓
OTS	Línea	✓
Text Compactor	Línea	✓
Summarizing	Línea	✓
Total		5

6.3.1 *Copernic Summarizer*

Copernic Summarizer tiene la opción de generar resúmenes de cien palabras (longitud requerida para TER), por lo que se seleccionó esta opción. A continuación, se muestran los resultados obtenidos por esta herramienta evaluados con *ROUGE*.

Gráfica 6.1 Resultados de *Copernic Summarizer* con TER en comparación con las diferentes heurísticas



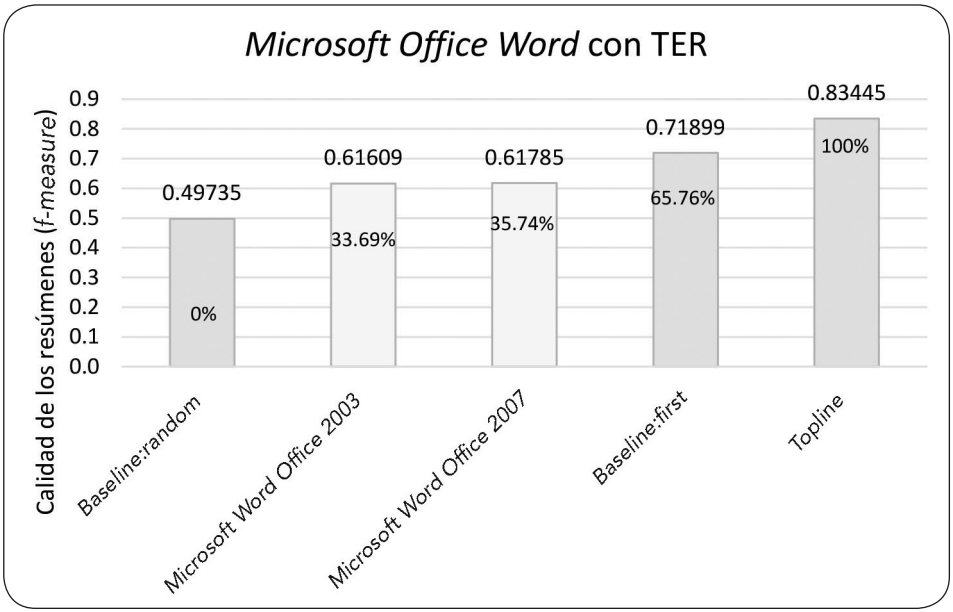
Como se puede observar, los resultados de *Copernic* no superan a la heurística *baseline:first* (**gráfica 6.1**).

6.3.2 *Microsoft Office Word*

A continuación, se muestran los resultados obtenidos de *Microsoft Office Word* evaluados con la herramienta *ROUGE*.

En la **gráfica 6.2** podemos observar que esta herramienta no supera a *baseline:first* y la diferencia entre versiones de *Microsoft Office Word* (2003 y 2007) no es significativa en cuanto a la GART.

Gráfica 6.2 Resultados de *Microsoft Office Word* usando TER en comparación con las diferentes heurísticas

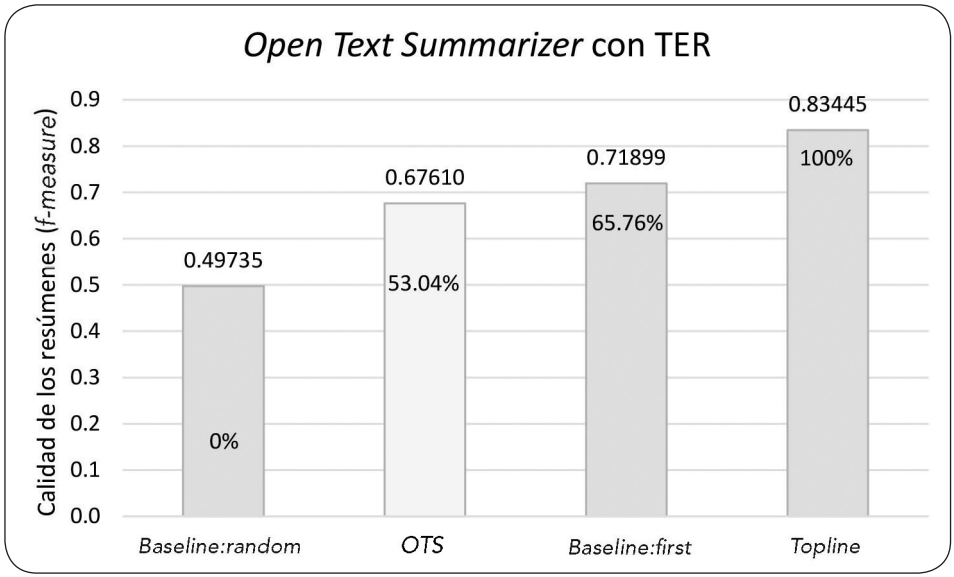


6.3.3 *Open Text Summarization*

Para la herramienta *Open Text Summarizer (OTS)* es necesario calcular el porcentaje correspondiente a cada documento, de tal forma que para cada uno se tengan cien palabras, como lo indican las especificaciones propias del *corpus* TER y, para ello, se hace uso de la fórmula número 10. A continuación, se muestran los resultados obtenidos por esta herramienta evaluados con *ROUGE*.

En la **gráfica 6.3** se muestran los resultados obtenidos con *Open Text Summarizer* usando el *corpus* TER. Como se puede observar, no se supera a *baseline:first*. Si se considera a *baseline:random* como el peor resumen que se puede generar y a *Topline* como el máximo a obtener, entonces *Open Text Summarizer* obtiene 53.04% que corresponde hasta ahora al porcentaje más bajo entre los métodos científicos novedosos.

Gráfica 6.3 Resultados de *Open Text Summarizer* con TER en comparación con las diferentes heurísticas



6.3.4 *Text Compactor*

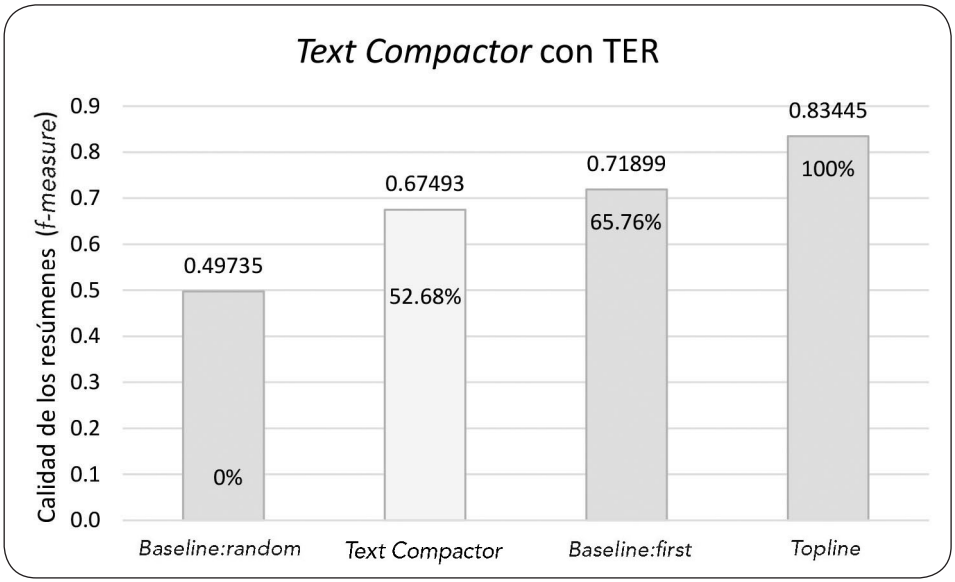
Para *Text Compactor* se tiene que calcular el porcentaje correspondiente a cada documento, de tal forma que para cada uno de ellos se tengan cien palabras. A continuación, se muestran los resultados obtenidos por esta herramienta evaluados con *ROUGE*.

En la **gráfica 6.4** se muestran los resultados para la herramienta *Text Compactor* con el *corpus* TER; como se puede observar, no superan la heurística *baseline:first*. Sin embargo, el porcentaje de avance que tiene respecto de la heurística *baseline:random* y *Topline* es de 52.68%.

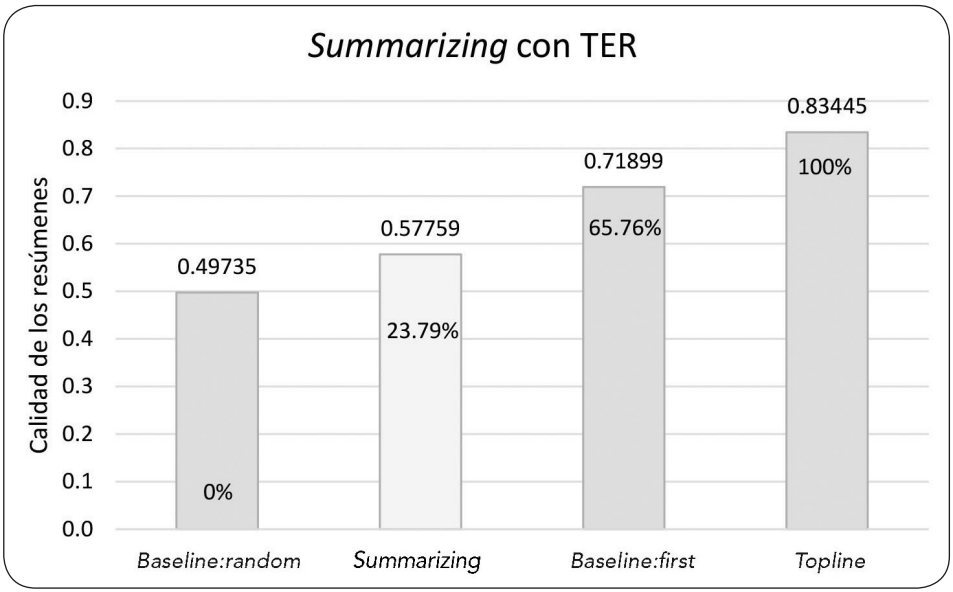
6.3.5 *Summarizing*

Summarizing tiene la opción de generar resúmenes a cien palabras (longitud requerida para TER), por lo que se seleccionó esta opción. A continuación, se muestran los resultados obtenidos evaluados con *ROUGE*.

Gráfica 6.4 Resultados de *Text Compactor* con TER en comparación con las diferentes heurísticas



Gráfica 6.5 Resultados de *Summarizing* con TER en comparación con las diferentes heurísticas



En la **gráfica 6.5** se muestran los resultados para *Summarizing* con el *corpus* TER. Como se puede observar, los resultados obtenidos con esta herramienta no superan la heurística *baseline:first,* y de las herramientas probadas es la que menos puntaje obtiene. Considerando a *baseline:random* como la peor forma de generar un resumen y a *Topline* como la mejor, *Text Compactor* obtiene 23.79% de avance.

6.4 Métodos científicos novedosos

Para la tarea de GART en español no se tienen investigaciones formales, con *corpus* y herramientas de evaluación que puedan ser comparables. Sin embargo, es importante saber cuál ha sido el esfuerzo realizado. A continuación, se describen algunos de los trabajos al respecto con *corpus* propios o ajustados de otras tareas.

6.4.1 Grafos semánticos

El trabajo de Plaza (2011) se completa con tres casos de estudio en los que el método diseñado se configura y utiliza para generar distintos tipos de resúmenes de textos de diversos dominios, y con unas características de estructura y estilo muy diferentes: artículos científicos de biomedicina, noticias periodísticas y páginas web de información turística en español.

El método que utiliza está basado en el uso de grafos semánticos, el cual contiene las siguientes etapas:

- Preprocesamiento.
- Traducción de las oraciones a conceptos.
- Representación de las oraciones como grados de conceptos, construcción del grafo del documento.
- *Clustering* de conceptos.
- *A*signación de oraciones a *clusters.*
- Selección de oraciones para el resumen.
- Construcción del resumen.

6.4.2 Compresión automática de frases

En el trabajo de Molina (2013) se propone la generación de resúmenes automáticos para el lenguaje español considerando las siguientes características del texto.

- La segmentación discursiva consiste en representar el documento a través de un árbol jerárquico que contiene información tipo retórico/discursivo.
- La comprensión de frases por eliminación de segmentos discursivos se basa en la gramaticalidad de la frase resultante; en su normatividad (entendida como la calidad de información importante retenida) y en la tasa de comprensión.
- La gramaticalidad, que consiste en determinar si una frase es correcta o no.
- La normatividad, basada en la frecuencia de las palabras.

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

6.4.3 Generación de resúmenes de múltiples documentos

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

Para la experimentación con el lenguaje en español se utiliza el *corpus* Desastres (Téllez *et al.*, 2009); el cual está diseñado para clasificación y fue adaptado para la GART.

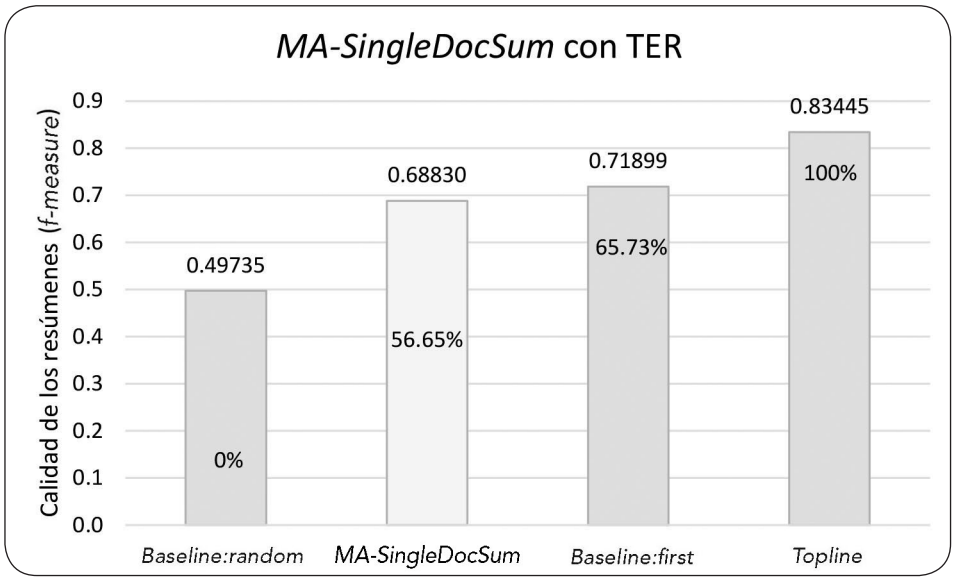
Para saber el estado en que se encuentra la investigación en el área de la GART, se probaron algunos de los mejores métodos científicos novedosos que

se han usado para el lenguaje inglés. A continuación se muestran los resultados obtenidos.

6.4.4 *MA-SingleDocSum*

Es uno de los mejores métodos presentados para el inglés. La descripción se hace en la sección 5.5.1.

Gráfica 6.6 Resultados de *MA-SingleDocSum* usando TER en comparación con las diferentes heurísticas



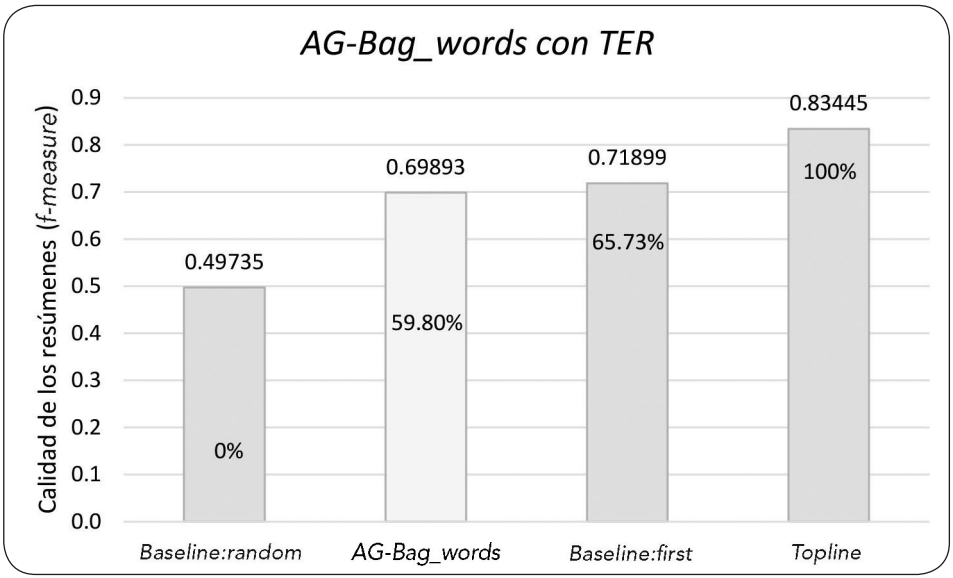
El método *MA-SingleDocSum* para el lenguaje español supera a la heurística *baseline:random* (**gráfica 6.6**), aunque no a *baseline:first* como lo hace para el inglés.

6.4.5 *AG-Bag_words*

AG-Bag_words es un método basado en un algoritmo genético aplicado únicamente al inglés. Sin embargo, por su composición puede trabajar con otros lenguajes, en este caso, para el español. La descripción de este método se hace en la sección 5.5.3.

El método *AG-Bag_words* para español no supera a la heurística *baseline:first* (**gráfica 6.7**).

Gráfica 6.7 Resultados de *AG-Bag_words* con TER en comparación con las diferentes heurísticas



6.4.6 *AG-Multi*

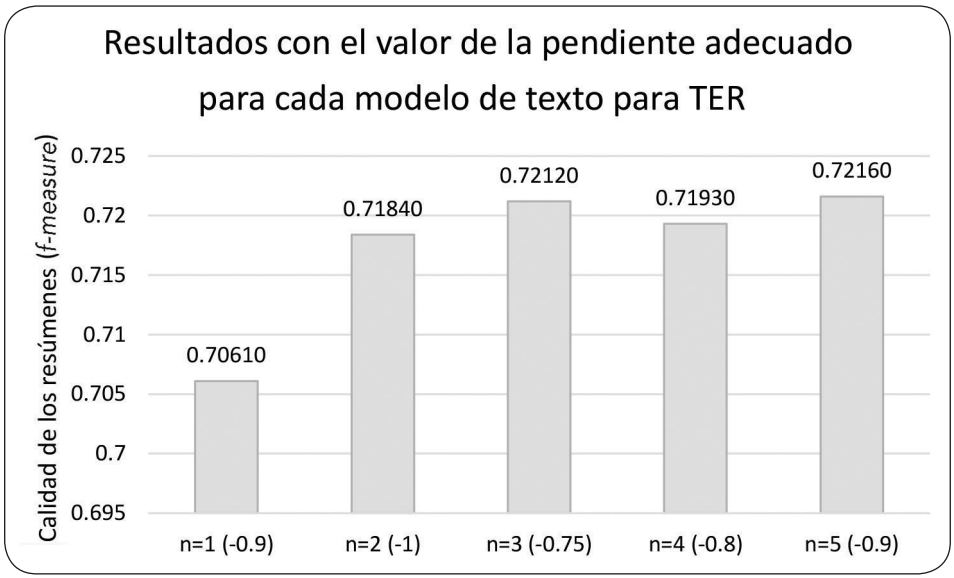
AG-Multi es un método basado en un algoritmo genético aplicado a varios lenguajes. La descripción se hace en la sección 5.5.5.

De acuerdo con las etapas realizadas en *AG-Multi* y propuestas por Matias (2016), los resultados obtenidos son los siguientes.

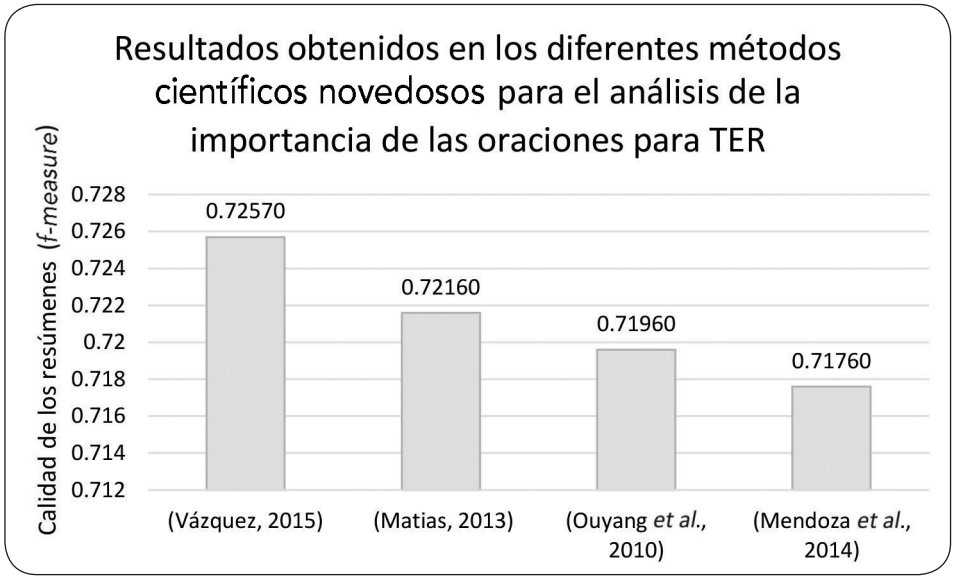
Para las pruebas por modelo de texto, el modelo *n-gramas* con $n = 5$ es el que mejor resultado obtiene para el lenguaje español y la mejor pendiente es $m = -5$ (**gráfica 6.8**). Cabe mencionar que para español los mejores resultados se obtuvieron sin preprocesamiento, y son lo que se presentan.

Al igual que para el lenguaje inglés, se hace un análisis de los métodos científicos novedosos que utilizan la característica posición de las oraciones (**gráfica 6.9**).

Gráfica 6.8 Resultados de la pendiente adecuada para cada modelo de texto para TER

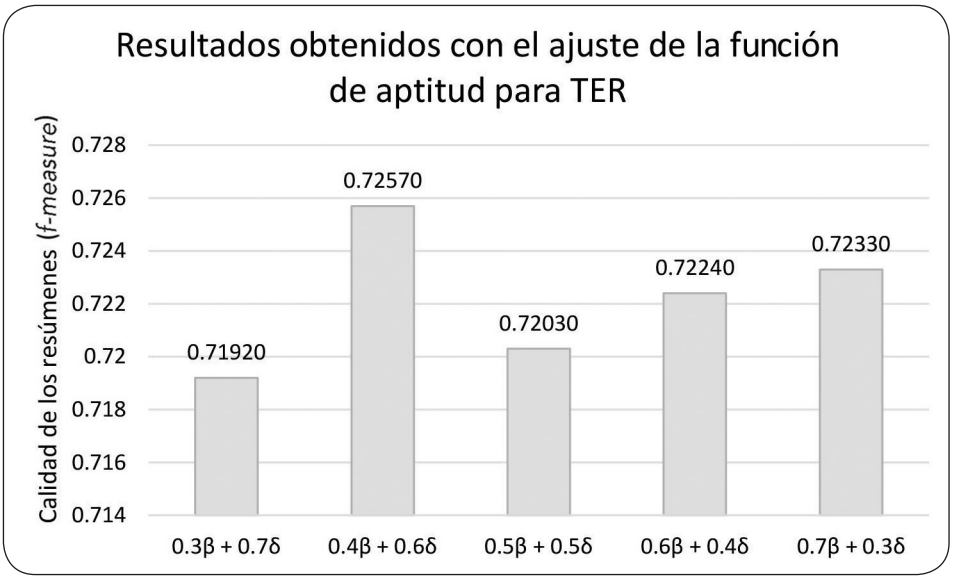


Gráfica 6.9 Resultados obtenidos en los diferentes métodos científicos novedosos para el análisis de la importancia de las oraciones



Para el español el ajuste de los parámetros β y δ para la colección TER, es de 0.4 para frecuencia de los términos y 0.6 en posición de las oraciones (**gráfica 6.10**). Esto significa que la posición de las oraciones importa más.

Gráfica 6.10 Resultados obtenidos con el ajuste de la función de aptitud



En la **gráfica 6.11** se muestra la comparación de los resultados obtenidos con el método *AG-Multi*, en comparación con las diferentes heurísticas.

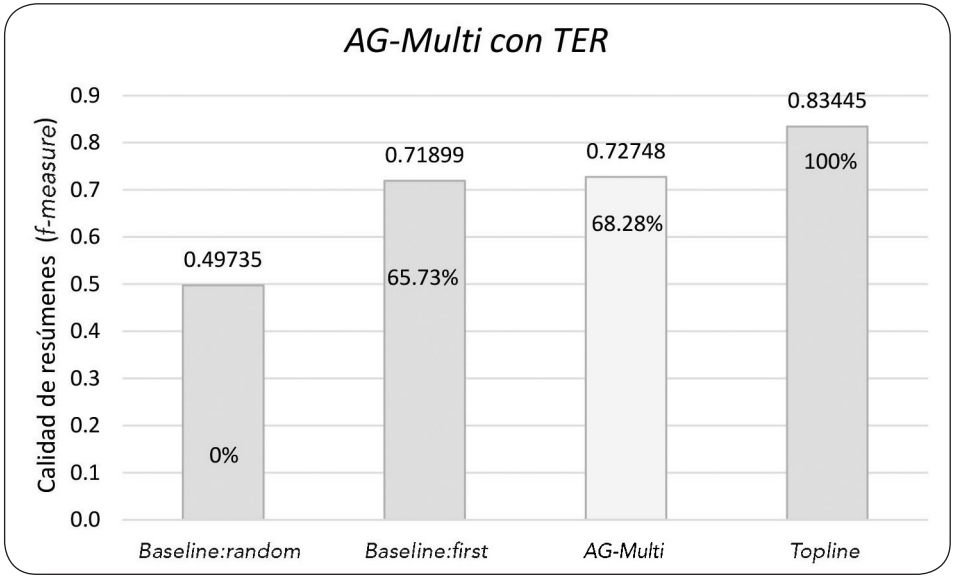
El método *AG-Multi* es el único que hasta ahora ha superado a las dos principales heurísticas: *baseline:random* y *baseline:first,* para el lenguaje español, con 67.75%.

6.4.7 *TextRank*

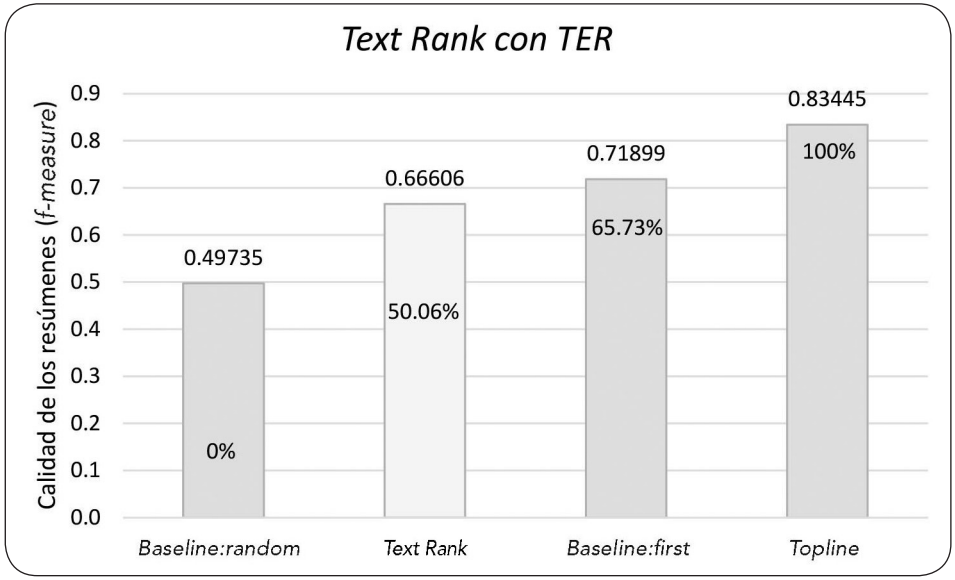
TextRank es un método basado en grafos, aplicado al inglés y al portugués. La descripción se hace en la sección 5.5.6.

TextRank es uno de los más utilizados para el estudio de la GART para el lenguaje inglés. Además de ser probado para aquél, se empleó en el portugués. Este método es independiente del lenguaje. Para español no supera a la heurística *baseline:first*. Sin embargo, sí rebasa a *baseline:random* (**gráfica 6.12**).

Gráfica 6.11 Resultados de *AG-Multi* usando TER en comparación con las diferentes heurísticas



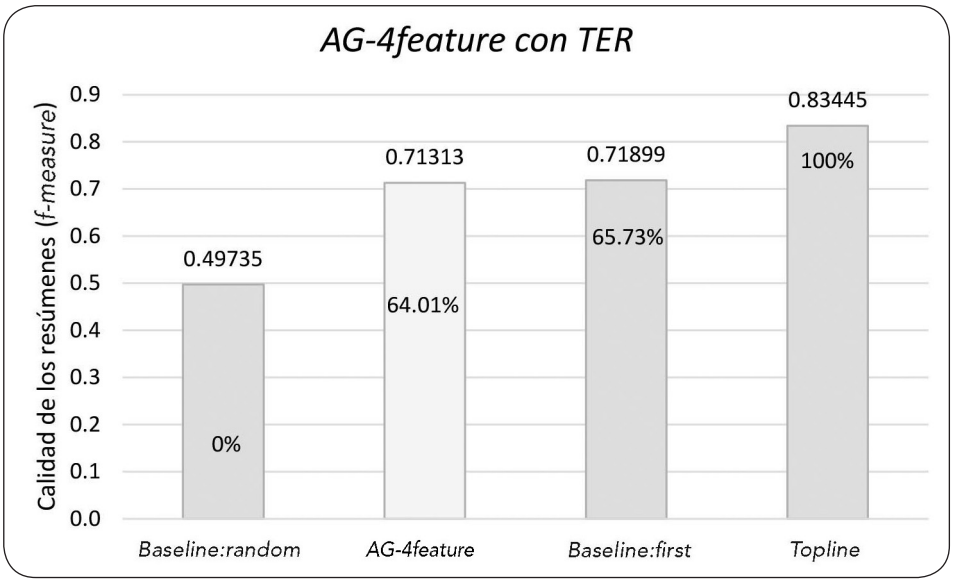
Gráfica 6.12 Resultados de Text Rank usando TER en comparación con las diferentes heurísticas



6.4.8 *AG-4feature*

Es un método para la GART en el lenguaje inglés. La descripción se hace en la sección 5.5.10.

Gráfica 6.13 Resultados de *AG-4feature* usando TER en comparación con las diferentes heurísticas



AG-4feature no supera a la heurística *baseline:first*. Sin embargo, no queda muy por debajo de ella, pues tiene 1.72% (**gráfica 6.13**).

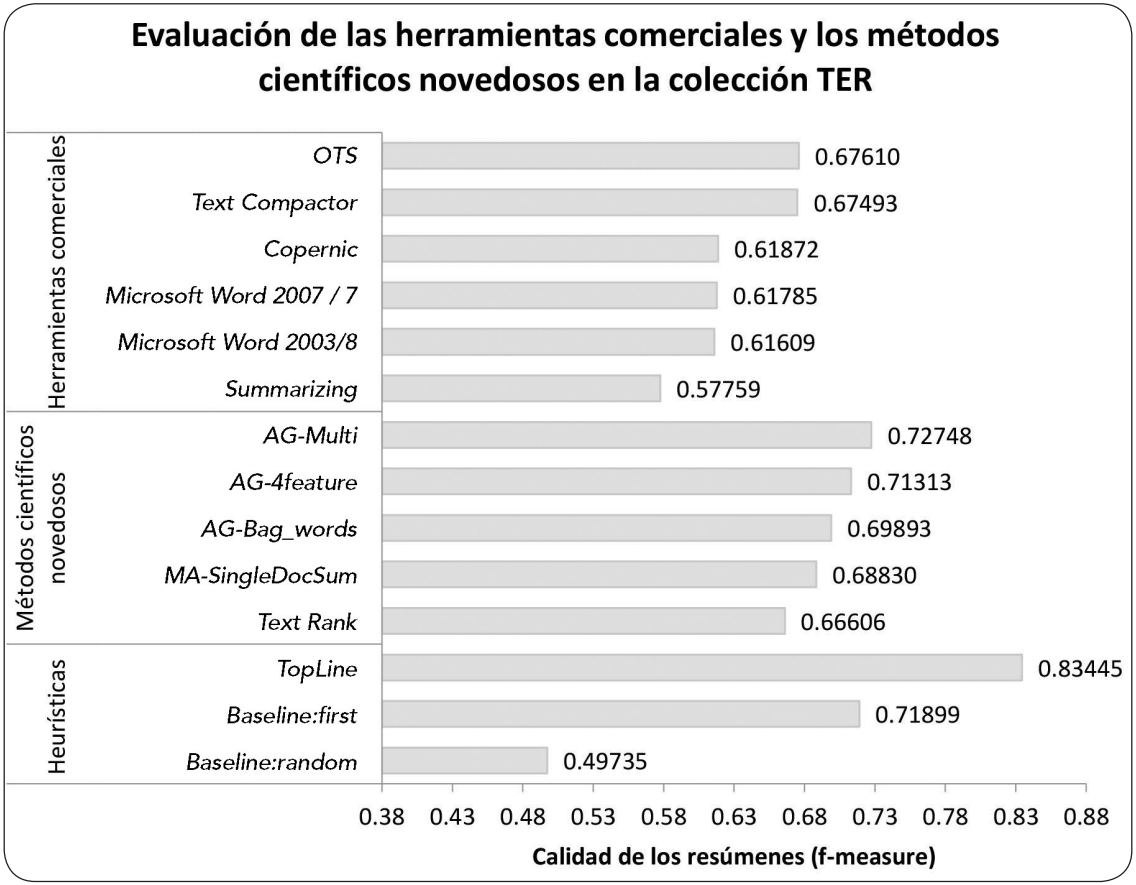
6.5 Resultados y análisis

Como se mencionó anteriormente, para el lenguaje español los recursos han sido limitados. Sin embargo, gracias al *corpus* TER, ahora se pueden probar los métodos científicos novedosos disponibles; se han probado aquellos que han tenido mejores resultados para el inglés.

La primera heurística por superar por los métodos y herramientas en la GART es *baseline:random* y, como se puede ver en la **gráfica 6.14**, todos los métodos y herramientas la superan. *Baseline:first,* especialmente para el español,

es de las más altas. Hasta el momento, sólo un método científico novedoso rebasa a esta heurística.

Gráfica 6.14 Evaluación de las herramientas comerciales y los métodos científicos novedosos en la colección TER



Para el *Test de Turing* realizado para español, se consideró a la herramienta comercial *Microsoft Office Word* 2007 y al método de Matias (2016). El resumen 1 corresponde a *Microsoft Office Word* y el resumen 6, al método de Matias (2016) (sección 1.2).

Capítulo VII

Generación automática de resúmenes para el lenguaje portugués

Este capítulo está dedicado a la presentación puntualizada del estudio de la tarea de GART para el lenguaje portugués. Se describe el *corpus TeMário*, el cual es utilizado para realizar las pruebas en portugués. También se muestran los resultados de las principales heurísticas, las herramientas comerciales y los métodos científicos novedosos probados al respecto. Finalmente, se brinda una comparación general de dichos elementos trabajados con el *corpus TeMário*.

El portugués es la sexta lengua hablada a nivel mundial, con doscientos millones de hablantes nativos y en doscientos dos países (**tabla 7.1**).

Tabla 7.1 Principales lenguas habladas en el mundo

No.	Lenguaje	Países	Hablantes
1	Chino	35	1302
2	Español	21	427
3	Inglés	106	339
4	Árabe	58	267
5	Hindi	4	260
6	Portugués	12	202
7	Bengalí	4	189
8	Ruso	17	171
9	Japonés	2	128
10	Lahnda	8	117
11	Javanés	3	84.3
12	Coreano	7	77.3
13	Alemán	26	76.9
14	Francés	53	75.9
15	Télugu	2	74.2
16	Marathi	1	71.4
17	Turco	8	71.4
18	Urdu	6	68.6
19	Vietnamita	3	68
20	Tamil	7	67.8
21	Italiano	13	63.4
22	Persa	30	61

Actualmente, el portugués ocupa el quinto lugar en el uso de Internet. Además de presentar un crecimiento significativo en el uso de las redes sociales, como Facebook y Twitter.

Tabla 7.2 Lenguas más usadas en Internet[27]

No.	Lenguaje	Usuarios de Internet
1	Inglés	1052
2	Chino	804
3	Español	337
4	Árabe	219
5	Portugués	169
6	Hindi	168
7	Francés	134
8	Japonés	118
9	Ruso	109
10	Alemán	92
	Otra	950

Cuando se desarrollan métodos independientes del lenguaje para la GART y se muestran los resultados para inglés y español, se plantea el problema de ¿Cómo funcionarían estos métodos para otros lenguajes? Por ejemplo, para el portugués. Como hemos visto en los capítulos anteriores, la mayoría de las investigaciones y conjuntos de datos se han llevado a cabo para el inglés. Como ya se sabe, la investigación en la tarea de GART tiene aproximadamente sesenta años. Para el portugués, las investigaciones surgen a principios del siglo XXI, de manera formal en el 2003, con el trabajo de Pardo (2003).

En este capítulo presentamos una investigación de los principales métodos científicos novedosos y herramientas comerciales para un *corpus* en portugués, implementándolos de manera independiente del lenguaje con su posterior evaluación.

[27]Según un estudio que revela las lenguas más usadas en internet: https://www.internetworldstats.com/stats7.htm

Al final, se demuestra qué métodos logran los mejores resultados del estado de arte a nivel internacional. Se hace un esfuerzo y la descripción del entorno experimental y teórico para fomentar la investigación de los métodos y herramientas comerciales para portugués.

7.1 Conferencias, talleres y *corpus*

Existen recursos, proyectos y herramientas para portugués que están descritos en el portal del Centro Institucional de Lingüística Computacional; se puede consultar en la web (NILC, 2018).[28]

7.1.1 *Corpus* CSTNews

El *corpus* CSTNews (Aleixo *and* Pardo, 2008) se utiliza en la tarea de la generación automática de múltiples documentos para el portugués. Cada una de las colecciones se etiqueta en una de las siguientes categorías:

- Noticias diarias
- Noticias mundiales
- Deportes
- Economía
- Política
- Ciencias

7.1.2 *Corpus* CSTNews-Update

CSTNews-Update (Cardoso *et al*., 2011) es una configuración diferente del *corpus* CSTNews, que tiene cincuenta colecciones de texto con dos o tres relacionados y que fueron recopilados de las principales agencias de noticias en Brasil.

[28]The Interinstitutional Center for Computational Linguistics. Research and development projects in Computational Linguistics and Natural Language Processing. Página web http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources. [Fecha de consulta 21 de Mayo de 2018].

7.1.3 *Corpus* utilizado para la evaluación y comparación

En este libro se describe la experimentación realizada sobre el *corpus TeMário*, cuyo nombre está formado por siglas tomadas de las siguientes palabras en portugués "**TE**xtos com su**MÁRIO**s". El *corpus* contiene cien artículos periodísticos y un resumen para cada uno de ellos generado por el mismo escritor del artículo (Pardo *and* Rino, 2003).

El objetivo principal del *corpus* es comparar los resúmenes hechos por los sistemas automáticos con los resúmenes elaborados por el humano. Además, puede servir para otras tareas de GART automáticos como, por ejemplo, el análisis lingüístico de textos y resúmenes, la construcción y la formación de resúmenes automáticos y la evaluación de aquellos construidos por el experto con los producidos automáticamente por los sistemas.

El uso del *corpus* se puede extender a las áreas de detección de tópicos y recuperación de información. Actualmente, se desarrollan investigaciones sobre cómo el experto reconoce la información relevante en un texto para componer sus resúmenes, o la identificación de parámetros que indican los criterios para resumir a efecto de construir el modelado de sistemas computacionales (Pardo *and* Rino, 2003) y (Martins *et al*., 2001). El *corpus* se compone por noticias sobre diferentes temas. La longitud requerida para los resúmenes es de 25-30% del tamaño de su texto-fuente.

A continuación, en la **tabla 7.3** se muestra cómo está compuesto el *corpus TeMário*.

Tabla 7.3 Características generales del *corpus* TeMário

Periódicos	Secciones	Número de textos	Número de palabras	Promedio de palabras/texto
Folha de São Paulo	Especial	20	12340	617
	Mundo	20	13739	686
	Opinión	20	10438	521
Jornal do Brasil	Internacional	20	12098	604
	Política	20	12797	639
	Total	100	61412	
	Promedios generales		12282	613

7.2 Heurísticas

Las heurísticas se calculan para dar una referencia y comparación de los métodos científicos novedosos y las herramientas comerciales.

Una de las problemáticas presentadas en el lenguaje portugués es la flexibilidad en el rango que se puede considerar para generar los resúmenes, ya que se da la opción de 25 a 30% del tamaño del documento original como el tamaño del resumen. Además de la apertura que se presenta para la longitud de los resúmenes, también surge la problemática de que el *corpus* no está etiquetado ni separado por oraciones, lo que dificulta llegar a un acuerdo en las diferentes heurísticas. En las pruebas realizadas para el libro se utilizó una longitud a 30% y el *corpus*, separando por oraciones el texto.

7.2.1 *Baseline:random*

Para calcular esta heurística para portugués se toman las oraciones de manera aleatoria y se forma un resumen hasta llegar a 30% de cada documento, con el fin de cumplir los requerimientos del *corpus TeMário*. Para esta heurística el valor reportado por Matias (2016) es de: 0.4574.

7.2.2 *Baseline:first*

Como se había mencionado anteriormente, por la flexibilidad que tiene el *corpus* para determinar la longitud de los resúmenes hay trabajos en el estado del arte que presentan valores diferentes para esta heurística, por lo que se supone como causa probable a la longitud que se consideró para realizar las pruebas.

Para la heurística *baseline:firts*, con una longitud a 30%, el valor obtenido en Matias (2016) es de: 0.4846. Sin embargo, en el trabajo de Mihalcea (2005) el valor reportado es de: 0.4963, debido al preprocesamiento adicional realizado.

7.2.3 *Topline*

El cálculo de *Topline* se realizó en el trabajo de Rojas J. (2017) utilizando algoritmos genéticos. Se calcula por medio de dos enfoques: combinaciones de oraciones y de párrafos como partes del texto a combinar.

Los mejores parámetros del AG propuesto por Rojas J. (2017) se presentan en la **tabla 7.4**.

Tabla 7.4 Parámetros del AG para calcular el Topline del *corpus* TeMário

Experimento	Elite	Generaciones	Individuos	Selección		Cruza	Mutación	
				Tipo	P		Tipo	P
1	Si	30	150	Torneo	3	CX	Inserción	8

A continuación, en la **tabla 7.5** se muestran los resultados para la heurística *Topline* en el *corpus TeMário* para el lenguaje portugués.

Tabla 7.5 Resultados de Topline para TeMário

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.6450	0.6059	0.6235
ROUGE-2	0.3328	0.3141	0.3225
ROUGE-SU4	0.3274	0.3078	0.3166

7.3 Herramientas comerciales

Para la GART en portugués se utilizaron las siguientes herramientas comerciales (**tabla 7.6**). Para la realización de las evaluaciones se usó el *corpus TeMário* a una longitud de 30%.

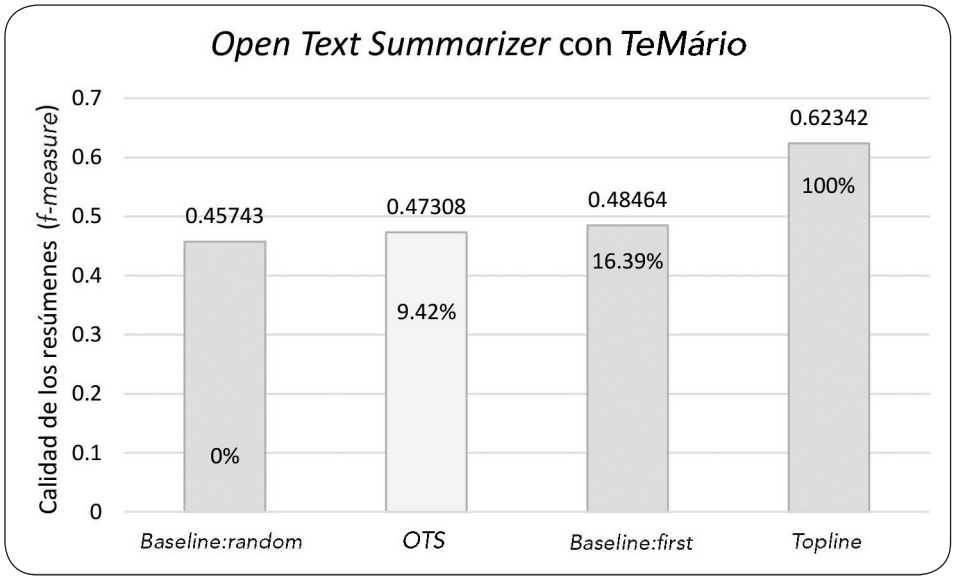
Tabla 7.6 Herramientas comerciales evaluadas para el lenguaje portugués

Herramienta	Tipo	Portugués
OTS	Línea	✓
Shvoong	Línea	✓
Total		2

7.3.1 TEXT *SUMMARIZER*

En esta herramienta se puede elegir el porcentaje que se requiere para el resumen. La longitud seleccionada fue de 30% para cada documento de la colección. A continuación, se muestran los resultados obtenidos en *TeMário* evaluados con *ROUGE*.

Gráfica 7.1 Resultados de *Open Text Summarizer* usando *TeMário* en comparación con las diferentes heurísticas

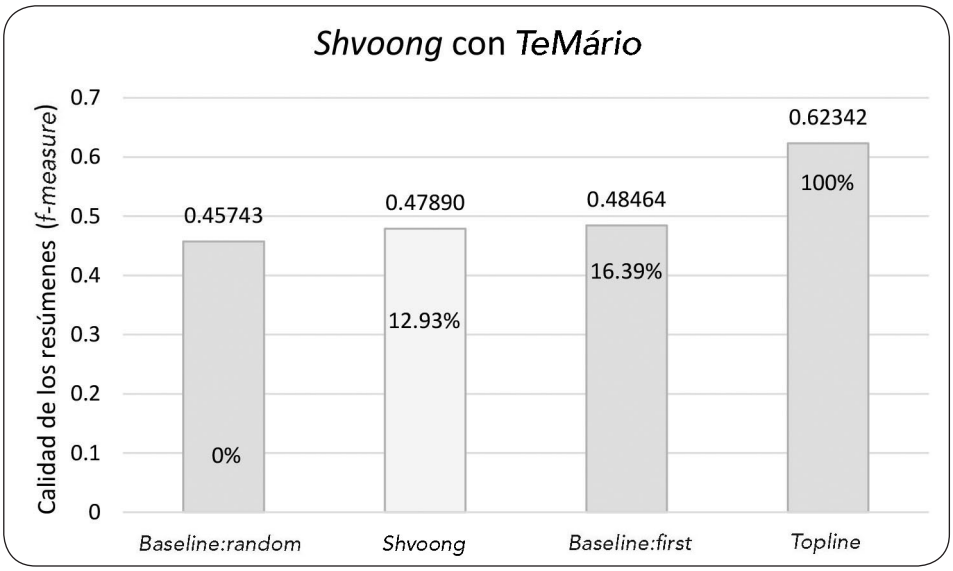


En la **gráfica 7.1** se muestran los resultados obtenidos con *Open Text Summarizer* usando el *corpus TeMário*. Como se puede observar, esta herramienta no supera a *baseline:first*. Si se considera a *baseline:random* como la peor forma de realizar un resumen y a *Topline* como el valor máximo que se puede obtener, entonces *Pertinence Summarizer* da 9.43%.

7.3.2 *SHVOONG*

A continuación, se muestran los resultados obtenidos por la herramienta *Shvoong* en la colección *TeMário* evaluados con *ROUGE*.

Gráfica 7.2 Resultados de *Shvoong* usando *TeMário* en comparación con las diferentes heurísticas



En la **gráfica 7.2** se muestran los resultados obtenidos con *Shvoong* usando el *corpus TeMário*. Como se puede observar, esta herramienta no supera a *baseline:first*. Sin embargo, *Shvoong* obtiene 12.92% de avance en la tarea de GART respecto de *baseline:random* y *Topline.*

Dentro de los proyectos de NILC se han desarrollado varios sistemas para generación y evaluación de los métodos de GART de *GistSumm* (Pardo *et al*., 2003), NeuralSumm (Pardo *et al*., 2003b), DMSumm (Pardo, 2002), *SuPor* (Modolo, 2003) y UNLSumm (Martins, 2002). Entre éstos, el más importante es el propuesto por Pardo, *GistSumm*, desarrollado en el año 2003; se ha mantenido actualizado y disponible de manera gratuita en la web hasta la fecha (Pardo *et. al*., 2003a).

A continuación, se describen algunas herramientas disponibles para el lenguaje portugués que no se probaron en este trabajo, pero que están disponibles para utilizarse, entre ellas están: Rsumm, ViSum y NILC-WISE.

RSumm[29]

Es una herramienta en línea que resuelve la tarea de generar automáticamente los resúmenes a partir de una búsqueda específica en Google News. Agrupa toda la información recabada por el usuario, con la posibilidad de agregar más noticias o restar algunas del resumen, considerando las necesidades del usuario.

ViSum[30]

Es un sistema para visualización de los resúmenes hechos para la tarea de la generación automática de múltiples documentos.

NILC-WISE[31]

Es una aplicación con una interfaz web desarrollada en NILC (Centro Interinstitucional de Lingüística Computacional) con el fin de proporcionar una forma y un repositorio para que los investigadores evalúen sus resúmenes automáticos.

7.4 Métodos científicos novedosos

En esta sección se presentan los métodos científicos novedosos que trabajan para la tarea de generación automática de resúmenes en portugués. Los dos primeros de la **tabla 7.7** no se encuentran disponibles, por lo cual no se pudieron replicar para probar el *corpus TeMário*.

[29]La extensión de sumarización en línea RSumm News consultado en la página web: http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/RSumm%20News%20-%20Tutorial/home.html

[30]Consultado el 15 de mayo de 2018, desde: http://conteudo.icmc.usp.br/pessoas/taspardo/

[31]NILC-WISE – Es una interfaz web para la evaluación de resúmenes. http://nilc.icmc.usp.br/nilcwise/login

Tabla 7.7 Métodos científicos novedosos evaluados para el lenguaje portugués

Herramienta	Portugués
SuPor	
SABio	
GistSumm	✓
AG-Multi	✓
TextRank	✓
Total	3

7.4.1 *SuPor*

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

7.4.2 *SABio*

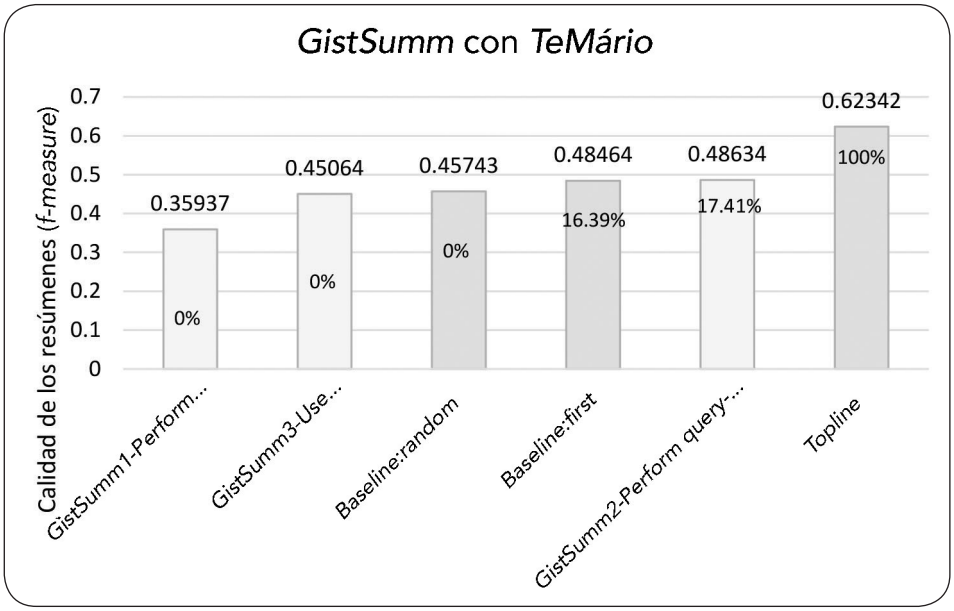
Automatic *Summarizer* for the Portuguese language with more Biologically plausible connectionist architecture and learning, o *SaBio*, está basado en una red neuronal entrenada con noticias del *corpus TeMário* (Orrú *et al*., 2006). Esta aplicación considera las siguientes características: tamaño de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

7.4.3 *GistSumm*

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

La longitud seleccionada fue de 30% para cada documento de la colección. A continuación, se muestran los resultados obtenidos por esta herramienta en *TeMário* evaluados con *ROUGE*.

Gráfica 7.3 Resultados de *GistSumm* usando *TeMário* en comparación con las diferentes heurísticas



El método *GistSumm* tiene tres configuraciones (GistSumm1-PerformIntrasetial summarization; GistSumm2-Performquery-based summarization y GistSumm3-Use averagekeywords ranking method) de las cuales GistSumm1 y GistSumm3

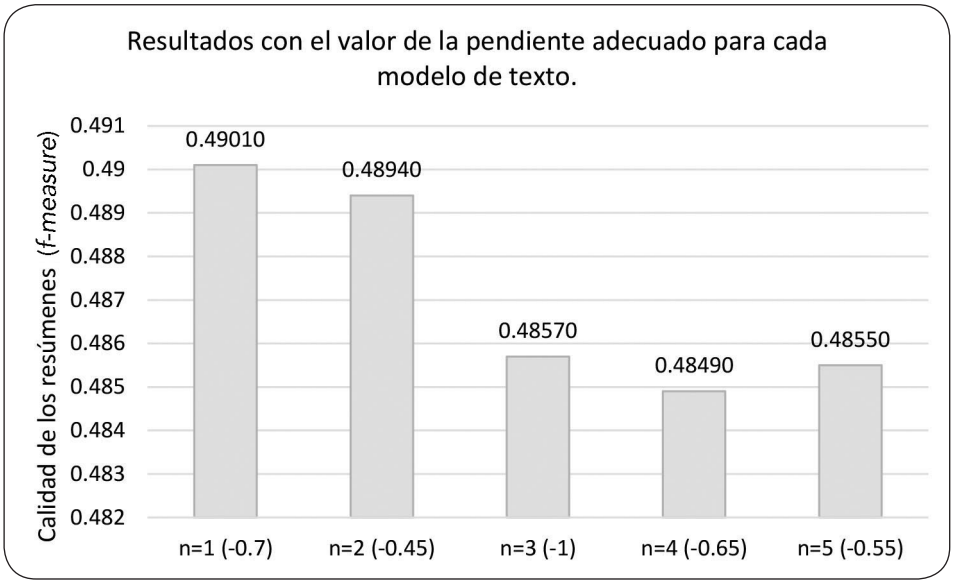
no superan a la heurística *baseline:random.* Sin embargo, la configuración para GistSumm3 supera tanto a *baseline:random*, como a *baseline:first*.

7.4.4 *AG-Multi*

En la sección 5.5.5 se describe el método propuesto por Matias (2016). Las pruebas realizadas con éste para el *corpus TeMário* se hicieron a 30% de longitud para los resúmenes. A continuación, se muestran los resultados obtenidos por esta herramienta en *TeMário* evaluados con *ROUGE*.

En la **gráfica 7.4** se observan los resultados que dio la aplicación del modelo de texto *n-gramas* ().

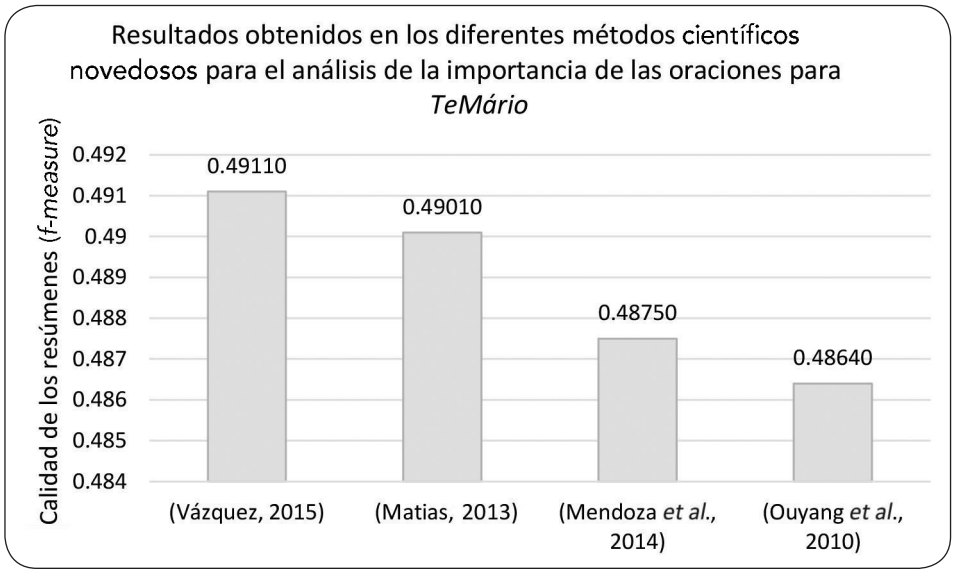
Gráfica 7.4 Resultados con el valor de la pendiente adecuada para cada modelo de texto para *TeMário*



El mejor modelo de texto para la colección *TeMário* es bolsa de palabras (*n-gramas* con).

En la **gráfica 7.5** se muestran los resultados del análisis realizado a los métodos científicos novedosos para determinar la mejor forma de calcular la característica de la posición de las oraciones.

Gráfica 7.5 Resultados obtenidos en los diferentes métodos científicos novedosos para el análisis de la importancia de las oraciones para *TeMário*



Como se puede observar, la fórmula propuesta en Vázquez (2015) es con la que se obtienen mejores resultados para el *corpus TeMário*.

Si se consideran las dos características (frecuencia de términos y posición de las oraciones) que utiliza el método de Matias (2016) como función de aptitud, para *TeMário* las dos tienen la misma importancia.

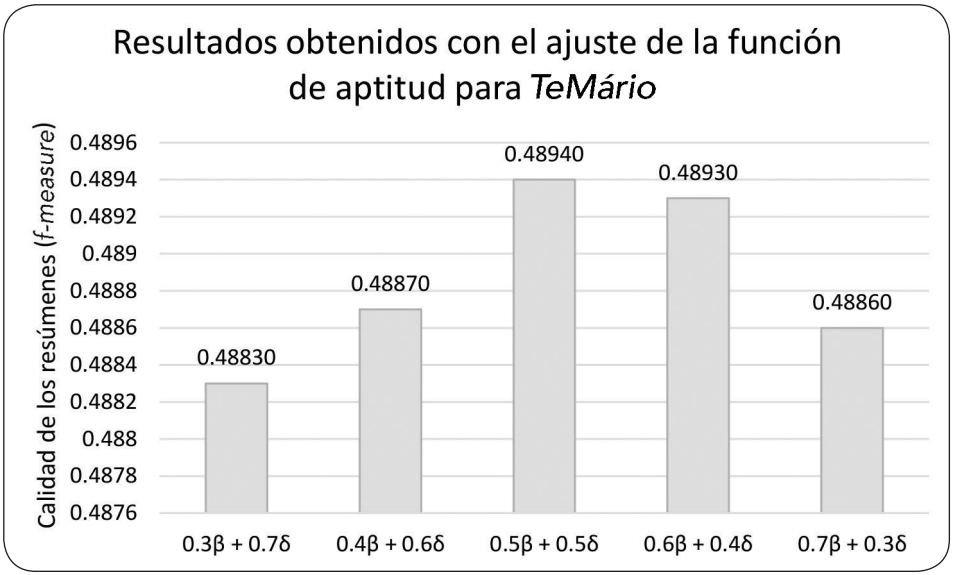
Considerando los resultados obtenidos en las **gráficas 7.5** y **7.6** se puede determinar que las características utilizadas en la función de aptitud tienen la misma importancia, por lo que el resultado que se elige para la comparación con las diferentes heurísticas es el de la **gráfica 7.**5.

Como se puede observar, el método *AG-Multi* supera las dos heurísticas *baseline:random* y *baseline:first* (**gráfica 7.7**).

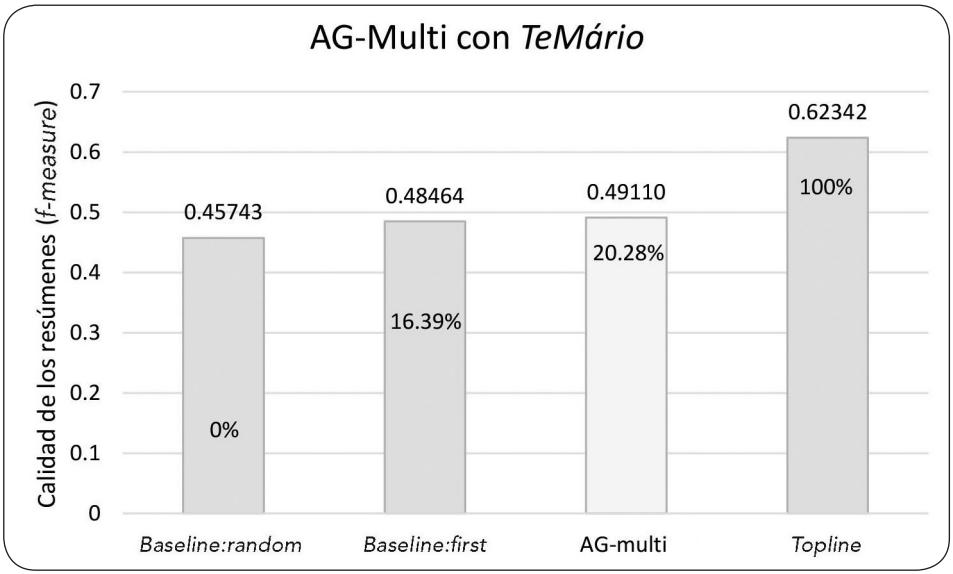
7.4.5 *TextRank*

Es un método basado en grafos propuesto por Mihalcea (2004), el cual usa los siguientes algoritmos.

Gráfica 7.6 Resultados obtenidos con el ajuste de la función de aptitud



Gráfica 7.7 Resultados de *AG-Multi* usando *TeMário* en comparación con las diferentes heurísticas



PageRank

Es uno de los algoritmos de clasificación más popular; fue diseñado como un método para el análisis de enlaces web. A diferencia de otros algoritmos de clasificación basados en grafos, integra la entrada y la salida de enlaces en un solo modelo y, por lo tanto, produce un solo conjunto de resultados (*Brin and Page*, 2012). El método *PageRankW* añade un peso entre los vértices del grafo formado. De esta manera, la clasificación de los algoritmos es adaptado para incluir aristas ponderadas.

HITS

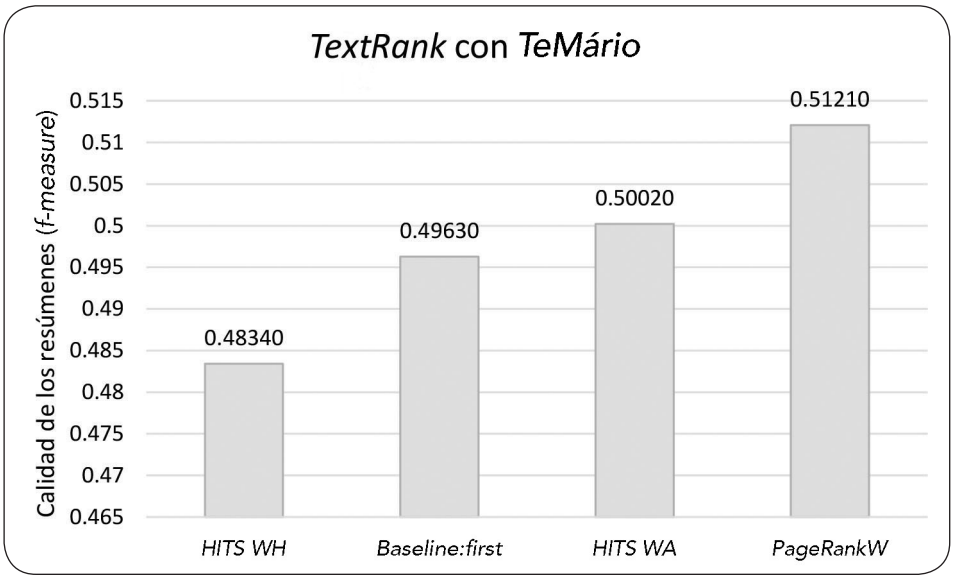
HITS (Hiperlinked Induced Topic Search) realiza una búsqueda de temas de enlaces inducidos. Es un algoritmo iterativo diseñado para la clasificación de páginas web de acuerdo con su grado de "autoridad". Además, hace una distinción entre las *"authorities"* (páginas con un gran número de enlaces entrantes) y *"hubs"* (páginas con un gran número de enlaces salientes) (Kleinberg, 1999). Para cada vértice, HITS coloca dos puntuaciones: una de "*authority"* y una de "*hub*". El método HITSW añade un peso entre los vértices del grafo formado. De esta manera, la clasificación de los algoritmos es adaptado para incluir aristas ponderadas.

Los algoritmos *PageRankW* e *HITS* fueron probados en el *corpus TeMário* por Mihalcea (2005) quien, en su trabajo, presenta que para la heurística *baseline:first* el valor que se considera es de 0.4963. Sin embargo, no menciona los parámetros que usó para llegar al resultado, por lo que en Matias (2016) se intenta replicar el experimento pero no se logra alcanzar el valor presentado por Mihalcea. Entonces los resultados de las dos autoras no pueden ser comparados, por lo que se presentan por separado. En la **gráfica 7.8** se muestran los datos obtenidos por Mihalcea (2005), tanto en la heurística *baseline:first* como en *PageRankW* e *HITS*.

7.5 Resultados y análisis

En la **gráfica 7.9** se muestran los resultados de los experimentos con el lenguaje portugués a través de las herramientas comerciales y los métodos científicos

Gráfica 7.8 Resultados de *TextRank* usando las diferentes configuraciones de PageRank e *HITS* en *TeMário*



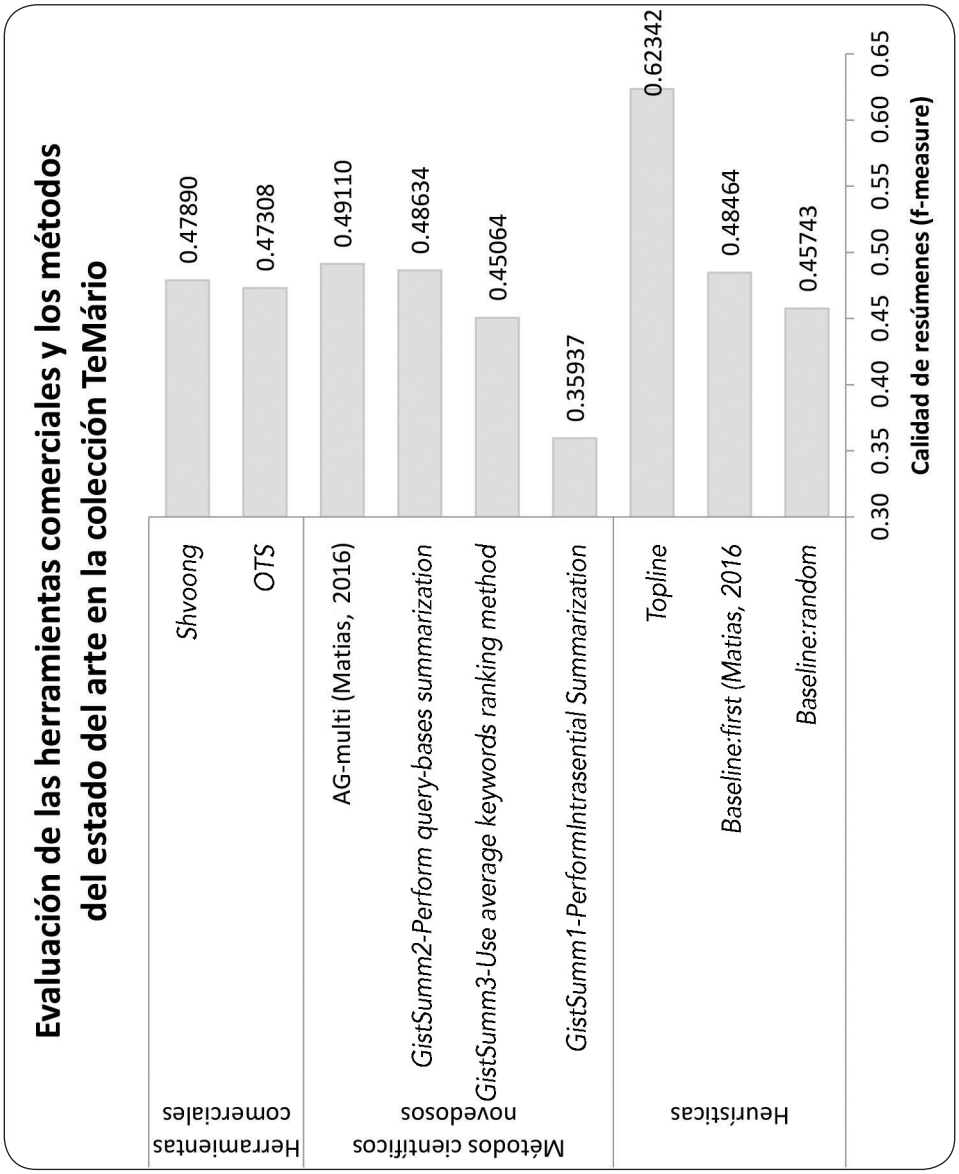
novedosos. Los experimentos se realizaron con la colección de documentos *TeMário* y se evaluaron con *ROUGE*. Para la colección *TeMário*, la extensión que se puede dar a los resúmenes va de 25 a 30%. Los resultados de los experimentos mostrados se hicieron a una extensión de 30%.

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

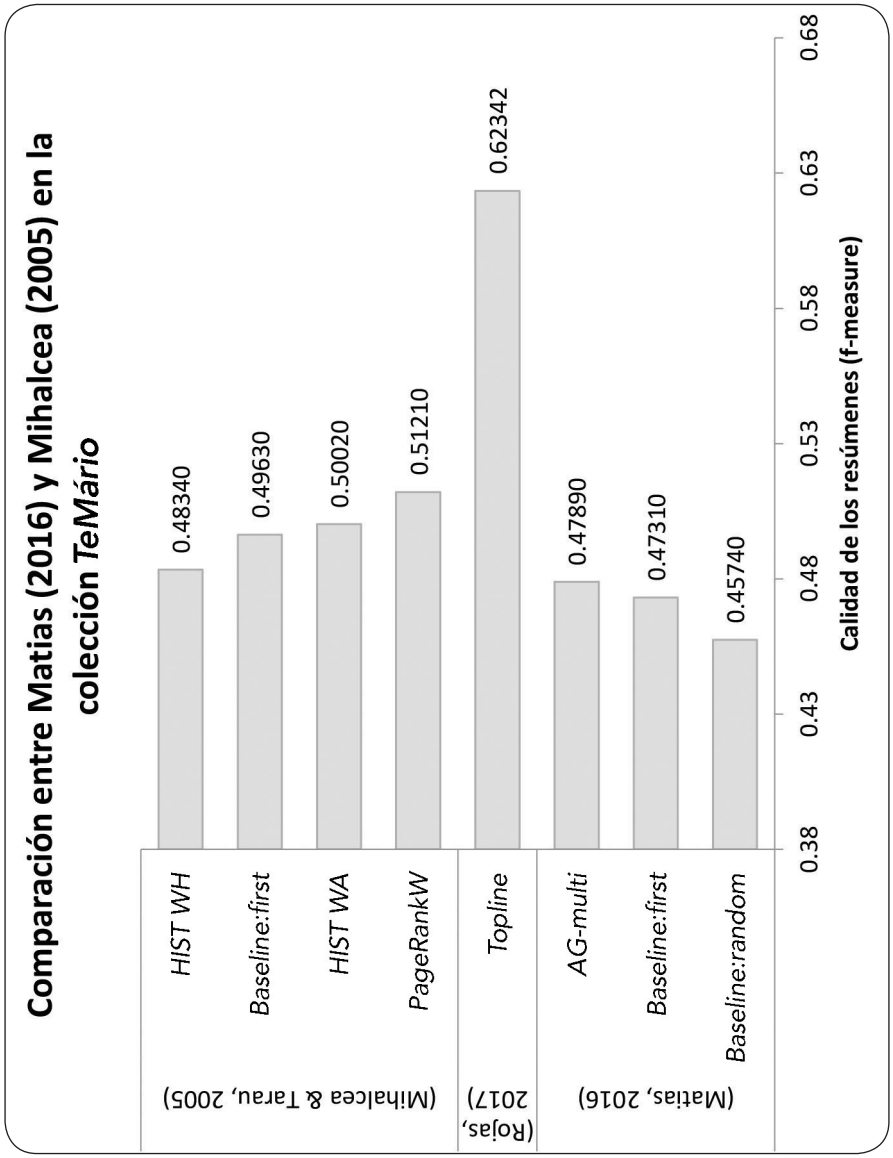
Como se puede observar en la **gráfica 7.10** el método de Matias (2016) supera a todas las herramientas comerciales en línea y obtiene los mejores resultados en el estado del arte.

En el estado del arte se encuentra el trabajo de Mihalcea y Tarau (2005), en donde se prueba la colección *TeMário*. Los resultados del trabajo de Mihalcea superan a los obtenidos en éste. Sin embargo, surge una problemática al comparar debido a que el rango de extensión que se usa para esta colección no está bien definido, por lo que para verificar la extensión utilizada en Mihalcea se obtuvo el resultado de *baseline* para *TeMário* en el rango establecido por la colección (25%

Gráfica 7.9 Resultados obtenidos con la colección en el lenguaje portugués con los métodos científicos novedosos y las herramientas comerciales



Gráfica 7.10 Comparación entre el trabajo actual y el de Mihalcea & Tarau (2005)



a 30% de la extensión del documento). Pero, habiendo probado con todas las extensiones posibles de 25% a 30%, no se llegó al resultado de *baseline* arrojado por Mihalcea. Por lo que se piensa que la problemática está en la forma en que se consideran las oraciones.

En la **gráfica 7.10** se muestra una comparativa entre los resultados de Mihalcea y Tarau (2005) y los del método propuestos en este trabajo. *Baseline* en el trabajo actual se calculó a 30%.

Como se puede observar en la **gráfica 7.10**, los resultados de la heurística *baseline* son diferentes. También, tanto el método de Matias (2016) como el de Mihalcea (2005) superan su *baseline* propuesto.

Capítulo VIII

Generación automática de resúmenes para el lenguaje ruso

Este capítulo está dedicado a la presentación puntualizada del estudio de la tarea de GART para el lenguaje ruso. Se describe el *corpus TEXTRUSS*, el cual es utilizado para realizar las pruebas en este lenguaje. También se muestran los resultados de las principales heurísticas, las herramientas comerciales y los métodos científicos novedosos empleados al respecto. Finalmente, se brinda una comparación general de dichos elementos probados con el *corpus TEXTRUSS*.

El ruso es una lengua indoeuropea hablada por más de ciento setenta y un millones de personas en forma nativa (**tabla 8.1**). Ocupa el octavo lugar de las lenguas más habladas a nivel mundial.

Tabla 8.1 Principales lenguas habladas en el mundo

No.	Lenguaje, origen	Países	Hablantes
1	Chino	35	1302
2	Español	21	427
3	Inglés	106	339
4	Árabe	58	267
5	Hindi	4	260
6	Portugués	12	202
7	Bengalí	4	189
8	Ruso	17	171
9	Japonés	2	128
10	Lahnda	8	117
11	Javanés	3	84.3
12	Coreano	7	77.3
13	Alemán	26	76.9
14	Francés	53	75.9
15	Télugu	2	74.2
16	Marathi	1	71.4
17	Turco	8	71.4
18	Urdu	6	68.6
19	Vietnamita	3	68
20	Tamil	7	67.8
21	Italiano	13	63.4
22	Persa	30	61

Pero, según el número de usuarios de Internet, Rusia ocupa el noveno lugar (**tabla 8.2**). Sin embargo, actualmente se registra un crecimiento respecto del uso

de Internet que lo posiciona en segundo lugar (Europe Internet Stats - Population Statistics, 2017).

Tabla 8.2 Lenguas más usadas en Internet[32]

No.	Lenguaje, origen	Usuarios de Internet
1	Inglés	1052
2	Chino	804
3	Español	337
4	Árabe	219
5	Portugués	169
6	Hindi	168
7	Francés	134
8	Japonés	118
9	Ruso	109
10	Alemán	92
	Otra	950

La GART para el lenguaje ruso no ha reportado en su estado de arte grandes avances durante los sesenta años de investigación. Por lo que el ruso es un campo de oportunidad para su estudio.

8.1 Conferencias, talleres y *corpus*

Para el ruso no existen conferencias o talleres dedicados a la tarea de GART. Sin embargo debido a la importancia que tiene y al crecimiento en usuarios de Internet, han surgido trabajos como los de Rojas (2016) y Hernández (2018), donde se hace un estudio sobre los métodos científicos novedosos y las herramientas comerciales para la producción de resúmenes. A continuación, se describe el *corpus* utilizado para los resúmenes generados en el lenguaje ruso.

[32]Según un estudio que revela las lenguas más usadas en internet: https://www.internetworldstats.com/stats7.htm

8.1.1 *Corpus* utilizado para la evaluación y comparación

TEXTRUSS está compuesto por artículos de noticias con sus resúmenes correspondientes, los cuales fueron conformados por personas expertas en el lenguaje ruso que, para el caso de este *corpus*, fueron los mismos periodistas que escribieron la noticia los que seleccionaron las oraciones más importantes.

Las noticias fueron descargadas del portal de noticias gazeta.ru ("Главные новости - Газета.Ru," 2015). El *corpus* es de diferentes dominios y contiene once categorías organizadas de la siguiente manera (se indica el nombre en ruso y su traducción al español):

1. ПОЛИТИКА (Política)
2. БИЗНЕС (Negocios)
3. ОБЩЕСТВО (Compañía)
4. МНЕНИЯ (Críticas)
5. КУЛЬТУРА (Cultura)
6. НАУКА (Ciencia)
7. ТЕХНОЛОГИИ (Tecnología)
8. НЕДВИЖИМОСТЬ (Bienes Inmuebles)
9. АВТО (Auto)
10. СТИЛЬ ЖИЗНИ (Estilo de vida)
11. СПОРТ (Deportes)

Cada categoría contiene veintidós artículos; de esta manera, en total tiene doscientos cuarenta y dos artículos. Los originales se llaman "textos-fuente".

Las partes de la estructura de cada artículo son las siguientes (**figura 8.1**):

8.1.2 Transliteración al idioma ruso

Para hacer uso del método científico novedoso presentado en el libro, hubo que transliterar los textos.

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, bajo el principio de letra por letra (Orozco, 1989). En la **figura 8.2** se encuentra una lista de letras en ruso transliteradas a letras latinas.

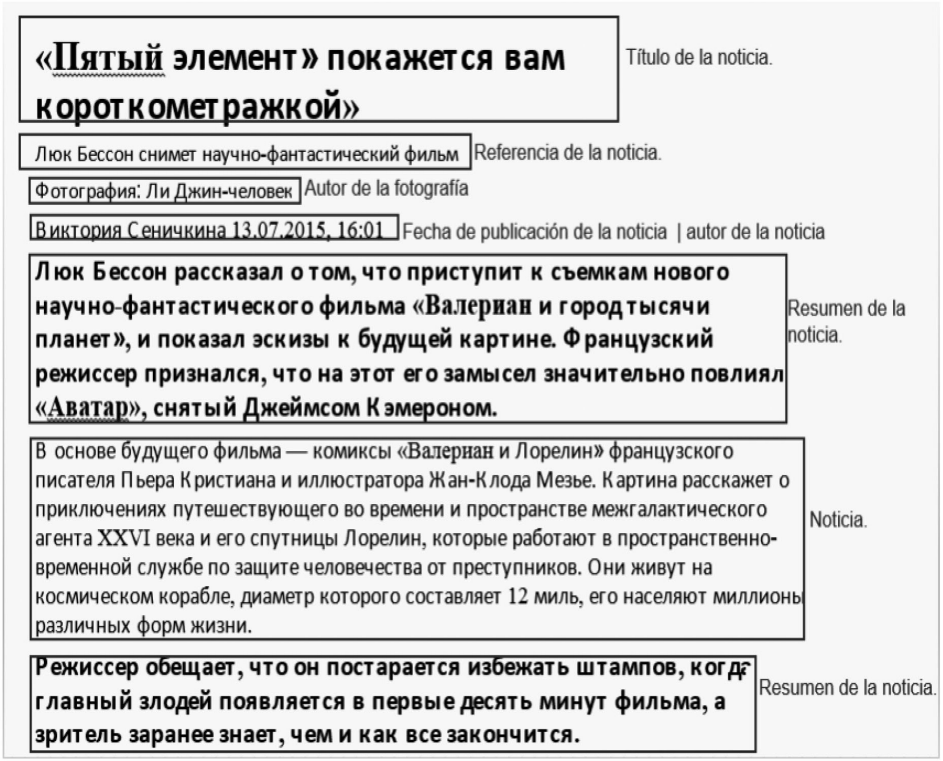


Figura 8.1 Ejemplo de noticia en el *corpus TEXTRUSS*

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 8.2 Transliteración de letras cirílicas a letras latinas ("Traslit," 2016)

8.2 Heurísticas

Para realizar el cálculo de las heurísticas, las herramientas comerciales y los métodos científicos novedosos, se utiliza el *corpus TEXTRUSS* debido a que es el único disponible y especial para la tarea de GART en el lenguaje ruso.

8.2.1 *Baseline:random*

Se seleccionan aleatoriamente n oraciones del texto original (Ledeneva, 2008). En la **tabla 8.3** se presentan los resultados obtenidos por la heurística *baseline:random* para *TEXTRUSS*. Cabe mencionar que para esta heurística se hicieron diez corridas como garantía del resultado mostrado.

Tabla 8. 3 Resultados de *baseline:random* para *TEXTRUSS*

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.8977	0.8540	0.8734
ROUGE-2	0.6221	0.5918	0.6053
ROUGE-SU4	0.7789	0.7407	0.7577

Para *baseline:random* los resultados tienden a ser bajos debido a que las oraciones son seleccionadas de forma aleatoria. Para el estado del arte, *baseline:random* sirve como referencia del peor resultado obtenido.

8.2.2 *Baseline:first*

Se seleccionan 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).

La **tabla 8.4** muestra los resultados obtenidos con *baseline:first* para el *corpus TEXTRUSS*.

Tabla 8. 4 Resultados de *baseline:first* para *TEXTRUSS*

Medida	Recuerdo	Precisión	*F-measure*
ROUGE-1	0.9332	0.8703	0.8994
ROUGE-2	0.7440	0.6940	0.7171
ROUGE-SU4	0.8477	0.7901	0.8168

Como se puede observar los datos generados a partir de *baseline:first* son muy altos, lo que nos muestra que las primeras oraciones para este *corpus* de noticias son muy importantes.

8.2.3 *Topline*

Para *TEXTRUSS* el *Topline* alcanzado es de: 1.0000. Esto debido a que sólo se tiene un *gold standard* y es completamente extractivo. Por esta razón, un método o herramienta puede generar el mismo resumen que el *gold standard*, lo cual permite definir un *Topline* de 1.0000.

8.3 Herramientas comerciales

Para el lenguaje ruso se utilizó el *corpus TEXTRUSS* y se empleó la longitud a cien palabras, misma que fue determinada considerando las características de *TEXTRUSS*. Para las herramientas que tienen únicamente la opción para elegir el porcentaje se usa la fórmula 11, la cual ayudará a calcular la cifra para generar el resumen.

$$\frac{\textit{número de palabras deseadas}}{\textit{número de palabras totales en el documento}} * 100 \qquad (11)$$

A continuación, se describen las herramientas comerciales usadas para las pruebas en el lenguaje ruso.

8.3.1 *Microsoft Office Word Summarizer*

Para realizar las pruebas con la herramienta *Microsoft Office Word* se utilizó el *corpus TEXTRUSS* transliterado. Para garantizar la longitud de los resúmenes a cien palabras se utilizó la fórmula 11.

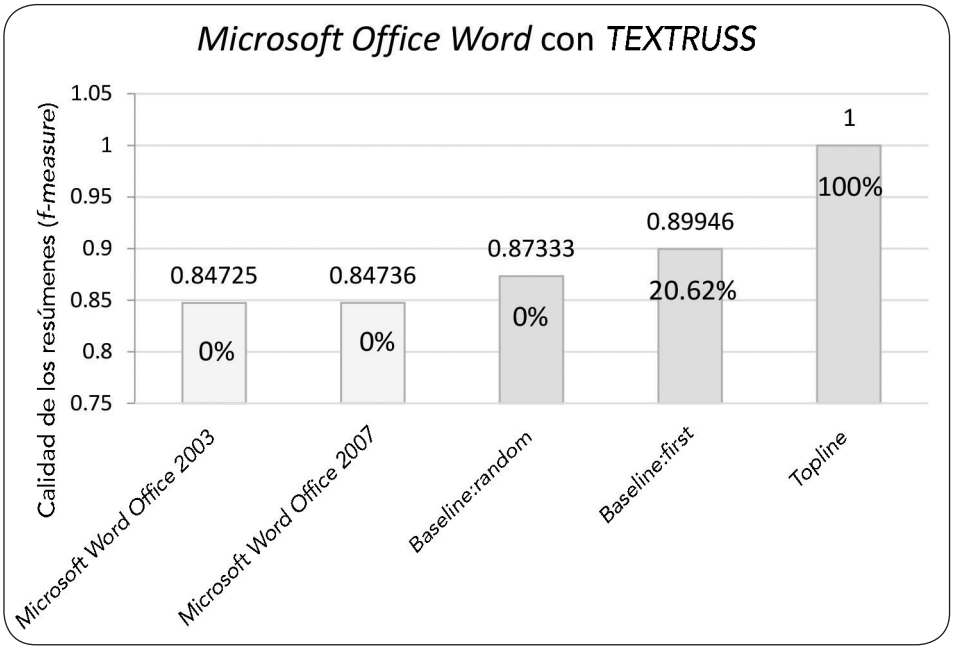
A continuación, en la **gráfica 8.1** se muestran los resultados con longitud a cien palabras.

La herramienta *Microsoft Office Word* no supera a la heurística *baseline:random*; sin embargo, esto pudo suceder por el número de pruebas que se realizan para determinar el valor de esta heurística.

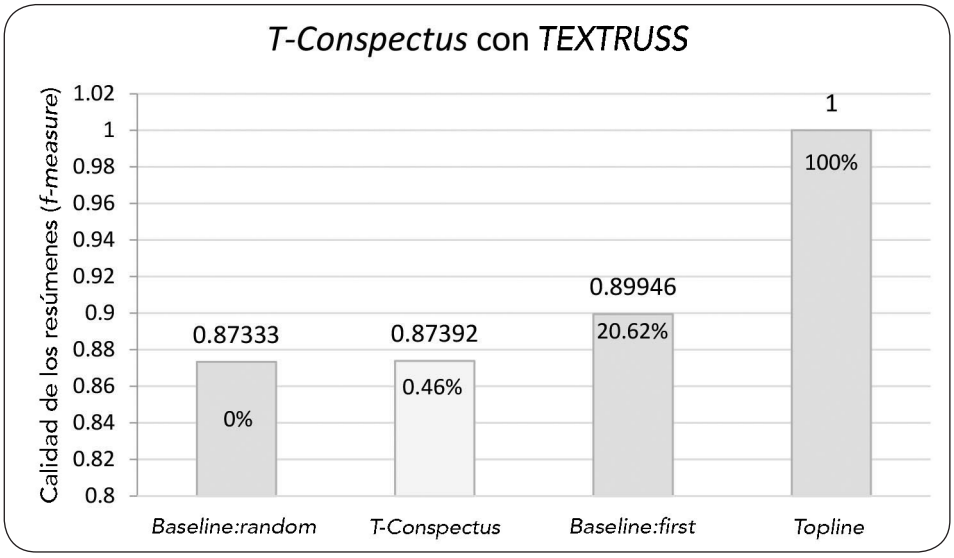
8.3.2 *T-Conspectus*

Para realizar las pruebas con la herramienta *T-Conspectus* se utilizó el *corpus TEXTRUSS* transliterado. Además, para generar los resúmenes se eligió el 20%

Gráfica 8.1 Resultados de *Microsoft Office Word* usando *TEXTRUSS* a cien palabras en comparación con las diferentes heurísticas



Gráfica 8.2 Resultados de *T-Conspectus* usando *TEXTRUSS* en comparación con las diferentes heurísticas



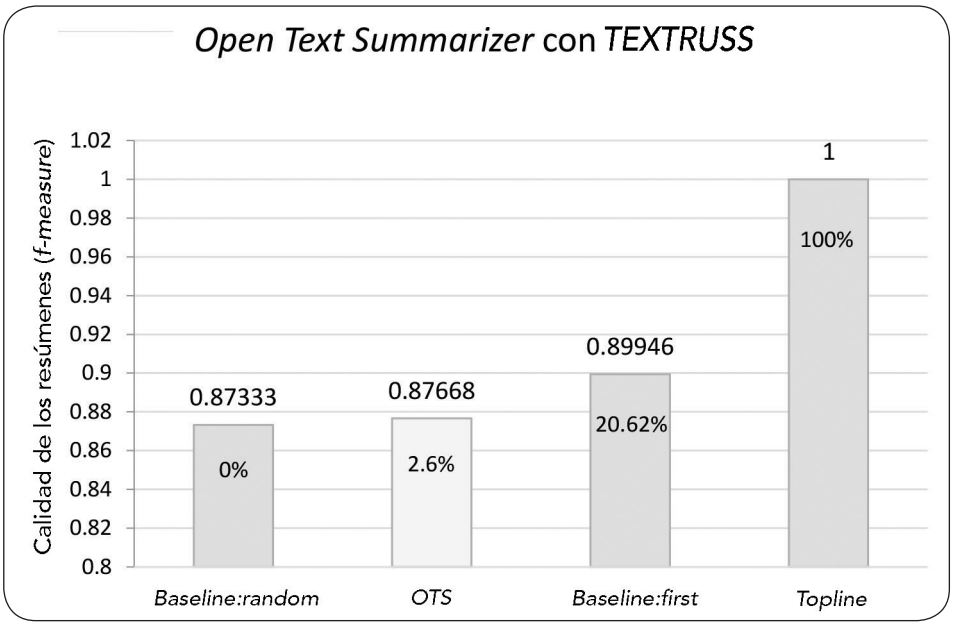
del tamaño del documento ya que es el porcentaje adecuado para conseguir textos con más de cien palabras en esta herramienta.

La herramienta *T-Conspectus* es una de las que obtiene mejores resultados para la GART en ruso, superando a *baseline:random* (**gráfica 8.2**).

8.3.3 *Open Text Summarizer (OTS)*

Para realizar las pruebas con la herramienta OTS se utilizó *TEXTRUSS* transliterado. Para garantizar la longitud de los resúmenes a cien palabras se utilizó la fórmula 11; con ello se logró determinar el porcentaje del resumen.

Gráfica 8.3 Resultados de *Open Text Summarizer* con *TEXTRUSS* en comparación con las diferentes heurísticas

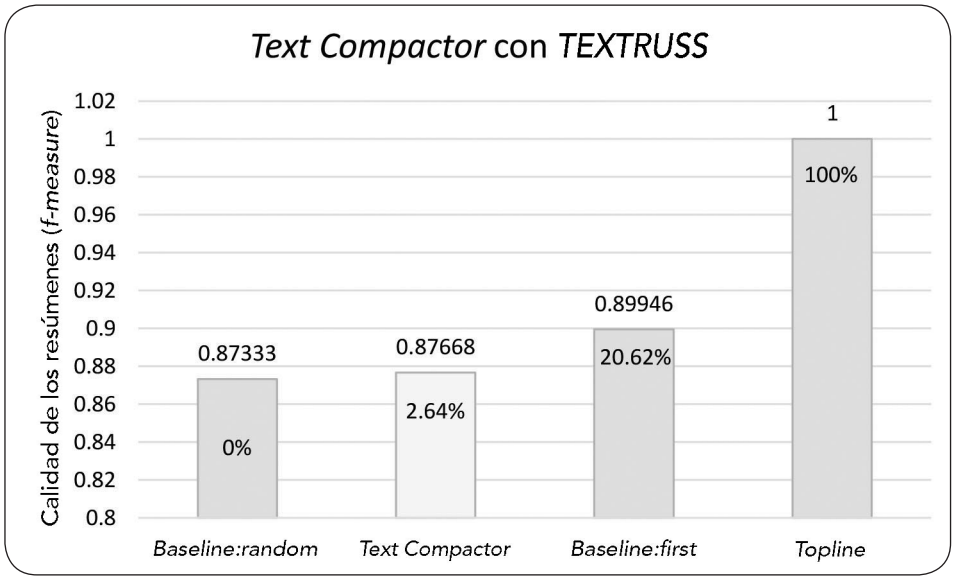


La herramienta OTS supera a la heurística *baseline:random*, mostrando un avance de 2.6% si se considera a *baseline:random* como 0 y a *Topline* como 100% (**gráfica 8.3**). Sin embargo, como se puede observar, los resultados obtenidos por OTS son muy bajos en comparación con el 20.62% que tiene la heurística *baseline:first.*

8.3.4 *Text Compactor*

Para realizar las pruebas con la herramienta *Text Compactor* se utilizó el *corpus TEXTRUSS* transliterado. Para garantizar la longitud de los resúmenes a cien palabras se utilizó la fórmula 11; con ello se logró determinar el porcentaje del resumen. En la **gráfica 8.4** se observa el resultado obtenido por esta herramienta: 2.64% con respecto a *baseline:random* y *Topline*.

Gráfica 8.4 Resultados de *Text Compactor* usando *TEXTRUSS* en comparación con las diferentes heurísticas

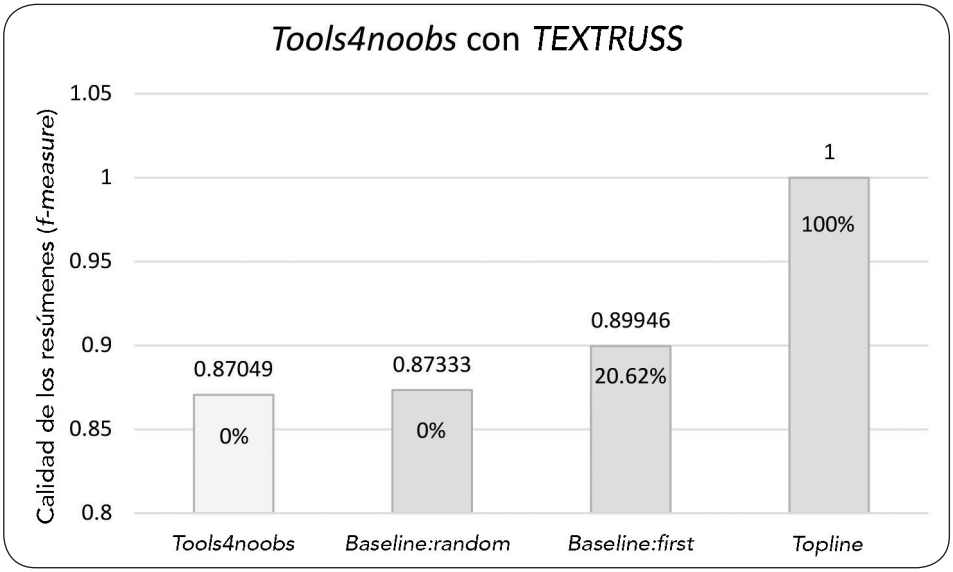


8.3.5 *Tools4noobs*

Para generar resúmenes con cien palabras en la herramienta *Tools4noobs* se aplicó la fórmula 12, esto debido a que, a diferencia de las otras herramientas, el umbral de *Tools4noobs* funciona de manera inversa.

$$\frac{\textit{número de palabras deseadas} * 100}{\textit{número de palabras totales en el documento} - 80} * -1 \qquad (12)$$

Gráfica 8.5 Resultados de *Tools4noobs* usando *TEXTRUSS* en comparación con las diferentes heurísticas



Con la herramienta *Tools4noobs*, a pesar de no superar a la heurística *baseline:random,* la diferencia es prácticamente nula, por lo que se puede considerar que obtienen el mismo valor (**gráfica 8.5**).

8.3.6 *Resumo*

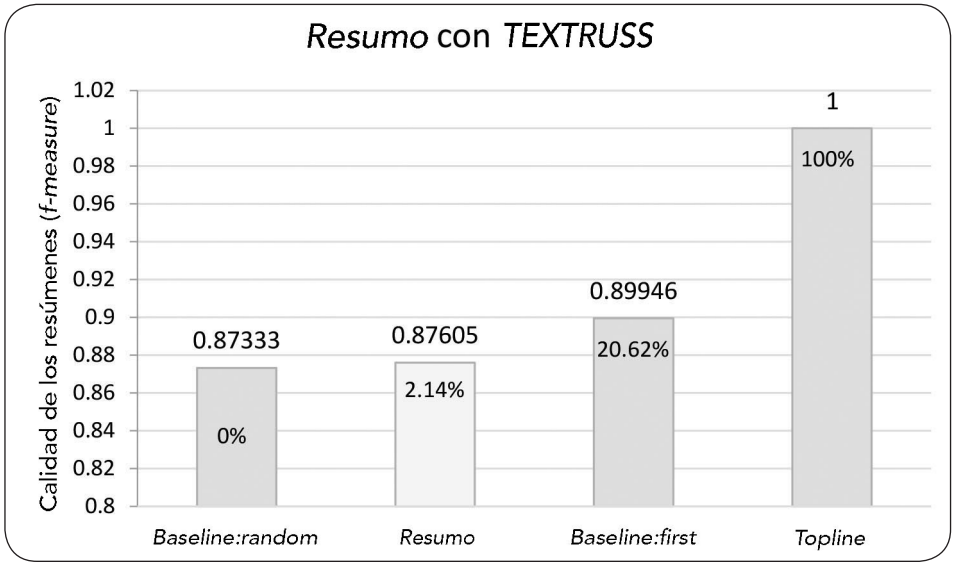
Para realizar las pruebas con la herramienta *Resumo* se utilizó el *corpus TEXTRUSS* en ruso. Para garantizar la longitud de los resúmenes a cien palabras se empleó la fórmula 1; con ello se logró determinar el porcentaje del texto.

Con la herramienta *Resumo* se puede trabajar con el texto en ruso y no es necesaria una transliteración; se tiene un avance de 2.14% considerando como referencia a las heurísticas *baseline:random* y *Topline* (**gráfica 8.6**).

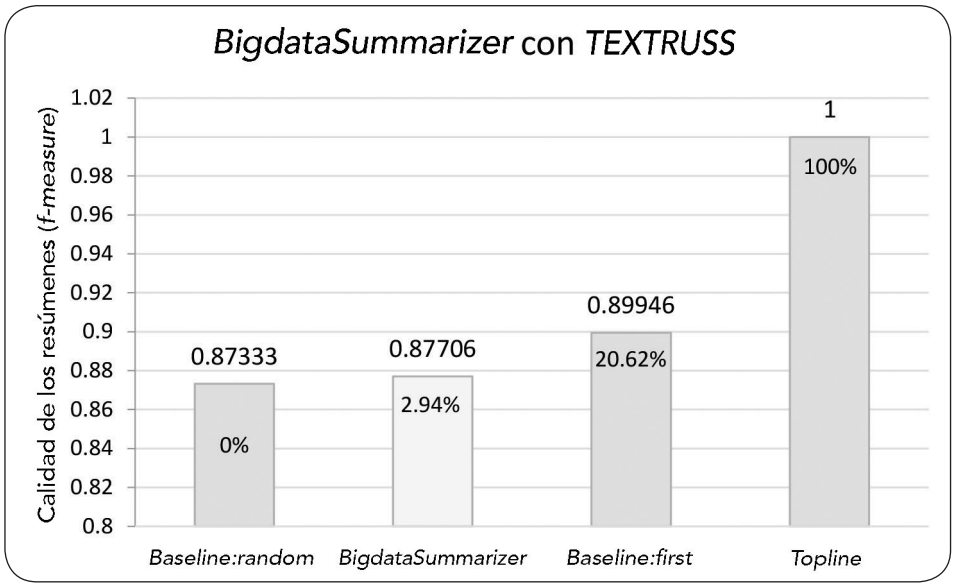
8.3.7 *BigdataSummarizer*

Para realizar las pruebas con la herramienta *BigdataSummarizer* se utilizó el *corpus TEXTRUSS* en ruso. Para garantizar la longitud de los resúmenes a cien palabras se empleó la fórmula 8; con ello se logró determinar el porcentaje del texto.

Gráfica 8.6 Resultados de *Resumo* usando *TEXTRUSS* en comparación con las diferentes heurísticas



Gráfica 8.7 Resultados de *BigdataSummarizer* usando *TEXTRUSS* en comparación con las diferentes heurísticas



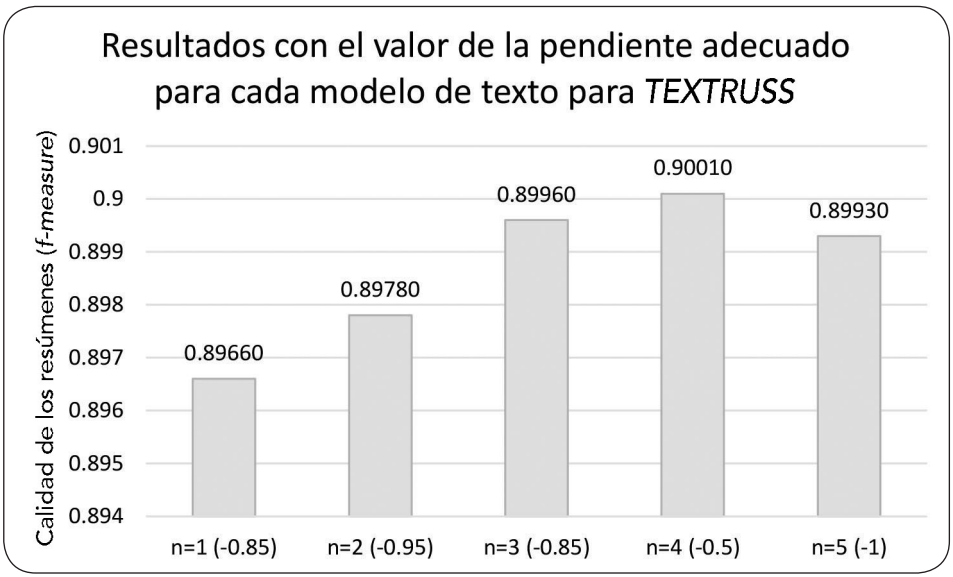
Al igual que para *Resumo*, para *BigdataSummarizer* el texto se puede usar en su forma normal sin transliteración. Para esta herramienta, se tiene un avance de 2.94% respecto de la heurística *baseline:random* (**gráfica 8.7**).

8.4 Métodos científicos novedosos

Se conoce y se tiene acceso al método de Matias (2016), el cual ha demostrado trabajar de forma correcta en diferentes lenguajes (inglés, español portugués), por lo que se utiliza para probar el ruso. La descripción está en la sección 5.5.5.

El método de Matias (2016) utiliza el modelo de texto n-gramas, por lo que para el lenguaje ruso se muestran los resultados obtenidos con (**gráfica 8.8**).

Gráfica 8.8 Resultado con el valor de la pendiente adecuada para cada modelo de texto

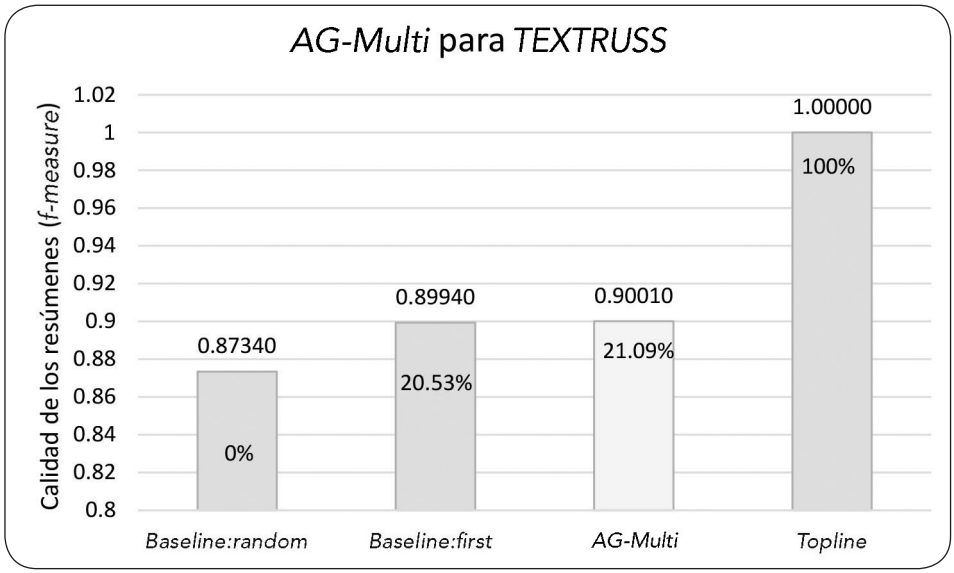


Como se puede observar, los mejores resultados se obtienen con $n = 4$ y un valor de la pendiente de $m = -0.5$. Cabe mencionar que aquellos se consiguieron con el *corpus* sin preprocesamiento. En Matias (2016) se probó un ajuste de importancia de las características para la función de aptitud. Sin embargo, para *TEXTRUSS*

se mantiene la importancia de las dos características utilizadas por el método (frecuencia de los términos y posición de las oraciones).

En la **gráfica 8.9** se muestra la comparación del resultado obtenido con AG-Matias y las diferentes heurísticas.

Gráfica 8.9 Resultados de *AG-Multi* usando *TEXTRUSS* en comparación con las diferentes heurísticas



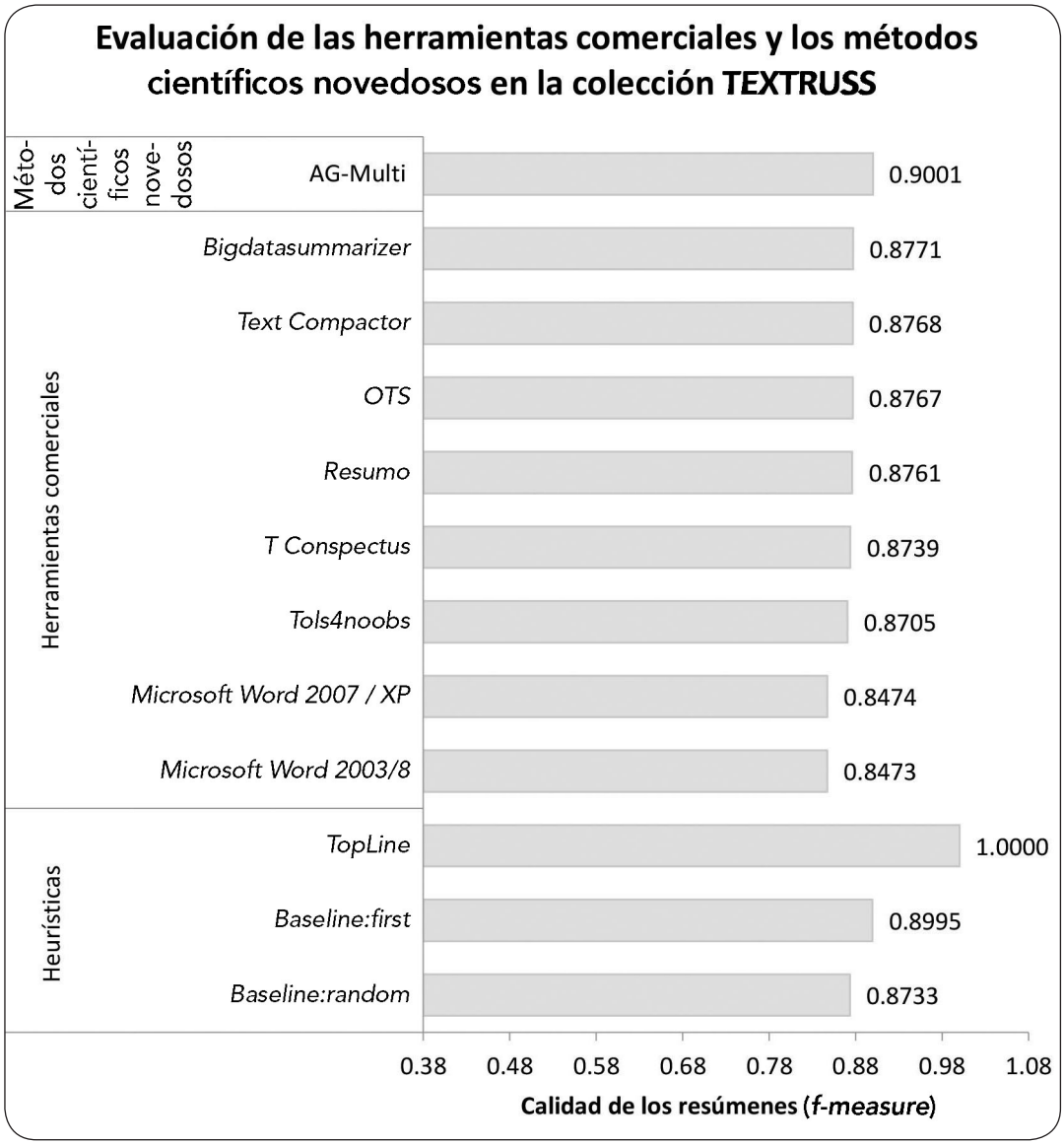
Como se puede observar, *AG-Multi* supera a *baseline:random* y por muy poco a la heurística *baseline:first*.

8.5 Resultados y análisis

Para el lenguaje ruso no se tienen investigaciones sobre la tarea de GART. En este libro se muestran los resultados de ocho herramientas comerciales y un método del estado del arte, además de presentar el resultado de las principales heurísticas. Considerando aquellos obtenidos con los métodos científicos novedosos y las herramientas comerciales, se puede decir que se han podido recuperar sesenta años de investigación para el lenguaje ruso en la tarea de generación automática

de resúmenes. Aunque hace falta mucho por hacer, ya se sabe de un método que supera a la heurística *baseline:first* para ruso (**gráfica 8.10**).

Gráfica 8.10 Resultados obtenidos con la colección en el lenguaje ruso con los métodos científicos novedosos y las herramientas comerciales



Capítulo IX

Conclusiones

En este capítulo se dan las conclusiones sobre la problemática en GART planteada y desarrollada a lo largo del libro. Se consideró como hipótesis el que un humano pudiera replicar el conocimiento necesario para generar un resumen de manera automática en una máquina; con las pruebas mostradas, se observó que ésta no sólo imitó al humano sino que lo superó. Se obtuvo a través del *Test de Turing* que el humano eligió mayormente los resúmenes hechos por una máquina. Sin embargo, quedan algunas cuestiones abiertas que habrán de resolverse en un futuro.

En este libro se presentó un estudio sobre la detección de ideas y composición de resúmenes en los lenguajes: inglés, español, portugués y ruso; con el objetivo de recuperar sesenta años de investigación en cada uno de los lenguajes mediante una actualización, principalmente en español, portugués y ruso.

Se habla de sesenta años de investigación ya que las primeras indagaciones sobre la tarea de GART se remontan a los años cincuenta (Lunh, 1953-1958) para el lenguaje inglés, y prácticamente hasta hace no más de diez años para otros lenguajes. Todas las pesquisas han estado enfocadas en el estudio cualitativo de la tarea de GART. Sin embargo, el estudio cuantitativo es importante y se ha dejado olvidado, puesto que la investigación se enfocó en los números y no en saber si una máquina podía ya realizar resúmenes similares a los del humano.

En este libro se presentaron una serie de pruebas del *Test de Turing* en la tarea de GART, con el objetivo de determinar si el humano ha sido capaz de transmitir a una máquina, por medio de modelos o métodos, los conocimientos necesarios para que pueda emular al humano. La conclusión, según los resultados de las pruebas del *Test de Turing*, es que una máquina se puede considerar inteligente para generar resúmenes. Tanto que en las pruebas hechas para los lenguajes inglés y español hubo entre 56% y 46% de confusión a la hora de seleccionar los resúmenes hechos por las personas.

Asimismo, se presentó una investigación puntualizada de la GART en los lenguajes inglés, español, portugués y ruso; calculando para cada uno los valores de las diferentes heurísticas (*baseline:random, baseline:first* y *Topline*), las herramientas comerciales y los métodos científicos novedosos; de lo cual, una vez considerados los sesenta años de investigación, se concluye lo siguiente para cada lenguaje.

- Para el lenguaje inglés se pudo observar que fue hasta el año 2008 cuando se logró superar a la heurística *baseline:first*. Los avances en la investigación respecto de *Topline* están a 41.04% de avance probando los *corpus* DUC01 y DUC02. Mientras que, en la herramientas comerciales, solamente *Copernic Summarizer* supera a *baseline:first* para DUC02, lo que nos muestra un gran avance en los métodos científicos novedosos pues solamente dos de ellos no rebasan a esta heurística para DUC 2002. De forma general, se puede observar para el lenguaje inglés un avance considerable respecto de las diferentes heurísticas y las herramientas comerciales.

- Acerca del lenguaje español, fue hasta el año 2001 que comenzaron las investigaciones sobre la tarea de GART. A lo largo de ese tiempo se han realizado esfuerzos; sin embargo, no se habían comparado pues no se contaba con un *corpus* especializado. En este libro se presenta el *corpus* TER, y se muestran los valores de las heurísticas *baseline:first*, *baseline:random* y *Topline*. Cabe mencionar que para el español, *baseline:firt* es muy alta y ello representa un reto a superar por los métodos científicos novedosos. Se evalúan las herramientas comerciales y los métodos científicos novedosos. Se concluye que para el lenguaje español se tiene 68.25% de avance respecto de las heurísticas *baseline:random* y *Topline.*
- Sobre el lenguaje portugués ya se habían realizado investigaciones desde el año 2003 con el *corpus TeMário*. Sin embargo, no se contaba con el valor de las heurísticas para determinar el grado de avance. Para el portugués se tiene un avance de 20.27% respecto de las heurísticas *baseline:random* y *Topline*. Se muestra que solamente dos métodos científicos novedosos logran superar a *baseline:first*, mientras que las herramientas comerciales no la rebasan.
- Finalmente, para el lenguaje ruso no se tenían trabajos comparables para la GART. En este libro se presenta un *corpus* especializado llamado *TEXTRUSS*. Anteriormente no había trabajos equiparables debido a que el lenguaje utiliza signos cirílicos, pero con *TEXTRUSS* también se puede obtener textos transliterados para su uso en métodos que trabajan con lenguajes latinos. Para el ruso se muestran las medidas de las heurísticas *baseline:first*, *baseline:random* y *Topline* para *TEXTRUSS*. Se evalúan las herramientas comerciales y los métodos científicos novedosos. Se concluye que para el lenguaje ruso se tiene un avance de 21.09% respecto de *baseline:random* y *Topline.*

Referencias

Acero, I., Alcojor, M., Díaz Esteban, A., Gómez Hidalgo, J.M., Maña López, M.J. (2001, septiembre). Generación automática de resúmenes personalizados. *Proces. Leng. Nat. Nº 27*. Pp. 281-290.

Al Saied, H., Dugué, N., Lamirel, J.-C. (2017). Automatic summarization of scientific publications using a feature selection approach. *Int. J. Digit. Libr*. Pp. 1-13.

Aleixo, P., Pardo, T.A.S. (2008). CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). *ICMC-USP*.

Alfonseca, E., Rodríguez, P. (2003). Generating extracts with genetic algorithms. *Presented at the European Conference on Information Retrieval*, *Springer*. Pp. 511–519.

Alvarado B., A. (2017). *Evaluación de la calidad de las herramientas comerciales y métodos del estado del arte para la generación de resúmenes del corpus DUC-2001*, México: Universidad Autónoma del Estado de México.

Amancio, D.R., Nunes, M.G., Oliveira Jr, O.N., Costa, L. da F. (2012). Extractive summarization using complex networks and syntactic dependency. *Phys. Stat. Mech. Its Appl. 391*. Pp. 1855-1864.

Antiqueira, L. (2007). *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos*, Brasil: Universidade de São Paulo.

Arévalo, J.A. (2017). El español una lengua viva. I*nforme 2017. Instituto Cervantes. infotra*.

Babar, S., Patil, P.D. (2015). Improving Performance of Text Summarization. *Procedia Comput. Sci. 46*. Pp. 354–363.

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D. (2000). An evaluation road map for summarization research. *TIDES July*.

Banko, M., Vanderwende, L. (2004). Using n-grams to understand the nature of summaries. *Proceedings of HLT-NAACL 2004: Short Papers. Association for Computational Linguistics*. Pp. 1–4.

Barzilay, R., Elhadad, M. (1999). Using lexical chains for text summarization. *Adv. Autom. Text Summ*. Pp. 111–121.

Benbrahim, M., Ahmad, K. (1995). Text summarisation: The role of lexical cohesion analysis. *New Rev. Doc. Text Manag*. Pp. 321–335.

Benítez, R., Escudero, G., Kanaan, S., Rodó, D.M. (2014). *Inteligencia artificial avanzada*. Editorial UOC.

Berker, M. (2011). Using genetic algorithms with lexical chains for automatic text summarization.

Bhargava, R., Sharma, Y., Sharma, G. (2016). Atssi: Abstractive text summarization using sentiment infusion. *Procedia Comput. Sci. 89*. Pp. 404–411.

Bing, L., Li, P., Liao, Y., Lam, W., Guo, W., Passonneau, R.J. (2015). Abstractive multi-document summarization via phrase selection and merging. *ArXiv Prepr. ArXiv150601597*.

Bossard, A., Généreux, M., Poibeau, T. (2008). Description of the LIPN System at TAC 2008: *Summarizing* Information and Opinions. *Presented at the TAC 2008*. Pp. 282–291.

Braslavski, P., Gustelev, V. (2007). News Summarization System Based On Machine Learning Approach. *Digit. Libr. Adv. Methods Tecnol. Digit. Collect*. Pp. 142–147.

Brin, S., Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. 56*. Pp. 3825–3833.

Briones, E.G., Cubino, R.L., Sobrino, B.L. (2012). *La noticia y el reportaje. Proyecto Mediascopio Prensa. La lectura de la prensa escrita en el aula*. Ministerio de Educación.

Cabral, L. de S., Lins, R.D., Mello, R.F., Freitas, F., Ávila, B., Simske, S., Riss, M. (2014). A platform for language independent summarization. *Proceedings of the 2014 ACM Symposium on Document Engineering. ACM*. Pp. 203–206.

Carbonell, J., Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*. Pp. 335–336.

Cardoso, P.C., Maziero, E.G., Jorge, M.L., Seno, E.M., Di Felippo, A., Rino, L.H., Nunes, M.G., Pardo, T.A. (2011). CSTnews-a discourse-annotated *corpus* for single and multi-document summarization of news texts in Brazilian Portuguese. *Proceedings of the 3rd RST Brazilian Meeting*. Pp. 88–105.

Cavalieri, D.C., Bastos-Filho, T., Palazuelos-Cagigas, S.E., Sarcinelli-Filho, M., (2015). On combining language models to improve a text-based human-machine interface. *Int. J. Adv. Robot. Syst. 12*. P. 170.

Coarite Choque, R. (2008). Areas de aplicación de la Inteligencia Artificial. *Rev. Inf. Tecnol. Soc*. Pp. 18–22.

Copernic Summarization-Technologies White Paper, 2003.

Corpus, 2014. *corpus*. Gran Dicc. Leng. Esp.

da Cunha Fanego, I. (2005). Hacia un modelo lingüístico de resumen automático de artículos médicos en español. *Proy. Investig. Univ. Pompeu Fabra Inst. Univ. Lingüíst. Apl. Dr. En Cienc. Leng. Lingüíst. Apl. Httpwww Upf Edupdiiulairia Dacunha 0 202*. 07–04.

Donaway, R.L., Drummey, K.W., Mather, L.A. (2000). A comparison of rankings produced by summarization evaluation measures. *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization. Association for Computational Linguistics*. Pp. 69–78.

Edmundson, H.P. (1969). New methods in automatic extracting. *J. ACM JACM 16*. Pp. 264–285.

Edmundson, H.P., Wyllys, R.E. (1961). Automatic abstracting and indexing—survey and recommendations. *Commun. ACM 4*. Pp. 226–234.

El-Haj, M., Rayson, P. (2013). Using a keyness metric for single and multi document summarisation. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-Document Summarization*. Pp. 64–71.

Europe Internet Stats - Population Statistics [WWW Document] (2017). URL https://www.internetworldstats.com/europa2.htm#ru

García-Hernández, R., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., Cruz, R. (2008). Text summarization by sentence extraction using unsupervised learning. *MICAI 2008 Adv. Artif. Intell*. Pp. 133–143.

García-Hernández, R.A., Ledeneva, Y. (2013). Single extractive text summarization based on a genetic algorithm. *Presented at the Mexican Conference on Pattern Recognition, Springer*. Pp. 374–383.

Genest, P.-E., Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. *Proceedings of the Workshop on Monolingual Text-To-Text Generation. Association for Computational Linguistics.* Pp. 64–73.

Gliozzo, D.A., Ackerson, C., Bhattacharya, R., Goering, A., Jumba, A., Kim, S.Y., Krishnamurthy, L., Lam, T., Littera, A., McIntosh, I., Murthy, S., Ribas, M. (2017).

Building Cognitive Applications with IBM Watson Services: *Volume 1 Getting Started*. P. 130.

Hassel, M., Dalianis, H. (2003). Text *Summarizer* (with PRM) [WWW Document]. URL http://*swesum*.nada.kth.se/index-eng-adv.html

Hernández, P.T. (2018). *Desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos para el corpus TEXTRUSS*. México: Universidad Autónoma del Estado de México, Tianguistenco.

Hirao, T., Isozaki, H., Maeda, E., Matsumoto, Y. (2002). Extracting important sentences with support vector machines. *Presented at the Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics*. Pp. 1–7.

Hsu, F. (1999). IBM's deep blue chess grandmaster chips. *IEEE Micro 19*. Pp. 70–81.

Igave, M.S., Gaikwad, C.M. (2016). Efficient Multi-Document Summary Generation Using Neural Network. *Int. J. Adv. Eng. Manag. Sci. IJAEMS*.

Jing, H. (2002). Using hidden Markov modeling to decompose human-written summaries. *Comput. Linguist. 28*. Pp. 527–543.

Jing, H. (2001). *Cut-and-paste text summarization*. EUA: Columbia University.

Jing, H., Barzilay, R., McKeown, K., Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. *AAAI Symposium on Intelligent Summarization. Palo Alto, CA*. Pp. 51–59.

Katragadda, R., Pingali, P., Varma, V. (2009). Sentence position revisited: a robust light-weight update summarization'baseline'algorithm. *Presented at the Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, Association for Computational Linguistics*. Pp. 46–52.

Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I., Paul, A. (2018). Abstractive Text Summarization based on Improved Semantic Graph Approach. *Int. J. Parallel Program*. Pp. 1–25.

Kiyoumarsi, F. (2015). Evaluation Of Automatic Text Summarizations Based On Human Summaries. *Procedia-Soc. Behav. Sci. 192*. Pp. 83–91.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM JACM 46*. Pp. 604–632.

Krishna, R.M., Reddy, C.S. (2016). Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis. *Computational Intelligence in Data Mining—Volume 1. Springer*. Pp. 261–272.

Kupiec, J., Pedersen, J., Chen, F. (1995). A trainable document *Summarizer*. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*. Pp. 68–73.

La Crónica de Hoy | La noticia hecha diario [WWW Document] (2014). URL http://www.cronica.com.mx/noticias.php

Last, M., Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. Pp. 207–237.

Ledeneva, Y. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization, *National Polytechnic Institute*.

Ledeneva, Y., García-Hernández, R.A. (2017). Generación automática de resúmenes. *Retos, propuestas y experimentos, 1st ed*.

Ledeneva, Y., García-Hernández, R.A. (2013). Automatic text summarization with Maximal Frequent Sequences.

Ledeneva, Y., Gelbukh, A., García-Hernández, R.A. (2008). Terms derived from frequent sequences for extractive text summarization. *Presented at the International Conference on Intelligent Text Processing and Computational Linguistics, Springer.* Pp. 593–604.

Ledeneva, Y., Hernández, R., Soto, R., Reyes, R., Gelbukh, A. (2011). EM clustering algorithm for automatic text summarization. *Adv. Artif. Intell*. Pp. 305–315.

Lehmam, A. (2010). Essential *Summarizer*: innovative automatic text summarization software in twenty languages. *Adaptivity, Personalization and Fusion of Heterogeneous Information. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE*. Pp. 216–217.

Leite, D., Rino, L.H. (2009). A Genetic Fuzzy Automatic Text *Summarizer*. *Anais Do Csbc 2009. SBC*. Pp. 779–788.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Presented at the Text summarization branches out: Proceedings of the ACL-04 workshop, Barcelona, Spain*.

Lin, C.-Y. (1999). Training a selection function for extraction. *Presented at the Proceedings of the eighth international conference on Information and knowledge management, ACM*. Pp. 55–62.

Lin, C.-Y., Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics*. Pp. 71–78.

Lin, C.-Y., Och, F.J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*. Pp. 605.

Lloret, E., Palomar, M. (2011). COMPENDIUM: Una herramienta de generación de resúmenes modular. *Proces. Leng. Nat*.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev. 2*. Pp. 159–165.

Luhn, H.P. (1953). Distributor and method for making the same. Google Patents.

Lynn, H.M., Choi, C., Kim, P. (2017). An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms. *Soft Comput*. Pp. 1–11.

Mani, I. (2001). Automatic Summarization. *Natural Language Processing, 3 (Paper*).

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B. (1999). The TIPSTER SUMMAC text summarization evaluation. *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*. Pp. 77–85.

Marcu, D. (1997). From discourse structures to text summaries. *Intell. Scalable Text Summ*.

Margarido, P.R., Pardo, T.A., Antonio, G.M., Fuentes, V.B., Aires, R., Aluísio, S.M., Fortes, R.P. (2008). Automatic summarization for text simplification: Evaluating text understanding by poor readers. *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. ACM*. Pp. 310–315.

Martins, C. (2002). *UNLSumm: Un resumen automático de textos UNL. Departamento de Informática*. Brasil: Universidad Federal de São Carlos. Sao Carlos - SP.

Martins, C., Pardo, T., Espina, A., Rino, L. (2001). *Introducción a los resúmenes automáticos*. Brasil: Universidad Federal de São Carlos.

Mateo, P.L., González, J.C., Villena, J., Martínez, J.L. (2003). Un sistema para resumen automático de textos en castellano. *Proces. Leng. Nat. 31*. Pp. 29–36.

Matias, G. (2016). *Generación Automática de Resúmenes Independientes del Lenguaje*. México: Universidad Autónoma del Estado de México.

Matias, G. (2013). *Generación automática de resúmenes usando algoritmos genéticos*. México: Universidad Autónoma del Estado de México.

McKeown, K., Radev, D.R. (1995). Generating summaries of multiple news articles. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*. Pp. 74–82.

Mendoza Becerra, M.E. (2015). Generación automática de resúmenes extractivos de múltiples documentos basada en algoritmos meméticos.

Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. *Expert Syst. Appl. 41*. Pp. 4158–4169.

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Presented at the Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics*. P. 20.

Mihalcea, R., Radev, D. (2011). *Graph-based natural language processing and information retrieval*. EUA: Cambridge university press.

Mihalcea, R., Tarau, P. (2005). A language independent algorithm for single and multiple document summarization.

Minel, J.-L., Nugier, S., Piat, G. (1997). How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN1.

Mingli, L.I., Sun, L., Han, X. (2016). Combining Relevance Clustering and Graph Model Methods for Automatic Summarization. *J. Residuals Sci. Technol. 13*.

Miranda, S. (2013). *Modelo para la Generación Automática de Resúmenes Abstractivos basados en gráfos conceptuales*. México: Instituto Politécnico Nacional.

Miranda-Jiménez, S., Gelbukh, A., Sidorov, G. (2013). *Summarizing* conceptual graphs for automatic summarization task. *International Conference on Conceptual Structures. Springer*. Pp. 245–253.

Modolo, M. (2003). Supongamos: un entorno para la exploración de métodos de extracción de resumen automático de texto en portugués. *Disert. Masters Dep. Informática UFSCar Sao Carlos - SP.*

Molina, A. (2013). Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. *Intel. Artif. 16*. Pp. 41–62.

Nandhini, K., Balasundaram, S.R. (2014). Extracting easy to understand summary using differential evolution algorithm. *Swarm Evol. Comput. 16.* Pp. 19–27.

Nenkova, A., Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Hlt-Naacl 2004*.

Orăsan, C. (2003). An evolutionary approach for improving the quality of automatic summaries. *Proceedings of the ACL 2003 Workshop on Multilingual Summarization*

and Question Answering-Volume 12. Association for Computational Linguistics. Pp. 37–45.

Orozco, A. (1989). Las RCA2 y la transliteración de nombres de autores personales rusos.

Orrú, T., Rosa, J.L.G., de Andrade Netto, M.L. (2006). SABIO: an automatic portuguese text *Summarizer* through artificial neural networks in a more biologically plausible model. *Presented at the International Workshop on Computational Processing of the Portuguese Language, Springer*. Pp. 11–20.

Ouyang, Y., Li, W., Lu, Q., Zhang, R. (2010). A study on position information in document summarization. *Presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics*. Pp. 919–927.

Over, P., Dang, H., Harman, D. (2007). DUC in context. *Inf. Process. Manag. 43*. Pp. 1506–1520.

Paice, C.D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Inf. Process. Manag. 26*. Pp. 171–186.

Pardo, T. (2002). DMSumm: A Resúmenes de generador automático. Disertación Masters. *Dep. Informática Univ. Fed. São Carlos SaoCarlos - SP*.

Pardo, T., Rino, L., Nunes, M.G. (2003b). NeuralSumm: Un enfoque Conexionista para los Textos resumen automático. *Actas XXIII Congr. Soc. Bras. Comput. VIII*. Pp. 203–245.

Pardo, T.A.S., Rino, L.H.M. (2003). *TeMário: Um corpus para sumarização automática de textos*. Brasil: São Carlos Universidade São Carlos Relatório Téc.

Pardo, T.A.S., Rino, L.H.M., Nunes, M. das G.V. (2003). GistSumm: A summarization tool based on a new extractive method. *Presented at the International Workshop on Computational Processing of the Portuguese Language, Springer*. Pp. 210–218.

Patel, A., Siddiqui, T., Tiwary, U.S. (2007). A language independent approach to multilingual text summarization. *Presented at the Large scale semantic access to content (text, image, video, and sound)*. Pp. 123–132.

Plaza, L. (2011). *Uso de grafos semánticos en la generación automática de resúmenes y estudio de su aplicación en distintos dominios: biomedicina, periodismo y turismo*. España: Universidad Complutense de Madrid, Madrid.

Polya, G., Zugazagoitia, J. (1965). Cómo plantear y resolver problemas. Trillas.

Qazvinian, V., Hassanabadi, L.S., Halavati, R., 2008. *Summarising text with a genetic algorithm-based sentence extraction. Int. J. Knowl. Manag. Stud. 2*. Pp. 426–444.

Rojas, J. (2018). Calculating the Significance of Automatic Extractive Text Summarization using a Genetic Algorithm. *J. Intell. Fuzzy Syst., Applications in Engineering and Technology.*

Rojas J. (2017). *Cálculo de Topline para la generación automática de resúmenes usando algoritmos genéticos*. México: Universidad Autónoma del Estado de México.

Rojas, J.M. (2016). *Evaluación de herramientas comerciales y métodos del estado del arte para la generación de resúmenes en idioma ruso*. México: Universidad Autónoma del Estado de México.

Saggion, H. (2011). Using SUMMA for Language Independent Summarization at TAC 2011. *Presented at the TAC*.

Salton, G., McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill Book Company.

Sparck Jones, K., Galliers, J.R. (1995). Evaluating natural language processing systems: An analysis and review. *Springer Science & Business Media*.

Suanmali, L., Salim, N., Binwahlan, M.S. (2011). GENETIC ALGORITHM BASED SENTENCE EXTRACTION FOR TEXT SUMMARIZATION. *Int. J. Innov. Comput. 1*.

Téllez, A., Montes, M., Villaseñor-Pineda, L. (2009). Using Machine Learning for Extracting Information from Natural Disaster News Reports. *Comput. Sist. Instituto Politécnico Nacional*. Pp. 33–44.

Toledo-Báez, M.C. (2010). Aproximación al resumen automático como herramienta de ayuda a la traducción jurídica en el ámbito del Derecho turístico1. *El Español, Lenguaje de Traducción Para La Cooperación y El Diálogo*. Madrid.

Torres-Moreno, J.-M. (2014). Automatic text summarization. *John Wiley & Sons*.

Traslit [WWW Document] (2016). Транслитератор Translitnet Бывший Translitru. URL https://translit.net/

Turing, A.M. (1950). Computing Machinery and Intelligence. *Computing Machinery and Intelligence.* Pp. 433–460.

Uddin, M.N., Khan, S.A. (2007). A study on text summarization techniques and implement few of them for Bangla language. *Computer and Information Technology, 2007. Iccit 2007. 10th International Conference On. IEEE*. Pp. 1–4.

UNE 50-103-90 (1990). Preparación de resúmenes.

Vázquez, E. (2015). *Modelo de relevancia de la posición de las oraciones en resúmenes de textos, mediante regresión simbólica*. México: Universidad Autónoma del Estado de México.

Venegas, R. (2011). Evaluación de resúmenes en español con Análisis Semántico Latente: Una implementación posible. *Rev. Signos 44.* Pp. 85–102.

Verma, R., Lee, D. (2017). Extractive Summarization: Limits, Compression, Generalized Model and Heuristics. *ArXiv Prepr. ArXiv170405550.*

Villar, A.M. (2005). *Microsoft Word 2003. Nociones básicas: Operaciones básicas, diseño, pruebas e impresión*, 1ra ed. ideaspropias.

Villatoro E. (2007). *Generación automática de resúmenes de múltiples documentos*. México: Instituto Nacional de Astrofísica, Óptica y Electrónica.

Wan, X. (2010). Towards a unified approach to simultaneous single-document and multi-document summarizations. *Presented at the Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics.* Pp. 1137–1145.

Wang, B., Zhang, J., Liu, Y., Zou, Y. (2017). Density peaks clustering based integrate framework for multi-document summarization. *CAAI Trans. Intell. Technol. 2*. Pp. 26–30.

Wang, L., Cardie, C. (2013). Domain-Independent Abstract Generation for Focused Meeting Summarization. *Presented at the ACL (1)*. Pp. 1395–1405.

Главные новости - Газета.Ru [WWW Document] (2015). URL https://www.gazeta.ru/

Anexo A

Test de Turing para el lenguaje español

En el anexo A se presentan dos pruebas más del *Test de Turing* para el lenguaje español. Se muestran los textos completos y los resúmenes que se proporcionaron a los humanos para hacer la identificación de los dos textos hechos por personas. Además, se muestran las tablas donde se menciona cuáles de los resúmenes fueron generados por la máquina y cuáles por el humano.

A continuación, se presenta el segundo texto utilizado en el *Test de Turing* para el lenguaje español así como los resúmenes hechos por el humano y aquellos generados por la máquina.

Analizan estados acciones contra el dengue

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de Programas Preventivos y Control de Enfermedades (Cenaprece), Jesús Felipe González, se revisó la situación epidemiológica del dengue de la región occidente del país, y los programas estatales de control y prevención. González Roldán subrayó la importancia del trabajo coordinado entre la federación, estados, municipios y la población para reducir el potencial de transmisión de esta enfermedad. Indicó que este trabajo de anticipación se debe enfocar en las medidas de prevención y promoción de la salud, para la eliminación de criaderos. Llamó a no bajar la guardia, pues la incidencia de letalidad en México está por debajo del indicador de la Organización Mundial de la Salud (OMS), mientras que en otras partes del mundo esta enfermedad va a la alza. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos, siempre y cuando la ciudadanía tome conciencia de participar en la eliminación de criaderos. A su vez Óscar Villaseñor Anguiano, secretario de Salud en Nayarit, agregó que el trabajo coordinado entre los estados compromete acciones como el control larvario, abatización y fumigación. "Los estados occidentales debemos desarrollar acciones conjuntas en el combate al dengue, para lograr disminuir el número de casos", mencionó.

Resumen 1

Llamó a no bajar la guardia, pues la incidencia de letalidad en México está por debajo del indicador de la Organización Mundial de la Salud (OMS), mientras que en otras partes del mundo esta enfermedad va a la alza. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos, siempre y cuando la ciudadanía tome conciencia de participar en la eliminación de criaderos. El dengue es un padecimiento que afecta a 27 estados del país…

Resumen 2

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. González Roldán subrayó la importancia del trabajo coordinado entre la federación, estados,…

Resumen 3

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. A su vez Óscar Villaseñor Anguiano, secretario de Salud en Nayarit, agregó…

Resumen 4

En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de Programas Preventivos y Control de Enfermedades (Cenaprece), Jesús Felipe González, se revisó la situación epidemiológica del dengue de la región occidente del país,

y los programas estatales de control y prevención. A su vez Óscar Villaseñor Anguiano, secretario de Salud en…

Resumen 5

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. Indicó que este trabajo de anticipación se debe enfocar en las medidas de prevención y promoción de la salud, para la eliminación de criaderos. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos,…

Resumen 6

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de…

De los resúmenes presentados, dos corresponden a los elaborados por el humano (*gold standard*), dos a las heurísticas y dos generados de forma automática por una máquina. A continuación, se da la correspondencia de cada uno de ellos.

- Resumen 1 — *Baseline:random (heurística)*
- Resumen 2 — Matias (2016) (máquina)
- Resumen 3 — Humano 1 (*gold standard*)
- Resumen 4 — *Microsoft Office Word* (máquina)
- Resumen 5 — Humano 2 (*gold standard*)
- Resumen 6 — *Baseline:first* (*heurística*)

Enseguida se presenta el tercer texto utilizado en el *Test de Turing* para el lenguaje español, así como los resúmenes generados por el humano y aquellos hechos por la máquina.

Público mexicano sensibiliza a la Oreja de Van Gogh en el Auditorio Nacional

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. "El último vals" se hizo sonar con Leire Martínez, quien al término de la canción expresó: "Buenas noches México, como saben nuestro último trabajo, Primera Fila, lo realizamos aquí", expresó la española antes de presentar "Cuando dices adiós", bajo luces multicolores. El ritmo de "Mi calle es Nueva York" dio paso a temas de la banda como "Vestido azul", balada de su álbum Lo que te conté mientras te hacías la dormida, misma que se ilumino con luces color pastel. Prosiguió "Inmortal", canción de su material discográfico, A las cinco en el Astoria, que fuera el disco debut de Martínez en el grupo. "Algo se nos ha quedado en tantas ocasiones que hemos visitado este país, la canción es para darles un poquito de nosotros", expresó la líder de grupo para presentar "Una y otra vez", segunda canción inédita del formato Primera Fila. Antes de que la cantante expresara: "La noche irá cargada de sorpresas, hoy no podía faltar este invitado a nuestro concierto", dijo para presentar "Mi vida sin ti" junto con Samo, quien no fue bien recibido. Aparecieron los clásicos temas de La Oreja de Van Gogh como "París" que se ligó con "Europa VII", melodía de la banda para crear conciencia sobre los problemas sociales que existen alrededor del mundo; la música de Xabi San Martín levitó en el aire mientras Leire realizó un cambio de ropa. Con *leggins*, blusón a rayas y botas negras, la española volvió a escena para expresar: "Queremos estar muy cerca de ustedes, esta canción habla de las sensaciones cuando te entregan por primera vez a tu bebé", aseguró para acercarse al público y cantar "Palabras para Paula". Leonel García ingresó al escenario para repetir con la agrupación el momento musical logrado en el nuevo disco y cantar "La playa"; el baladista vestido con traje gris oscuro agradeció la invitación. La velada continuó y la banda rememoró el primer sencillo de toda su carrera musical: "El 28". No se fueron las sorpresas y el tema "Adiós", poco tocado en sus conciertos, resonó en el Auditorio Nacional para seguir con "María", "Deseos de cosas imposibles" y "Jueves", tema donde bajaron las luces, el público prendió los

celulares y proyectó en todo el espacio figuras que ante la vista panorámica crearon un cielo estrellado. La respuesta de Leire fue espontánea: lágrimas que cayeron sobre sus mejillas y expresó: "Gracias por recordarnos que estos son los momentos que más valen la pena, por confiar en la música en directo", aseguró Leire con voz entrecortada y visiblemente emocionada por los gritos y los aplausos de los seguidores. La intensidad musical de "Muñeca de trapo" y "La niña que llora en tus fiestas" inundó el espacio, luego llegó "El primer día del resto de mi vida" y "Pálida luna". El final se acercó en una fiesta de globos de colores que aventaron sus seguidores en todo el recinto y que dieron vida a los temas "Cometas por el cielo", "Pop", "20 de enero" y "Puedes contar conmigo". Al término, La Oreja de Van Gogh levantó la bandera de México y España y se abrazaron para celebrar la noche y despedirse de sus fans.

Resumen 1

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. "El último vals" se hizo sonar con Leire Martínez, quien al término de la canción expresó: "Buenas noches México, como saben nuestro último trabajo, Primera Fila, lo realizamos aquí", expresó la española antes de presentar "Cuando dices adiós", bajo luces multicolores. El ritmo de "Mi calle es Nueva York" dio paso a temas…

Resumen 2

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. Aparecieron los clásicos temas de La Oreja de Van Gogh como "París" que…

Resumen 3

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con

una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. "El último vals" se hizo sonar con Leire Martínez, quien al término de…

Resumen 4

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Aparecieron los clásicos temas de La Oreja de Van Gogh como "París" que se ligó con "Europa VII", melodía de la banda para crear conciencia sobre los problemas sociales que existen alrededor del mundo; la música de Xabi San Martín levitó en el aire mientras Leire realizó un cambio de ropa. La velada…

Resumen 5

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. "El último vals" se hizo sonar con Leire Martínez, quien al término de…

Resumen 6

La respuesta de Leire fue espontánea: lágrimas que cayeron sobre sus mejillas y expresó: "Gracias por recordarnos que estos son los momentos que más valen la pena, por confiar en la música en directo", aseguró Leire con voz entrecortada y visiblemente emocionada por los gritos y los aplausos de los seguidores Al término, La Oreja de Van Gogh levantó la bandera de México y España y se abrazaron para celebrar la noche y despedirse de sus fans El ritmo de "Mi calle es Nueva York" dio paso a temas de la banda como "Vestido azul", balada de su álbum Lo…

De los resúmenes presentados para esta prueba dos corresponden a los hechos por el humano (*gold standard*), dos a las heurísticas y dos generados de forma automática por una máquina. A continuación, se da la correspondencia de cada uno de ellos.

- Resumen 1 — Matias (2016) (máquina)
- Resumen 2 — Humano 1 (*gold standard*)
- Resumen 3 — *Microsoft Office Word* (máquina)
- Resumen 4 — Humano 2 (*gold standard*)
- Resumen 5 — *Baseline:first* (*heurística*)
- Resumen 6 — *Baseline:random* (*heurística*)

Anexo B

Test de Turing para el lenguaje inglés

En este anexo se presentan dos pruebas más del *Test de Turing* para el lenguaje inglés. Se muestran los textos completos y los resúmenes que se proporcionaron a las personas para hacer la identificación de los dos resúmenes hechos por el humano. Además, se dan las tablas donde se observa cuáles de los textos fueron hechos por la máquina y cuáles por el humano.

A continuación, se presenta el segundo texto utilizado en el *Test de Turing* para el lenguaje inglés así como los resúmenes hechos por el humano y aquellos generados por la máquina.

Gilbert Reaches Jamaican Capital With 110 Mph Winds

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said. Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica. Meanwhile, Havana Radio reported today that 25,000 people were evacuated from Guantanamo Province on Cuba's southeastern coast as strong winds fanning out from Gilbert began brushing the island. All Jamaica-bound flights were canceled at Miami International Airport, while flights from Grand Cayman, the main island of the three-island chain, arrived packed with frightened travelers. "People were running around in the main lobby of our hotel (on Grand Cayman) like chickens with their heads cut off," said one vacationer who was returning home to California through Miami. Hurricane warnings were posted for the Cayman Islands, Cuba and Haiti. Warnings were discontinued for the Dominican Republic. "All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane," the service said, adding, "Little change in strength is expected for the next several hours as the hurricane moves westward over Jamaica." The Associated Press' Caribbean headquarters in San Juan, Puerto Rico, was unable to get phone calls through to Kingston, where high winds and heavy rain preceding the storm drenched the capital overnight, toppling trees, causing local flooding and littering streets with branches. Most Jamaicans stayed home, boarding up windows in preparation for the hurricane. Some companies broadcast appeals for technicians and electricians to report to work. The weather bureau predicted Gilbert's center, 140 miles southeast of Kingston before dawn, would pass south of Kingston and hit the southern parish of Clarendon. Flash flood warnings were issued for the parishes of Portland on the northeast and

St. Mary on the north. The north coast tourist region from Montego Bay on the west and Ocho Rios on the east, far from the southern impact zone and separated by mountains, was expected only to receive heavy rain. Officials urged residents in the higher risk areas along the south coast to seek higher ground. "It's certainly one of the larger systems we've seen in the Caribbean for a long time," said Hal Gerrish, forecaster at the National Hurricane Center. Forecasters at the center said the eye of Gilbert was 140 miles southeast of Kingston at dawn today. Maximum sustained winds were near 110 mph, with tropical-storm force winds extending up to 250 miles to the north and 100 miles to the south. Prime Minister Edward Seaga of Jamaica alerted all government agencies, saying Sunday night: "Hurricane Gilbert appears to be a real threat and everyone should follow the instructions and hurricane precautions issued by the Office of Disaster Preparedness in order to minimize the danger." Forecasters said the hurricane had been gaining strength as it passed over the ocean after it dumped 5 to 10 inches of rain on the Dominican Republic and Haiti, which share the island of Hispaniola. "We should know within about 72 hours whether it's going to be a major threat to the United States," said Martin Nelson, another meteorologist at the center. "It's moving at about 17 mph to the west and normally hurricanes take a northward turn after they pass central Cuba." Cuba's official Prensa Latina news agency said a state of alert was declared at midday in the Cuban provinces of Guantanamo, Holguin, Santiago de Cuba and Granma. In the report from Havana received in Mexico City, Prensa Latina said civil defense officials were broadcasting bulletins on national radio and television recommending emergency measures and providing information on the storm. Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti. The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night. Flights were canceled Sunday in the Dominican Republic, where civil defense director Eugenio Cabral reported some flooding in parts of the capital of Santo Domingo and power outages there and in other southern areas.

Resumen 1

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said. Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica. The weather bureau predicted Gilbert's center, 140 miles southeast…

Resumen 2

national radio and television recommending emergency measures and The weather bureau predicted Gilbert's center, 140 miles was declared at midday in the Cuban provinces of Guantanamo, strong winds fanning out from Gilbert began brushing the island the northeast and St Mary on the north. The north coast tourist In the report from Havana received in Mexico City, Prensa Latina vacationer who was returning home to California through Miami "Right now it's actually moving over Jamaica," said Bob. The storm ripped the roofs off houses and flooded coastal areas "We should know within about 72 hours whether it's going to be…

Resumen 3

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. All Jamaica-bound flights were canceled at Miami International Airport, while flights from Grand Cayman, the main island of the three-island chain, arrived packed with frightened travelers. Hurricane warnings were posted for the Cayman Islands, Cuba…

Resumen 4

Hurricane Gilbert hit Jamaica today with 110 mph winds and torrential rain, causing serious damage in Kingston overnight. The storm center is expected to hit land at Clarendon parish, then move lengthwise across the island. The government is preparing for the worst with government agencies on alert and coastal residents directed to move to higher ground. Communications have already been affected. Gilbert, described as one of the larger systems, has already caused some damage

in Puerto Rico, the Dominican Republic, Haiti, and Cuba. Fears are high on the Cayman Islands, the next target on its track…

Resumen 5

Gilbert Reaches Jamaican Capital With 110 Mph Winds Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said…

Resumen 6

Hurricane Gilbert, packing 110mph winds and torrential rain, moved over the Jamaican capital city of Kingston today after skirting Puerto Rico, Haiti and the Dominican Republic. It's tropical-storm force winds extend up to 250 miles to the north and 100 miles to the south. Hal Gerrish, a forecaster with the National Hurricane Center said it is one of the larger systems seen in the Caribbean for a long time. Warnings were posted for the Cayman Islands, Haiti and Cuba but discontinued for the Dominican Republic. Jamaicans are expecting to bear the brunt of Gilbert as its eye moves lengthwise across…

De los resúmenes presentados para esta prueba, dos corresponden a los hechos por el humano (*gold standard*), dos a las heurísticas y dos generados de forma automática por una máquina. Enseguida, se da la correspondencia de cada uno de ellos:

- Resumen 1 — *Copernic* (máquina)
- Resumen 2 — *Baseline:random* (heurística)
- Resumen 3 — Matias (2016) (máquina)
- Resumen 4 — Humano 1 (gold standard)
- Resumen 5 — *Baseline:first* (heurística)
- Resumen 6 — Humano 2 (gold standard)

Enseguida se presenta el tercer texto utilizado en el *Test de Turing* para el lenguaje inglés así como los resúmenes generados por el humano y aquellos hechos por la máquina.

Hurricane Hits Jamaica With 115 mph Winds; Communications Disrupted

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. For half an hour, the hurricane lashed the city, tearing branches from trees, blowing down fences and whipping paper through the air. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm moved across the city. Skies brightened, the winds died down and people waited for an hour before the second blow of the hurricane arrived. All Jamaica-bound flights were canceled at Miami International Airport. Flights from the Cayman Islands, reportedly next in the path of the hurricane, arrived in Miami packed with travelers cutting short their vacations. "People were running around in the main lobby of our hotel (on Grand Cayman Island) like chickens with their heads cut off," said one man. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. It said Jamaica would receive up to 10 inches of rain that would cause flash floods and mud slides. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," he said. Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic. Hurricane warnings were issued Monday for the south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic. High winds and heavy rain preceding the storm drenched Kingston overnight, toppling trees, causing local flooding and littering streets with branches. Most of Jamaica's 2.3 million people stayed home, boarding up windows in preparation for the hurricane. The popular north coast resort area, on the other side of the mountains, was expected to receive heavy rain but not as much damage from the hurricane as the south coast, where officials urged residents to seek higher ground. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. In Washington, the Navy reported its bases at Guantanamo Bay, Cuba, and Roosevelt Roads, Puerto Rico, had taken various precautionary steps but appeared to be safe from

the brunt of the hurricane. Lt. Ken Ross, a spokesman, said the Navy station at Guantanamo reported that as of 2:30 p.m. EDT, the brunt of the storm appeared to be passing southeastern Cuba. "They have reported maximum winds of 25 knots and gusts up to 50 knots," said Ross. "But there are no reports of injuries or damage." The spokesman said earlier in the day, Guantanamo had moved to "Condition Two," meaning electrical power usage was cut back to only essential uses and "all non-essential personnel sent to their barracks." The storm also skirted Puerto Rico without causing any damage to military facilities, Ross said. Sheets said Gilbert was expected next to sweep over the Cayman Islands, on its westward track, and in two to three days veer northwest into the southern Gulf of Mexico. Residents of the neighboring Caymans, a British dependency to the northwest, were urged to "rush all preparatory actions." The National Weather Service warned that the Caymans could expect high waters and large waves "which may undermine buildings along the beaches." "All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane," the service advised. Forecaster Hal Gerrish on Sunday described Gilbert "certainly one of the larger systems we've seen in the Caribbean for a long time."

❧ Resumen 1

The full force of Hurricane Gilbert slammed into Kingston, Jamaica, at noon on Monday. Torrential rain and 115 mph winds severely damaged the city and its airport. No casualties or injuries have been reported. The storm is expected to move lengthwise across the island, dropping as much as 10 inches of rain. Next landfall is expected to be the Cayman Islands, where tourists are attempting to evacuate and residents were making preparations. The United States Navy reported its bases on Puerto Rico and at Guantanamo Bay, Cuba, were not damaged by Gilbert as it moved through the central Caribbean…

❧ Resumen 2

hurricane, "he said evacuated from coastal areas in Guantanamo Province on the nation's residents to seek higher ground Havana Radio, meanwhile, reported Monday that 25,000 people were Jamaica would receive up to 10 inches of rain that would cause reported that as of 2:30 p m EDT, the brunt of the storm appeared an hour before the second blow of the hurricane arrived southeastern coast as Gilbert's winds and rain began to brush the For half an hour, the hurricane lashed the city,

tearing overnight, toppling trees, causing local flooding and littering Cuba, and Roosevelt Roads, Puerto Rico, had taken…

Resumen 3

Hurricane Gilbert slammed into Kingston, Jamaica on Monday, with torrential rains and 115mph winds that ripped roofs off buildings, uprooted trees, downed power lines and did heavy damage to the airport and parked aircraft. No fatalities in this city of 750,000 people have been reported. Jamaica's popular north coast is not expected to receive as much damage as the south coast, where officials urged residents to seek higher ground. The storm skirted Puerto Rico and is now tracking toward the Cayman Islands. The U.S. Navy reports that its Cuban bases at Guantanamo Bay and Roosevelt Island appear to be relatively …

Resumen 4

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. For half an hour, the hurricane lashed the city, tearing branches from trees, blowing down fences and whipping paper through the air. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm…

Resumen 5

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm moved across the city. Skies brightened, the winds died down and people waited for an hour before the second blow of the…

Resumen 6

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. Hurricane warnings were issued Monday for the

> south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic High winds and heavy rain preceding the storm drenched Kingston overnight, toppling trees, causing local flooding and littering streets with branches. The popular north coast…

De los resúmenes presentados para esta prueba, dos corresponden a los hechos por el humano (*gold standard*), dos a las heurísticas y dos generados de forma automática por una máquina. A continuación, se da la correspondencia de cada uno de ellos.

- Resumen 1 — Humano 1 (*gold standard*)
- Resumen 2 — *Baseline:random* (*heurística*)
- Resumen 3 — Matias (2016) (máquina)
- Resumen 4 — Humano 2 (*gold standard*)
- Resumen 5 — *Baseline:first* (*heurística*)
- Resumen 6 — *Copernic* (máquina)

Anexo C

Ejemplo de resumen en el lenguaje portugués

En el anexo C se presenta un ejemplo de una noticia en el lenguaje portugués, así como dos resúmenes: uno generado por una herramienta comercial y otro por un método científico novedoso.

Para el lenguaje portugués aún no se realiza un *Test de Turing*. Sin embargo, en esta sección se presenta un ejemplo de una noticia del *corpus TeMário*. Además, se muestran dos resúmenes: uno hecho por una herramienta comercial y el otro por un método científico novedoso.

CIDADE É CANTADA EM MAIS DE 1.800 MÚSICAS

Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Essa história de citar a cidade começou em 1750, quando dois compositores (Calixto e Anchieta Arzão) decidiram fazer "Missa à São Paulo", partitura recuperada e gravada pela primeira vez em 1970 com regência de Júlio Medaglia. Depois disso, parece que não foi mais possível conter homenagens e desilusões musicais dos compositores pela que é hoje a terceira maior metrópole do mundo. Até a primeira frase do Hino Nacional menciona São Paulo: "Ouviram do Ipiranga..." Esses dados foram pesquisados por um paraibano de João Pessoa, radicado em São Paulo desde 75, que há cinco anos levanta a história musical da cidade em sebos e livrarias. O escritor e jornalista Assis Angelo, 41, está agora preparando o que ele chama de a primeira enciclopédia musical sobre São Paulo. Os primeiros 300 verbetes já estão escritos. Angelo acredita que o material resultante da sua pesquisa é suficiente para 900 verbetes e umas 600 páginas de livro, à espera de patrocinadores. E como seu trabalho é enciclopédico, vale dizer que a cidade já foi cantada de "A" a "Z", passando por "X" e "Y", por intérpretes e compositores de todos os Estados brasileiros. Por exemplo, com "Z", "Zona Leste Total" (de Luiz Carlos, 1991); com "X", "Xamego Paulista" (de Arlindo Bettio e Nhozinho, 1987); com "Y", "Yayá do Peruche" (de L. Correa e Rodolfo Vila) e com "A", entre outras, "A Baixada do Glicério" (de Enerdino Ortiz, João Marques e Manoel Lourenço). Se quantidade significar amor, Adoniram Barbosa foi o mais apaixonado dos cantores. Adoniram lidera o ranking com nada menos que 22 músicas sobre São Paulo. Tom Zé e a dupla Tonico e Tinoco também têm lá sua quedinha pela cidade. Cada um gravou 11 músicas. Quem pensa que Caetano Veloso parou na "arroz de festa" Sampa quando quis falar da cidade, mostra ou que não entende nada desta cidade ou que não sabe nada de Caetano. O doce bárbaro fala de São Paulo em outras cinco músicas, menos que Itamar Assumpção (dez músicas) e mais que Alvarenga e Ranchinho, que gravaram quatro composições. O ano em que mais se cantou São Paulo foi o do 4.º Centenário, 1954. Foram gravadas 72 músicas, com versões até japonês

enaltecendo a cidade. Nem Hebe Camargo deixou de gravar a sua. "A Hebe vai ficar uma arara, mas a música que ela gravou, "Paulicéia em Festa", era um horror, muito ruim", diz o pesquisador Angelo. Já a composição "IV Centenário", de Mário Zan e J.M. Alves, fez tanto sucesso na época, que vendeu, numa São Paulo quase provinciana, mais de cinco milhões de discos. Só para registro, Mário Zan é o autor do bolero que só estourou e ficou bastante conhecido depois de gravado em espanhol. A estrofe é esta: "Dizem que os homens/ não devem chorar/ por uma mulher/ que não soube amar..." Se o Campeonato Paulista algum dia for definido pela quantidade de música composta para os times, só vai dar Corinthians. Das 140 músicas compostas para os clubes paulistas, 85 foram dedicadas ao time do Parque São Jorge.

❧ *Shvoong* (Herramienta comercial)

CIDADE É CANTADA EM MAIS DE 1.800 MÚSICAS VICTOR AGOSTINHO Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Essa história de citar a cidade começou em 1750, quando dois compositores (Calixto e Anchieta Arzão) decidiram fazer "Missa à São Paulo", partitura recuperada e gravada pela primeira vez em 1970 com regência de Júlio Medaglia. Até a primeira frase do Hino Nacional menciona São Paulo: "Ouviram do Ipiranga..." Adoniram lidera o ranking com nada menos que 22 músicas sobre São Paulo. O doce bárbaro fala de São Paulo em outras cinco músicas, menos que Itamar Assumpção (dez músicas) e mais que Alvarenga e Ranchinho, que...

❧ *AG-Multi* (Método científico novedoso)

Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Depois disso, parece que não foi mais possível conter homenagens e desilusões musicais dos compositores pela que é hoje a terceira maior metrópole do mundo. Esses dados foram pesquisados por um paraibano de João Pessoa, radicado em São Paulo desde 75, que há cinco anos levanta a história musical da cidade em sebos e livrarias. Os primeiros 300 verbetes já estão escritos. Angelo acredita que o material resultante da sua pesquisa é suficiente para 900 verbetes e umas 600 páginas de livro, à espera de patrocinadores...

Anexo D

Ejemplo de resumen en el lenguaje ruso

En el anexo D se presenta un ejemplo de una noticia en el lenguaje ruso, así como dos resúmenes: uno generado por una herramienta comercial y otro por un método científico novedoso.

Para el lenguaje ruso aún no se realiza un *Test de Turing*. Sin embargo, en esta sección se presenta un ejemplo de una noticia del *corpus TEXTRUSS*. Además, se presentan dos resúmenes: uno hecho por una herramienta comercial y el otro por un método científico novedoso.

Как не стать жертвой автоподставы

Как узнать автоподставщика на дороге и что делать при встрече с автомошенниками

Фотография: Shutterstock

17.08.2015, 20:58 | Алина Распопова

Автоподставщики сами бросаются под колеса, а после вымогают у водителей деньги. Они просят компенсировать вред здоровью или поломку дорогих часов. Другие мошенники подставляются на своих старых иномарках под водителей-новичков и выманивают по пять тысяч евро за раз. Им удается убеждать водителей, что случай не страховой, и угрожают расправой. Как не попасться на уловки автоподставщиков, «Газете.Ru» рассказали эксперты ГУ МВД России по Москве. Чтобы добраться до кошельков наивных водителей, автоподставщики используют как новые, так и классические способы обмана. Так, некоторое время назад сотрудники Московского уголовного розыска задержали 43-летнего мужчину, который ловко изобразил, как его якобы сбил проезжающий мимо автомобиль. На самом деле он специально бросался под машины, а его сообщник наносил удар по автомобилю жертвы для имитации звука удара. Далее «пешеход» демонстрировал водителю сломанные дорогостоящие часы или планшетный компьютер. Для решения проблемы он требовал выплатить компенсацию. Для усиления психологического давления на водителя аферист представлялся адвокатом и показывал поддельное удостоверение. Чтобы дополнительно надавить на свою жертву, он звонил по громкой связи своему подельнику, выступающему в роли инспектора ГИБДД. Его сообщник уверенным голосом заявлял, что такой проступок влечет за собой лишение водительских прав. Обманутые люди отдавали «потерпевшему» крупные суммы, порой доходящие до миллиона рублей. По такой проверенной схеме действовали еще несколько злоумышленников, которые также попались в руки полиции. Еще одна организованная группа из трех человек обманом выманивала деньги у водителей, убеждая их, что они повредили их дорогую машину, а случай нестраховой. Доверчивые автомобилисты выкладывали до €5 тыс., только бы избавить себя от неприятностей. По просьбе «Газеты.Ru» специалисты в ГУ МВД России по Москве рассказали о том, как вычислить

автоподставщиков. Используя свой опыт, они объяснили, как работают автомошенники и как правильно себя вести при встрече с ними. Как выбирают жертву «Подставлялы — хорошие психологи, и чаще всего их жертвами становятся неопытные автолюбители, — рассказали «Газете.Ru» в ГУ МВД России по Москве. — В первую очередь это «чайники» и любители болтать за рулем по мобильному телефону. Еще один тип легкой добычи — начинающий водитель со знаком «У» на стекле. В числе потенциальных клиентов — те, кто водит агрессивно и постоянно перестраивается из ряда в ряд». Со слов самих преступников, в качестве жертвы они выбирают только мужчин. Женщины начинают звонить мужьям и друзьям, после чего приезжают люди и начинаются ненужные разборки. Вне зоны риска обладатели новых дорогих иномарок. Редко подставляются под машины, в которых едут несколько человек — свидетели мошенникам ни к чему. Для работы автоподставщики используют подержанные машины, но известных престижных марок. Это могут быть старенькие Mercedes, BMW, Audi, Volvo. На деле цена таких средств передвижения не выше $10 тыс. Неповрежденные машины практически никогда не подставляются.

Одиночные разводки

«Самый грубый способ одиночной подставы — обогнать жертву, подрезать и резко оттормозиться, подставив под удар корму, — отметили в ГУ МВД России по Москве. — Если предполагаемый «спонсор» не успел затормозить — мошенники будут стараться повесить на него всю вину. Ведь всегда виноват тот, кто сзади». Еще один распространенный вариант работы в паре: «охотник» заходит по правому борту жертвы сзади. Этот водитель ведет себя так, как будто и не собирается приближаться, и ждет перестроения «спонсора» в правый ряд. Следуя ПДД, «жертва» заблаговременно показывает правый «поворотник». В ответ на это «подстава» всем своим поведением дает понять «спонсору», что пропускает его. Но как только «жертва» подает вправо, «подстава» резко ускоряется и подставляет свой левый борт под удар. Для удобства инсценировки такого ДТП «подстава» выбирает темное время суток, машину темных цветов и едет только с «габаритами». Попасть на удочку может и водитель, который пытается выехать из левого или среднего ряда, в том числе и на круговом движении. В этот момент автомобилиста подсекают справа. Как правило, не до, а сразу же после удара следует возмущенное «бибиканье». Это тоже часть спектакля, рассчитанная на возможных свидетелей. Мошенники также любят подставлять правый борт под тех, кто бодро мчится по правому ряду, не уступая дорогу соседям слева от себя при объезде припаркованных машин.

Работа в паре

Есть и классические примеры парной работы. Потенциальная жертва должна двигаться по крайнему левому ряду. Ей на хвост плотно садится спешащий водитель и начинает сигналить дальним светом. Логика большинства нормальных водителей — уступить. Не ожидая подвоха, жертва начинает перестраиваться правее. В этот момент ее цепляет машина, которая до того спокойно двигалась чуть поодаль. Ранее водитель мог не обращать на нее внимания, либо она просто находилась в мертвой зоне. Согнавшая жертву с полосы машина якобы уезжает, жертва остается один на один с «невинно пострадавшим». Еще пример. На относительно свободной дороге, двигаясь по крайнему левому ряду с хорошей скоростью, потенциальная жертва «развода» догоняет вяло ползущего впереди якобы «чайника», который упорно отказывается уступить дорогу. Жертва делает рывок вправо, а там его уже поджидает перехватчик. Или такой способ: потенциальная «жертва» едет в среднем ряду. Справа подкатывает «подстава», слева – «сгоняющий» – дорогая машина, возможность контакта с которой инстинктивно отметается всяким нормальным водителем. Так они и едут втроем – параллельно и рядом. Вдруг, «сгоняющий» делает резкий поворот руля в сторону жертвы. Та, чтобы избежать контакта, тоже уходит вправо. Даже если «жертва» контролирует свой правый борт, то «подстава» может неожиданно подвинуться к «жертве», оставаясь при этом в пределах своей полосы движения. В результате «жертва» въезжает правым бортом в левый борт «подставы». По ПДД в ДТП виновата «жертва». «Сгоняющий» уезжает, его задние номера не читаются.

Как понять, что вас разводят?

После инцидента подставщики немедленно начинают убеждать якобы виноватого водителя, что он должен заплатить добровольно и прямо на месте. Нередко злоумышленники устанавливают жесткие временные ограничения. Например, «деньги нужны через час, через 40 минут сервис закроется, у тебя есть два часа собрать деньги». Некоторые мошенники практикуют хорошо зарекомендовавший себя метод «злого-доброго». «Злой» требует максимальную сумму, а «добрый» соглашается на существенно меньшую, мотивируя тем, что, так уж и быть, ремонтироваться они будут у знакомых.

Что делать?

«Законный путь подстащикам изначально невыгоден. Они не получат желаемую сумму, и у них нет желания светиться в правоохранительных органах и страховых компаниях, — предупреждают в ГУ МВД России по Москве. — Договориться с ними по-хорошему невозможно. За один только поцарапанный бампер с вас потребуют баснословную сумму. А уж если вы поедете с

«подставлялами» в их сервис, будьте готовы, что мастер объявит о поврежденных лонжеронах и выставит внушительный счет. Поэтому закройтесь в машине и не выходите до приезда ГИБДД. Позвоните домой, друзьям, обрисуйте ситуацию, в которую попали, опишите марки, номера, приметы автомашин «подставлял» и их самих. Фотографируйте «подставлял» и номера их машины. В большинстве случаев уже одно это заставит их спешно залезть в свой автомобиль и уехать. При этом обязательно продержитесь до приезда сотрудников ГИБДД».

Tool4noobs (Herramienta comercial)

Как не попасться на уловки автоподставщиков, «Газете.Ru» рассказали эксперты ГУ МВД России по Москве. Мошенники также любят подставлять правый борт под тех, кто бодро мчится по правому ряду, не уступая дорогу соседям слева от себя при объезде припаркованных машин. Как выбирают жертву «Подставлялы — хорошие психологи, и чаще всего их жертвами становятся неопытные автолюбители, — рассказали «Газете. Ru» в ГУ МВД России по Москве. Одиночные разводки «Самый грубый способ одиночной подставы — обогнать жертву, подрезать и резко оттормозиться, подставив под удар корму, — отметили в ГУ МВД России по Москве. Они не получат желаемую сумму, и у них нет желания светиться в правоохранительных органах и…

AG-Multi (Método científico novedoso)

Они просят компенсировать вред здоровью или поломку дорогих часов. Как не попасться на уловки автоподставщиков, «Газете.Ру» рассказали эксперты ГУ МВД России по Москве. Цхтобы добраться до кошельков наивных водителей, автоподставщики используют как новые, так и классические способы обмана. На самом деле он специально бросался под машины, а его сообщник наносил удар по автомобилю жертвы для имитации звука удара. Еще одна организованная группа из трех человек обманом выманивала деньги у водителей, убеждая их, что они повредили их дорогую машину, а случай нестраховой. Используя свой опыт, они объяснили, как работают автомошенники и как правильно себя вести при встрече с ними. В первую очередь это…

Anexo E

Palabras vacías en el lenguaje inglés

En el anexo E se presenta la lista de palabras vacías para la GART en el lenguaje inglés.

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

MAINLY, MANY, MAY, MAYBE, ME, MEAN, MEANWHILE, MERELY, MIGHT, MORE, MOREOVER, MOST, MOSTLY, MUCH, MUST, MY, MYSELF, N, NAME, NAMELY, ND, NEAR, NEARLY, NECESSARY, NEED, NEEDS, NEITHER, NEVER, NEVERTHELESS, NEW, NEXT, NINE, NO, NOBODY, NON, NONE, NOONE, NOR, NORMALLY, NOT, NOTHING, NOVEL, NOW, NOWHERE, O, OBVIOUSLY, OF, OFF, OFTEN, OH, OK, OKAY, OLD, ON, ONCE, ONE, ONES, ONLY, ONTO, OR, OTHER, OTHERS, OTHERWISE, OUGHT, OUR, OURS, OURSELVES, OUT, OUTSIDE, OVER, OVERALL, OWN, P, PARTICULAR, PARTICULARLY, PER, PERHAPS, PLACED, PLEASE, PLUS, POSSIBLE, PRESUMABLY, PROBABLY, PROVIDES, Q, QUE, QUITE, QV, R, RATHER, RD, RE, REALLY, REASONABLY, REGARDING, REGARDLESS, REGARDS, RELATIVELY, RESPECTIVELY, RIGHT, S, SAID, SAME, SAW, SAY, SAYING, SAYS, SECOND, SECONDLY, SEE, SEEING, SEEM, SEEMED, SEEMING, SEEMS, SEEN, SELF,SELVES, SENSIBLE, SENT, SERIOUS, SERIOUSLY, SEVEN, SEVERAL, SHALL, SHE, SHOULD, SHOULDN'T, SINCE, SIX, SO, SOME, SOMEBODY, SOMEHOW, SOMEONE, SOMETHING, SOMETIME, SOMETIMES, SOMEWHAT, SOMEWHERE, SOON, SORRY, SPECIFIED, SPECIFY, SPECIFYING, STILL, SUB, SUCH, SUP, SURE, T, T'S, TAKE, TAKEN, TELL, TENDS, TH, THAN, THANK, THANKS, THANX, THAT, THAT'S, THAT'S, THE, THEIR, THEIRS, THEM, THEMSELVES, THEN, THENCE, THERE, THERE'S, THEREAFTER, THEREBY, THEREFORE, THEREIN, THERES, THEREUPON, THESE, THEY, THEY'D, THEY'LL, THEY'RE, THEY'VE, THINK, THIRD, THIS, THOROUGH, THOROUGHLY, THOSE, THOUGH, THREE, THROUGH, THROUGHOUT, THRU, THUS, TO, TOGETHER, TOO, TOOK, TOWARD, TOWARDS, TRIED, TRIES, TRULY, TRY, TRYING, TWICE, TWO, U, UN, UNDER, UNFORTUNATELY, UNLESS, UNLIKELY, UNTIL, UNTO, UP, UPON, US, USE, USED, USEFUL, USES, USING, USUALLY, UUCP, V, VALUE, VARIOUS, VERY, VIA, VIZ, VS, W, WANT, WANTS, WAS, WASN'T, WAY, WE, WE'D, WE'LL, WE'RE, WE'VE, WELCOME, WELL, WENT, WERE, WEREN'T, WHAT, WHAT'S, WHATEVER, WHEN, WHENCE, WHENEVER, WHERE, WHERE'S, WHEREAFTER, WHEREAS, WHEREBY, WHEREIN, WHEREUPON, WHEREVER, WHETHER, WHICH, WHILE, WHITHER, WHO, WHO'S,

WHOEVER, WHOLE, WHOM, WHOSE, WHY, WILL, WILLING, WISH, WITH, WITHIN, WITHOUT, WON'T, WONDER, WOULD, WOULDN'T, X, Y, YES, YET, YOU, YOU'D, YOU'LL, YOU'RE, YOU'VE, YOUR, YOURS, YOURSELF, YOURSELVES, Z, ZERO

Anexo F

Palabras vacías en el lenguaje español

En el anexo F se presenta la lista de palabras vacías para la GART en el lenguaje español.

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

Anexo G

Palabras vacías en el lenguaje portugués

En el anexo G se presenta la lista de palabras vacías para la GART en el lenguaje portugués.

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

Anexo H

Documentación del *Corpus* TER

En el anexo H se presenta la documentación del *corpus* TER creado para la tarea de GART para el lenguaje español. Se describe cada una de las etapas para su construcción y, finalmente, la composición del *corpus*.

H.1 Introducción

En este documento se describe la construcción de un *corpus* de textos en español para resúmenes, con el fin de servir como apoyo al área de procesamiento del lenguaje natural en español, principalmente en el área de la generación de resúmenes automáticos. El *corpus* fue creado bajo el proyecto de la Red Temática en Tecnologías del Lenguaje (Red TTL).

Se compone por noticias en español mexicano y dos resúmenes generados por dos humanos. El objetivo principal del documento es servir como *corpus* para evaluar los métodos científicos novedosos y las herramientas comerciales para la generación de resúmenes extractivos. Sin embargo, puede ser utilizado para diversos fines como los análisis lingüísticos de textos, ya sea para resúmenes o sólo análisis de texto, sistemas de recuperación de información o detección de tópicos.

Actualmente, existen métodos científicos novedosos que trabajan con la generación de resúmenes extractivos, como: Ledeneva (2008), Ledeneva, (2008ª), García (2008), Montiel, (2009), García, (2013), Mendoza, (2014), Meena, (2015), Bhargava, (2016), entre otros. Sin embargo, todos ellos trabajan sólo para el lenguaje inglés. Existen otros que son independientes del lenguaje y prueban más de una colección de documentos, como Mihalcea, (2005), Patel, (2007), Last, (2010), Saggion, (2011). Pero, a pesar de probar con más de una colección de documentos en lenguajes como inglés, portugués, chino, entre otros, dejan de lado a uno de los más importantes: el español.

Según Cervantes (2013), los expertos predicen que para el año 2050 habrá más de 530 millones de hispanohablantes, de los cuales 100 millones estarán viviendo en los Estados Unidos. Esto nos muestra un amplio campo de trabajo para el PLN en español. Por eso la importancia de contar con un *corpus* en este sentido, y que además sea en español mexicano para conocer mejor el comportamiento de los diferentes métodos y herramientas para la generación de resúmenes extractivos en nuestro lenguaje.

H.2 *Corpus* de textos en español para resúmenes

H.2.1 Características generales

El *corpus* creado es en español mexicano exclusivo para la generación de resúmenes extractivos para un solo documento. Se presenta en formato digital sobre noticias periodísticas.

H.2.1.1 Protocolo de compilación

Búsqueda y acceso a la información

El *corpus* fue creado a partir de noticias electrónicas disponibles en la red. Las noticias fueron obtenidas de la página oficial del periódico "Crónica" ("La Crónica de Hoy | La noticia hecha diario," 2014). Se seleccionaron 20 noticias de las siguientes categorías: academia, bienestar, ciudad, cultura, deportes, espectáculos, estados, mundo, nacional, negocios, opinión y sociedad; con un total de 240 noticias. Aquellas seleccionadas fueron del mes de abril de 2014. Una de las consideraciones más importantes para la elección de las noticias fue que tuvieran diferentes longitudes, pero siempre más de cien palabras.

Preprocesamiento

Las noticias se descargaron de la Web en un formato .html, por lo que se realizó el siguiente proceso de limpieza y normalización.

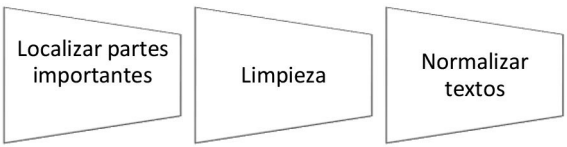


Figura H.1 Fases del preprocesamiento

- Localizar partes importantes. Las noticias que están disponibles en la red, además del texto de la noticia, pueden contener más información como anuncios, fotografías, ligas a otras páginas, etc. Por ello fue necesario detectar las partes que nos brindaran información relevante y necesaria para la construcción del *corpus*.

Los segmentos que se eligieron fueron: la clave de la noticia, la cual consiste en un número único que la identifica y que también forma parte del nombre del archivo, el título de la noticia, la categoría a la que pertenece, la fecha de publicación y el texto de la misma.

- Limpieza. El proceso de limpieza consiste en eliminar todas las etiquetas html, texto, ligas, imágenes, etc., dejando solamente el título de la noticia, la categoría a la que pertenece, la fecha de publicación y el texto. La limpieza se realizó mediante un programa en Java, utilizando expresiones regulares para que se aplicara de manera automática para todos los textos.
- Normalizar textos. Una vez limpios los textos, se realizó el etiquetado para identificar de manera más sencilla las partes de la noticia. A continuación, en la tabla H.1 se describen las etiquetas utilizadas y posteriormente se da un ejemplo de etiquetado.

Tabla H.1 Descripción de etiquetas para etiquetar los textos completos

Etiquetas	Descripción
<DOC></DOC>	Etiqueta que indica el inicio y final de documento
<DOCNO> </DOCNO>	Etiqueta que indica el nombre del documento
<FILEID></FILEID>	Etiqueta que indica un número único del documento
<HEAD></HEAD>	Etiqueta que indica el título del documento
<CATEGORY> </ CATEGORY >	Etiqueta que indica la categoría a la que pertenece el documento
<DATE></DATE>	Etiqueta que indica la fecha de expedición del documento
<TEXT></TEXT>	Etiqueta que indica cual es el texto a resumir
<s><\s>	Etiquetas que indican el inicio y fin de una oración

Ejemplo de texto completo etiquetado

<DOC>
<DOCNO>
<s docid=”09ED020414_825542” num=”1” wdcount=”1”> 825542 </s>
</DOCNO>

<FILEID>
<s docid=”09ED020414_825542” num=”2” wdcount=”1”>09ED020414_825542</s>
</FILEID>
<HEAD>
<s docid=”09ED020414_825542” num=”3” wdcount=”10”> Atención digna a grupos vulnerables distingue a Toluca, afirma alcaldesa </s>
</HEAD>
<CATEGORY>
<s docid=”09ED020414_825542” num=”4” wdcount=”1”> Estados </s>
</CATEGORY>
<DATE>
<s docid=”09ED020414_825542” num=”5” wdcount=”1”> 2014-04-02 </s>
</DATE>
<TEXT>
<s docid=”09ED020414_825542” num=”6” wdcount=”61”> La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría</s>

<s docid=”09ED020414_825542” num=”7” wdcount=”82”> En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan</s>

<s docid=”09ED020414_825542” num=”8” wdcount=”40”> En presencia de miembros del Cabildo y de autoridades municipales, la presidenta del sistema DIF de Toluca, Diana Elisa González Calderón, indicó que en esta ocasión se entregaron 110 auxiliares auditivos, 100 juegos de lentes y 30 sillas de ruedas.</s>

```
</TEXT>
</DOC>
```

Almacenamiento

Para nombrar a cada uno de los archivos se observaron las siguientes consideraciones; al tratarse de 20 archivos por categoría, se asignó un número consecutivo (1-20), posterior a esto se tomaron dos letras para cada categoría. A continuación se describen: academia (AC), bienestar (BI), ciudad (CI), cultura (CU), deportes (DE), espectáculos (ES), estados (ED), mundo (MU), nacional (NA), negocios (NE), opinión (OP) y sociedad (SO). Seguido de la abreviación de la categoría, se colocó la fecha de la noticia y luego separado por un guion bajo la clave de la noticia. Ejemplo de nombre de archivo: 01AC010414_825278.txt. Finalmente, se tienen 12 carpetas con 20 archivos en cada una; en total son 240 archivos.

H.2.2 Construcción de los resúmenes

Una vez construido el *corpus* de noticias en español, para cada archivo se crearon dos resúmenes hechos por dos humanos.

Selección de los humanos. Las consideraciones tomadas para seleccionar a un humano fueron que tuviera nacionalidad mexicana, educación mínima de licenciatura y se le dieron las siguientes indicaciones.

H.2.2.1 Construcción de los resúmenes

A las personas se les proporcionó la noticia dividida en oraciones con el número de palabras de cada una de ellas. Se les pidió que leyeran completamente la noticia y seleccionaran las oraciones que consideraban importantes. De aquellas seleccionadas, se les pidió que crearan un resumen mayor a 100 palabras. En el apartado de anexos se muestra un ejemplo de las instrucciones dadas a los humanos, además de la lista con los nombres de cada uno de ellos.

A continuación, en la tabla H.2 se describen las etiquetas empleadas en los resúmenes generados por los humanos y posteriormente se da un ejemplo de etiquetado.

Tabla H.2 Descripción de etiquetas para etiquetar los resúmenes

Etiquetas	Descripción
<SUM></SUM>	Etiqueta que indica el inicio y final del resumen hecho por el humano.
CATEGORY	Indica la categoría a la que pertenece la noticia.
TYPE	Indica el tipo de resumen, en este caso es por documento.
SIZE	Indica el tamaño mínimo de palabras que debe tener el resumen.
DOCREF	Muestra el nombre del documento base para la generación del resumen extractivo.
SELECTOR	Indica las iniciales del humano que realizó el resumen.
Summarizer	Indica cuál de los dos resúmenes generados es A- el primero, B- el segundo.

Ejemplo de resumen etiquetado

<SUM
CATEGORY=”ESTADOS”
TYPE=”PERDOC”
SIZE=”100”
DOCREF=”09ED020414_825542”
SELECTOR=”EX”
Summarizer=”B”>

La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría. En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan.

</SUM>

H.2.2.2 Recopilación de los resúmenes generados por los humanos

Una vez creado el resumen extractivo por el humano, se asignó una clave a cada uno de los textos para nombrar los archivos de resumen de la siguiente manera. Ejemplo de nombre de archivo: SUM_01AC010414_825278_LX.sum.

Como se puede observar, para identificar que el archivo pertenece a los resúmenes modelos, se agregó la etiqueta SUM a cada uno; seguido de esto, se colocó el nombre de la noticia original y finalmente se agregó la clave asignada al humano. La extensión de estos archivos es .sum.

Finalmente resultaron 12 carpetas con 40 archivos cada una, dando un total de 480 archivos de resúmenes modelo.

H.2.3 Descripción del *corpus*

Como se había mencionado, el *corpus* está compuesto por 240 noticias de diferentes categorías. La colección se presenta de forma etiquetada, donde se describe en qué consiste cada parte que conforma el texto. Es importante mencionar que el *corpus* fue dividido en oraciones, las cuales también son etiquetadas para facilitar el análisis del texto.

A continuación, en la **tabla H.3** se muestran las categorías en las que está dividido el *corpus*, el número de documentos que lo componen y el número de oraciones.

Los resúmenes generados por los humanos son de más de cien palabras. Sin embargo, para la evaluación de los métodos y las herramientas para la generación de resúmenes se recomienda que la evaluación se realice a cien palabras.

H.2.4 Organización del *corpus*

Como se puede observar en la **figura H.2**, el *corpus* está compuesto por dos carpetas. Resúmenes, donde se localizan los dos resúmenes hechos por los dos humanos para cada uno de los documentos originales. Finalmente, como se muestra en la **figura H.3**, la carpeta Textos Originales se divide en dos carpetas, Textos por archivos y Textos por categoría, donde se encuentran los textos completos originales etiquetados.

Tabla H.3 Características de los textos completos del *corpus*

	Categoría	Número de textos	Número de palabras	Promedio de palabras	Número de oraciones	Promedio de oraciones
Periódico Crónica	Academia	20	10966	548,3	382	19,1
	Bienestar	20	11801	590,05	405	20,25
	Ciudad	20	7568	378,4	219	10,95
	Cultura	20	8631	431,55	297	14,85
	Deportes	20	9519	475,95	363	18,15
	Espectáculos	20	8869	443,45	311	15,55
	Estados	20	7471	373,55	185	9,25
	Mundo	20	7108	355,4	247	12,35
	Nacional	20	7533	376,65	186	9,3
	Negocios	20	7523	376,15	229	11,45
	Opinión	20	12716	635,8	443	22,15
	Sociedad	20	6507	325,35	228	11,4
	Total	240	106212		3495	
	Media			442,55		14,5625

El *corpus* está organizado de la siguiente manera:

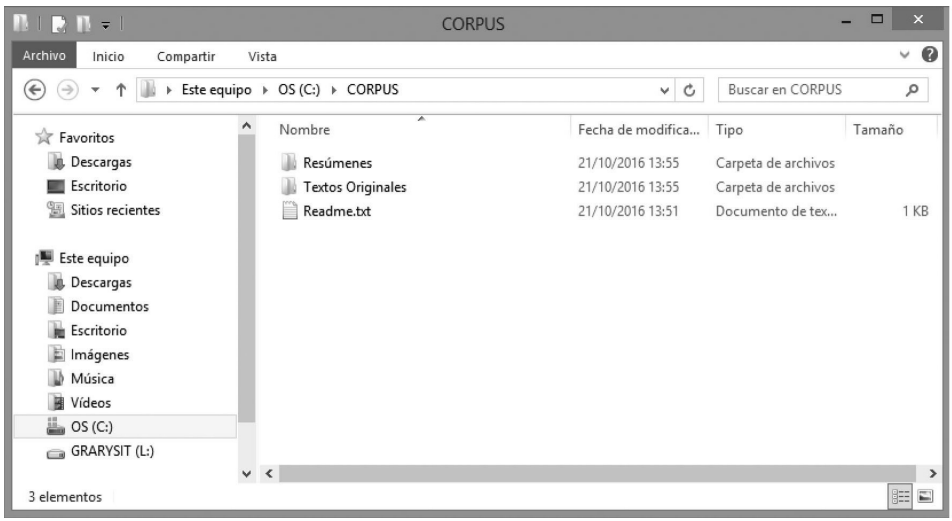


Figura H.2 Directorio de *corpus*

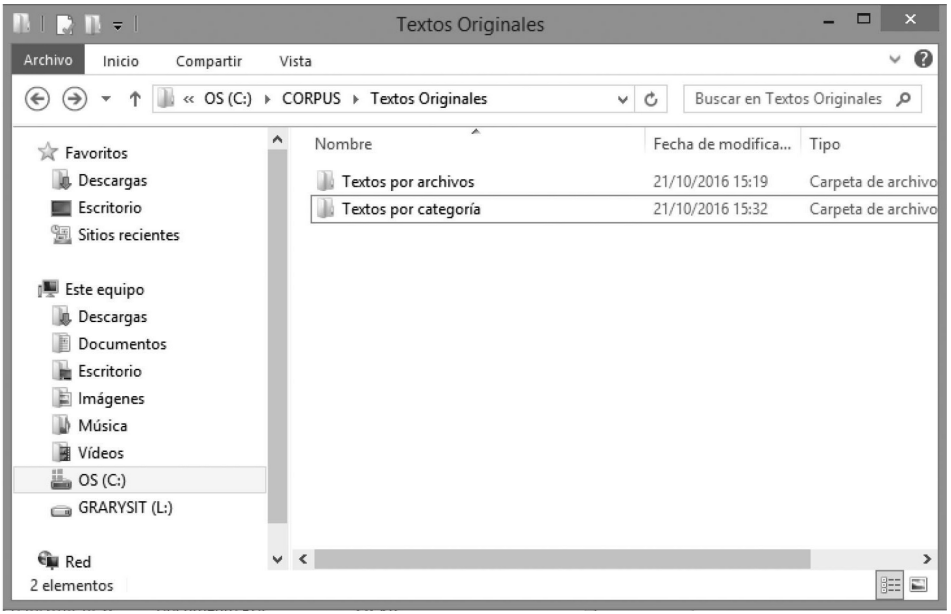


Figura H.3 Directorio de textos originales

La única diferencia entre estas dos últimas carpetas es que en Textos por categoría se encuentran 12 carpetas, una para cada categoría del *corpus* donde hay 20 archivos pertenecientes a la categoría. Mientras que en la carpeta Texto por archivos están los 240 archivos. En la **figura H.4** se muestra el contenido de estas carpetas.

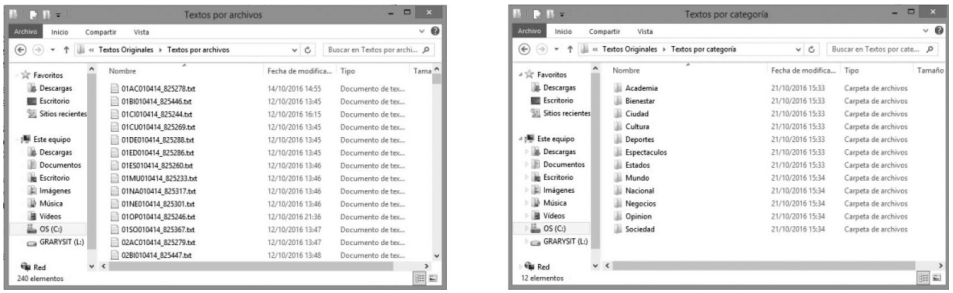


Figura H.4 Directorio de textos por archivos y textos por categoría

H.3 Consideraciones finales

Como se mencionó anteriormente, el *corpus* propuesto en este trabajo está construido para ser utilizado principalmente en el estudio de la generación de resúmenes extractivos para el lenguaje español. El *corpus* presenta este etiquetado, de tal manera que sólo se muestran los textos completos y los resúmenes en diferentes carpetas. Sin embargo, el etiquetado permite eliminar las partes que no se consideren importantes para el análisis de esta colección, por ejemplo, si se desea sólo trabajar con el texto, se considera únicamente el contenido en la etiqueta <TEXT></TEXT>. Una de las aportaciones importantes de este trabajo es que el texto está separado por oraciones, lo que muestra una estandarización para futuros usos.

Referencias (Anexo H)

Bhargava, 2016	Bhargava, R., Sharma, Y., & Sharma, G. (2016). ATSSI: Abstractive Text Summarization Using Sentiment Infusion. Procedia Computer Science, 89, 404-411.
Crónica. (s.f.)	© La Crónica Diaria S.A. de C.V. Obtenido de © La Crónica Diaria S.A. de C.V: http://www.cronica.com.mx/noticias.php
García, 2008	García R., Montiel, R., Ledeneva, Y., Rendón, e., Gelbukh, A. & Cruz, R. (2008). Text Summarization by Senten ce Extraction Using Unsupervised Learning. 7° Conferencia Internacional Mexicana de Inteligencia Artificial (MICAI08); Notas de la conferencia de Inteligencia Artificial, Springer-Verlag, Vol 5317, pp133-143.
García, 2013	García-Hernández, R. A., & Ledeneva, Y. (2013, June). Single extractive text summarization based on a genetic algorithm. In Mexican Conference on Pattern Recognition (pp. 374-383). Springer Berlin Heidelberg.
Last, 2010	Last, M. & Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. NATO Science for Peace and Security Series - D: Information and Communication Security. Vol. 27: Web Intelligence and Security, pp. 207-237.

Ledeneva, 2008	Ledevena, Y. N. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. México, D.F.: Presentada en el Instituto Politécnico Nacional, para obtención del grado de Doctor.
Ledeneva, 2008[a]	Ledeneva, Y., Gelbukh, A. & García, R. (2008). Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. Research in Computing Science, Vol. 34, pp.163-174.
Meena, 2015	Meena, Y. K., & Gopalani, D. (2015). Feature Priority Based Sentence Filtering Method for Extractive Automatic Text Summarization. Procedia Computer Science, 48, 728-734.
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, 41(9), 4158-4169.
Mihalcea, 2005	Mihalcea, R. & Taran, P.. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Vol. 1, pp. 602-607.
Montiel, 2009	Montiel, R. (2009). Generación automática de resúmenes mediante aprendizaje no supervisado. Edo. de México: Presentada en el Instituto Tecnológico de Toluca, para obtención del Título de Ingeniero en Sistemas Computacionales.
Patel, 2007	Patel, A., Siddiqui, T & Tiwary, U. (2007). A language independent approach to multilingual text summarization. Conference RIA2007, Pittsburgh PA, U.S.A., 123-132.
Saggion, 2011	Saggion, H., Szasz, S., & Grupo, T. A. L. N. (2011). A Bilingual Summary *Corpus* for Information Extraction and other Natural Language Processing Applications. on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011), 28.

Anexo I

Documentación del *Corpus TeMário*

En el anexo I se presenta la documentación del *corpus TeMário* para portugués. Se hace una traducción al lenguaje español del documento donde se muestra dicho *corpus* (Pardo *and* Rino, 2003) que originalmente fue escrito en portugués. Esto se hace para entender los detalles de *TeMário*; aunque se pueden detectar algunas discrepancias en la traducción, pues no se domina el portugués.

Universidad de San Paulo - USP
Universidad Federal de San Carlos - UFSCar
Universidad Estadual Paulista - UNESP

TeMário: Un *corpus* para resúmenes automáticos de textos

Thiago Alexandre Salgueiro Pardo
Lucia Helena Machado Rino

NILC-TR-03-09
Octubre de 2003
Series de Informes del Centro Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Código Postal 668, 13560-970 San Carlos, SP, Brasil

Resumen

En este anexo se describe *TeMário*, el cual es un *corpus* orientado a resúmenes de textos automáticos. Desarrollado para diversos fines, como análisis lingüístico, la formación de resúmenes automáticos y su evaluación posterior, el *TeMário* es compuesto básicamente por textos periodísticos y sus resúmenes son en portugués. Éstos se construyeron por un experto y escritor para la publicación de textos en portugués. Este *corpus* se utiliza principalmente para investigaciones específicas de los métodos de resumen automático en el proyecto EXPLOSA.[33]

ÍNDICE

[33]Desarrollado con el apoyo de la FAPESP (Proc. Nro. 01/08849-8).

I.3. CONSIDERACIONES FINALES
REFERENCIAS BIBLIOGRÁFICAS
APÉNDICE A — ESPECIFICACIONES DEL MANUAL DE TRABAJO DE RESÚMENES

I.1 Introducción

Este anexo describe el *TeMário* (siglas de '***TE**xtos com su**MÁRIOs***'), el cual es un *corpus* construido con el fin de obtener los resúmenes automáticos bajo el proyecto EXPLOSA[34] (exploración de varios métodos para el resumen automático).

Este *corpus* se compone básicamente de los textos periodísticos y sus respectivos resúmenes manuales, generados por un experto y escritor para su publicación en portugués.[35] Además de servir para diferentes propósitos, por ejemplo, el análisis lingüístico, la construcción y la formación de resúmenes automáticos y la evaluación de estos sistemas, que también se ocuparán en otras tareas relacionadas, cuyas áreas de interés actual implican la Recuperación de Información y Detección de Temas.

Del proyecto EXPLOSA hay varios sistemas que pueden beneficiarse de este *corpus* de formación y evaluación, como *GistSumm (*Pardo *et. al.,* 2003a), el NeuralSumm (Pardo *et al.* 2003b)*,* el DMSumm[36] (Pardo, 2002), *SuPor* (Módolo, 2003) y el UNLSumm (Martins, 2002*)*. Además de estos sistemas, cuya información genérica se puede encontrar en el sitio NILC (http:// www.nilc.icmc.usp.br/) otras actividades pueden ser desarrolladas con *TeMário*. Por ejemplo, los estudios del modo en que el experto reconoce la información relevante en un texto para componer sus resúmenes, o la identificación de parámetros que indican los criterios para resumir en el modelado de sistemas computacionales. Los detalles sobre estas tareas y su relación con el resumen automático se encontraban inicialmente en Pardo y Rino (2003) y Martins *et al*. (2001).

[34]http://www.dc.ufscar.br/lucia~/projects/EXPLOSA.htm (FAPESP, Proc. Emisión. 01/08849-8).

[35]Los resúmenes manuales utilizados aquí para indicar los textos escritos construidos por un profesional. En inglés, el nombre de resúmenes profesionales o resumidores humanos también es utilizado por algunos autores.

[36]Todos estos están disponibles para su descarga en http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23#resource

Además de las tareas directamente relacionadas con el resumen automático, actualmente otros proyectos en el NILC pueden hacer uso de *TeMário*, como el proyecto LAZIO-WEB que es una construcción de recursos para diversas investigaciones, entre ellas la recuperación de la información así como el etiquetado de texto o información en portugués. En un contexto más amplio, el programa formará parte de Linguateca[37] que es un gran repositorio internacional de recursos que contiene datos e información para procesamiento automático de la lengua portuguesa.

I.2 *TeMário*

I.2.1 Características generales

El nombre de *TeMário* para centrarse en el *corpus* fue elegido por dos razones: por hacer referencia a los objetos que lo componen -resúmenes y textos- y por la palabra -tema- en su nombre, cuyo reconocimiento es esencial en la tarea de resumir.

Para construir el *TeMário*, 100 artículos de prensa han sido recolectados, con un total de 61.412 palabras. 60 textos en línea del periódico Folha de Sao Paulo (en adelante, identificado por la abreviatura **FSP**) distribuidos uniformemente en las secciones: especial, mundo y opinión; los 40 textos restantes fueron publicados en el periódico de Brasil (en adelante, identificado por la sigla **JB**) también en línea, y son de igual manera uniformemente distribuidos en secciones: internacional y política. La **tabla I.1** resume estos datos, mostrando el número de palabras por sección y el promedio de ellas para cada sección del texto. Según la **tabla I.1**, las palabras promedio por sección son 12.282 y las promedio por texto son de 613, siendo esto correspondiente a los textos que van de 1 a 2 páginas y media.

Los textos periodísticos fueron elegidos para componer el *corpus* porque tenían un lenguaje dirigido a una amplia audiencia de lectores, y por lo tanto una cobertura portuguesa en términos de vocabulario y en términos de construcciones gramaticales. Así, se excluyeron automáticamente a partir de la selección más suplementos tales como comentarios de periódicos; FSP, por ejemplo, es destinado

[37] http://www.linguateca.pt/

Tabla I.1 Características del *corpus* de "textos-fuente"

Periódicos	Secciones	Número de textos	Número de palabras	Promedio de palabras/texto
Folha de São Paulo	Especial	20	12340	617
	Mundo	20	13739	686
	Opinión	20	10438	521
Jornal do Brasil	Internacional	20	12098	604
	Política	20	12797	639
	Total	100	61412	
	Promedios generales		12282	613

a un público más culto. Esta limitación tiene el objeto principal de facilitar las tareas relacionadas con resúmenes automáticos: es habitual el uso de mano de obra especializada para elaborar evaluaciones de resultados automáticos. Un estilo más exagerado impone una mayor dificultad en la lectura, la comprensión y evaluación, lo que lleva a resultados equívocos sobre el centro de la tarea.

Esta relación también se evidencia por el hecho de que el género periodístico es actualmente el más utilizado en evaluaciones a gran escala en resúmenes automáticos: los concursos internacionales de evaluación de resúmenes automáticos, como text SUMMArization evaluation conference (SUMMAC) y DUC, que han utilizado los textos periodísticos de grandes volúmenes de datos.

Para la producción del *TeMário*, una vez recolectados los artículos periodísticos se procedió a la construcción de los resúmenes correspondientes, por lo que los textos son llamados "textos-fuente".

I.2.2 Construcción de resúmenes

Los textos recopilados se envían al experto y escritor para su publicación en portugués y para llevar a cabo dos tareas: la construcción de resúmenes correspondientes a la Tarea 1 (apéndice del *TeMário*) y para cada texto-fuente su idea principal Tarea 2 (apéndice del *TeMário*). Así, en la Tarea 1 el profesor produjo resúmenes informativos. En la Tarea 2 se asumió el cargo de lector de textos simple, aprendiendo lo que es más importante.

Relacionar ambas tareas en la toma de la idea principal de un texto es esencial para la producción de buenos resúmenes informativos. Es decir, para identificar frases que hacen referencia a la idea principal, éstas serán las que conducirán al experto a la construcción de los resúmenes ya que deben tener todos (una parte significativa) información principal del "texto-fuente", e incluso pueden remplazar (la principal condición de resúmenes informativos). Se considera aquí la alternancia entre funciones del experto, del lector y el escritor, la tarea comúnmente conocida como "reescribir texto-fuente en forma condensada" *(Mani*, 2001).

Además de la necesidad de producir resúmenes informativos, el sistema automático generador de resúmenes tuvo una restricción adicional: el tamaño de cada resumen debe ser de aproximadamente 25-30% del tamaño de su texto-fuente. Desde el punto de vista del resumen automático, esto es equivalente a establecer tasas de compresión de "textos-fuente" para el intervalo de 70-75%, es decir, 70 o 75% de los contenidos de estos textos deben ser descartados para la elaboración de resúmenes.

Las instrucciones que el experto realiza para ambas tareas, el resumen y el marcado de las frases, se pueden encontrar en el apéndice A.

I.2.3 Complementación del *corpus*

El *TeMário* compuesto por 100 "textos-fuente" y sus manuales de resúmenes constituye un conjunto significativo (aunque relativamente pequeño) de datos de texto para diversas tareas de resumen automático, como la formación de los sistemas automatizados y la personalización de los resúmenes de textos manuales del mismo género y dominio. Sin embargo, para las tareas de evaluación, considerando que los resúmenes manuales son el resultado de un proceso de reescritura del contenido del texto-fuente que el escritor considera más relevante, utilizar manuales como resúmenes "ideales" con el fin de comparar con aquellos generados automáticamente, no es tarea fácil: difícilmente es una correspondencia explícita. Por lo tanto, las revisiones de los resúmenes manuales y automáticos, en general, se basan en el criterio humano y por ende la revisión de los experimentos es cara y compleja en su aplicación. Para minimizar este problema, es común utilizar extractos de "ideas" en lugar de resúmenes ideales para la comparación con resultados automáticos, especialmente cuando se está tratando de realizar la extracción de resumen automático.

En este caso, la terminología indica que tanto los extractos de ideas como los extractos resultantes del proceso de resumen, en sí, se derivan de la metodología de extracción, la aplicación de segmentos textuales seleccionados para el texto abreviado, cuya principal característica es la de reproducir literalmente partes del "texto-fuente". Así, se hace posible considerar simplemente los patrones similares entre los extractos ideales y sus extractos correspondientes de evaluación, para determinar si son buenos representantes de la idea principal de los "textos-fuente", y los extractos son ideales. Claramente, esta etapa puede ser realizada automáticamente en la mayoría de los casos y la evaluación anterior puede ser de forma diferente, con considerables beneficios en términos de costo y complejidad. Por esta razón, el *TeMário* fue complementado con extractos producidos con los resúmenes manuales por un generador de extractos de ideas.

El generador de extractos de ideas identifica y extrae frases de oraciones yuxtapuestas de "textos-fuente" que tienen el mismo contenido de las frases de los resúmenes manuales correspondientes. Para eso, utiliza la medición coseno de Salton (1989), según la metodología descrita por Rino y Pardo (2003). Es importante decir que los extractos de ideas no pueden ser realmente ideales para aprovechar en su totalidad las ideas de forma completa y de contenido relevante del "texto-fuente", como lo haría el experto: una medida del coseno, ya que se basa puramente en la concurrencia de palabras entre el resumen manual y el "texto-fuente", puede producir extractos con frases inapropiadas. Sin embargo, estos extractos se consideran ideales para ser utilizados lo mejor posible, desde el punto de vista del costo/beneficio de la producción automática.

La **tabla I.2** relaciona el tamaño de los resúmenes manuales y extractos de ideas. Cabe señalar que el número promedio de palabras de los resúmenes manuales es significativamente menor que el promedio de palabras de extractos de ideas. Esa diferencia puede ser debida al hecho de que el experto será capaz de resumir el contenido que desea de la mejor manera posible para satisfacer las restricciones de resumen, utilizando el proceso de reescritura. En el caso de la extracción de ideas que satisface a estas restricciones, no siempre es trivial pues está previamente fijada en la unidad mínima a extraer los "textos-fuente" en general; las frases se extraen en su totalidad para componer los resúmenes. Por esta razón, es más común tener más extractos de resúmenes a mano.

Tabla I.2 Características de los resúmenes manuales y extractos de ideas

		Resúmenes a mano		Extracto de ideas	
Periódicos	**Secciones**	**Número de palabras**	**Promedio de palabras/ secciones**	**Número de palabras**	**Promedio de palabras/ secciones**
Folha de São Paulo	Especial	4313	215	4450	222
	Mundo	4234	211	4706	235
	Opinión	3373	168	3980	199
Jornal do Brasil	Internacional	3734	186	5676	283
	Política	3791	189	4451	222
	Total	19445		23263	
	Promedios generales	3889	193	4652	232

I.2.4 Organización del *TeMário*

Teniendo en cuenta un entorno jerárquico donde los archivos pueden ser almacenados por el uso de Microsoft Windows, el programa se organiza en una sola carpeta con dos subcarpetas agregadas respectivamente, los "textos-fuente" y los resúmenes, como se muestra en la **figura I.1**.

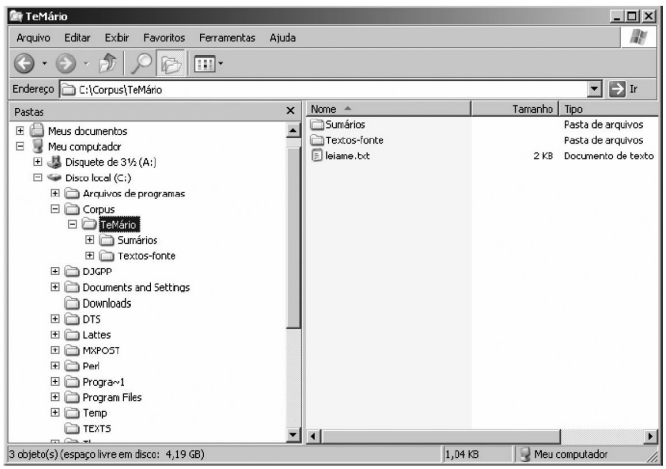


Figura I.1 Directorio *TeMário*

En la carpeta de "textos-fuente" hay tres carpetas organizadas (**figura I.2**):

a) La primera contiene los textos originales con sus títulos, organizados por sus fuentes; en otras palabras, los textos se agrupan por periódico (FSP o JB, como se muestra en la **figura I.3**) y sección (especial, mundo y opinión para los textos de la FSP, y política internacional, para JB), por un total de 60 textos de FSP y 40 de la JB;
b) La segunda contiene todos los "textos-fuente" con sus títulos, sin discriminación de origen;
c) La tercera contiene los "textos-fuente" sin ningún tipo de información de fuente o título.

Los archivos de texto están sin formato con extensión .txt ya que permite su procesamiento automático. Aparte de sus prefijos, todos los nombres de archivo incluyen el año (NN), el mes (AA) y el día de la publicación (del 1-31).[38] Los prefijos indican las secciones de periódicos correspondientes como sigue:

- "textos-fuente" de la sección especial de la **FSP tienen el prefijo "ce"** (de la sección especial);
- "textos-fuente" de la sección de **FSP mundial tienen el prefijo "mu".**

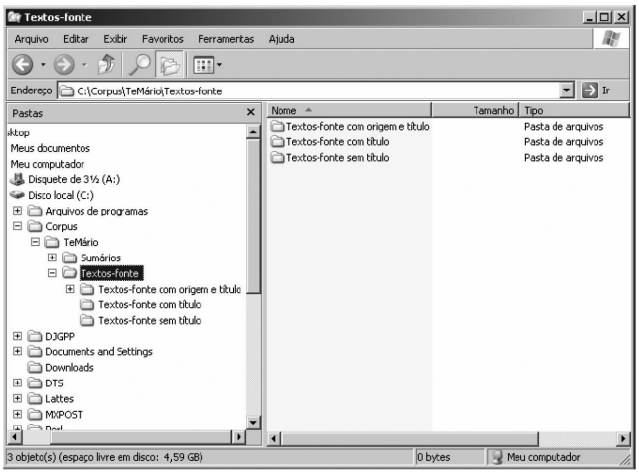


Figura I.2 Directorio "textos-fuente"

[38]NN para dos dígitos numéricos y AA para las dos primeras letras del mes correspondiente.

- "textos-fuente" de la sección de opinión de la **FSP tienen el prefijo "op**";
- "textos-fuente" de la sección **JB Internacional tienen el prefijo "in";**
- "textos-fuente" de la sección de Política **JB tienen el prefijo "po**".

Así, por ejemplo, el nombre de archivo 'in96fe29-a.txt' indica un texto de sección JB Internacional, publicado el 29 de febrero de 1996; el archivo 'mu94ag07- b.txt 'indica una sección de texto de FSP Mundial, publicado el 7 de agosto de 1994. Además, los "textos-fuente" tienen sus archivos sin título identificados **por "ST-"** antes de los prefijos anteriormente mencionados.

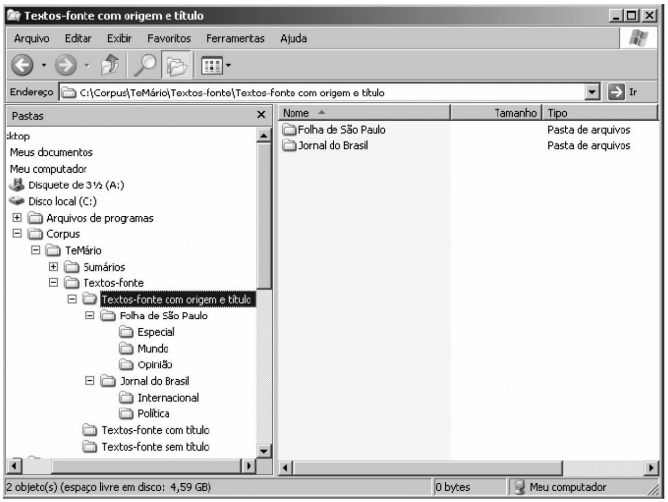


Figura I.3 Directorio de "textos-fuente" completos

Esta subdivisión del texto-fuente en las carpetas de diferente naturaleza tiene el propósito de ayudar a la recuperación de los datos por intereses específicos. Por ejemplo, si el objetivo es desarrollar un proceso de evaluación, los "textos-fuente" con título permiten asociar resúmenes o extractos generados automáticamente con el título para la comparación. En este caso, se puede considerar que el título es un representante legítimo de la idea principal del texto-fuente, que ha sido elegido por el escritor con el fin de verificar los resultados automáticos a preservar. Ya para los estudios lingüísticos, como la comprobación de las características particulares de un género o de dominio, un analista puede ser dirigido a la recuperación de libros específicos, los que ya indican una clasificación genérica.

Además, en el mismo proceso para el resumen, manual o automático, es conveniente que los "textos-fuente" sean visibles sin ningún título. No en el caso del resumen manual, de modo que no induzca el escritor a colocar información explícita relacionada con el título. En el caso automático, la razón es diferente: el resumen de un texto no incluye el procesamiento de su título.

En la carpeta de ***sumários*** también hay tres subcarpetas (**figura I.4**): una con resúmenes a mano (***sumários manuais, palabra en portugués***), otra con resúmenes manuales marcados (***sumários manuais marcados, palabra en portugués***), finalmente, otra con los extractos ideales producidos por el generador de extractos de ideas, como se ha descrito anteriormente.

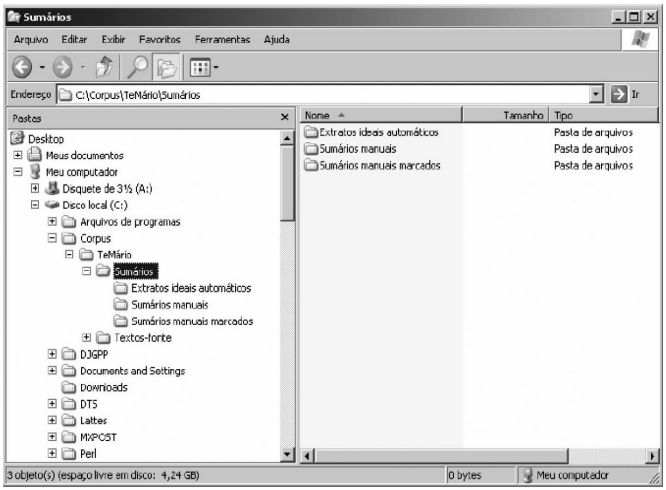


Figura I.4 Subcarpetas de la carpeta "Resúmenes"

En la carpeta de resúmenes manuales (*sumários manuais)* se encuentran los archivos sin formato (extensión .txt) que contienen los resúmenes construidos por expertos. Sus nombres contienen los nombres exactos de archivos de los "textos-fuente" correspondientes más el prefijo "Suma", para indicar que se trata de resúmenes en lugar de textos completos. La carpeta de resúmenes manuales marcados (*manuais marcados*) son el resumen de lo mismo, pero ahora son marcadas con color rojo las frases que indicaban el resumen profesional y las ideas principales de los "textos-fuente" correspondientes (Tarea 2 conforme a lo solicitado por el experto, ver apéndice del *TeMário*). Estos archivos se llaman

también "textos-fuente", pero con el prefijo "Summ-" (Resúmenes marcados manualmente). Su extensión .doc es precisamente para que el formato se conserve y, por lo tanto, las frases marcadas no se alteren. En consecuencia, se recomienda que los datos siempre sean preservados.

Para diferenciar los resúmenes manuales (prefijo "Sum") y los resúmenes marcados de forma manual (prefijo "Summ"), los archivos en la carpeta "Extractos de ideas automáticos" tienen el prefijo "Ext-" (éstos son también los archivos sin ningún tipo de formato, es decir, archivos con extensión .txt).

I.3 Consideraciones finales

En este anexo se describe el *TeMário*, el cual es un *corpus* de 100 artículos periodísticos más sus resúmenes manuales correspondientes y extractos de ideas. Los resúmenes manuales fueron construidos por un experto y escritor en portugués, mientras que los extractos de las ideas se producen automáticamente. Debido a la naturaleza específica del repositorio de datos, también contiene los textos originales con sus títulos y resúmenes manuales, y ahora se marcan con las ideas principales de los "textos-fuente" que guiaron las decisiones del resumen realizado por el experto. La generación de dicha información no representa una carga significativa para el experto humano y el mantenimiento de los títulos de los "textos-fuente" consta de una sola representación, desde el punto de vista del resumen automático que constituye un buen depósito de datos, tanto para los estudios comparativos en evaluación automática de los resultados como para la exploración de otras técnicas de resumen automático. Por ejemplo, los propios títulos pueden servir como base para la elección de los correspondientes segmentos de texto para componer un resumen: un título puede considerarse una frase esencial, como lo hace el *GistSumm* (Pardo *et al.*, 2003a), en este caso.

El uso potencial del *TeMário* se puede aumentar más cuando es considerado un lenguaje como XML (eXtensible Markup Language), conforme el proyecto LACIOWEB.[39]

En este caso, los archivos con la extensión .doc (que indican las ideas que guían las decisiones del experto) también se pueden convertir para la notación XML sin pérdida de significado, ya que en este lenguaje las etiquetas de estilo se conservan.

[39] http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Referencias bibliográficas (Anexo I)

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Martins, C.B. (2002). *UNLSumm: Um Sumarizador Automático de Textos UNL*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Relatório Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos.

Módolo, M. (2003). *SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português*. Dissertação de Mestrado. Departamento de Computação, UFSCar. São Carlos - SP.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003a). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* - PROPOR, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003b). NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. *Anais do IV Encontro Nacional de Inteligência Artificial*. Campinas-SP.

Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. *Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII*: *III Jornada de Minicursos de Inteligência Artificial (III MCIA),* pp. 203-245. Campinas-SP.

Salton, G. (1989) *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

Anexo J

Documentación del *Corpus TEXTRUSS*

En el anexo J se presenta la documentación del *corpus TEXTRUSS* creado para la tarea de GART para el lenguaje ruso. Se describen la estructura y composición.

J.1 Creación del *corpus TEXTRUSS*

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

ПОЛИТИКА (POLÍTICA)
БИЗНЕС (NEGOCIOS)
ОБЩЕСТВО (COMPAÑIA)
МНЕНИЯ (CRITICAS)
КУЛЬТУРА (CULTURA)
НАУКА (CIENCIA)
ТЕХНОЛОГИИ (TECNOLOGÍA)
НЕДВИЖИМОСТЬ (BIENES INMUEBLES)
АВТО (AUTO)
СТИЛЬ ЖИЗНИ (ESTILO DE VIDA)
СПОРТ (DEPORTES)

De cada categoría se obtuvieron 22 artículos de noticias. En total la colección contiene 242 artículos de noticias.

Las partes de la estructura de cada artículo son las siguientes (**figura J.1**):

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

J.2 Organización del *corpus*

El *corpus* contiene 3 formatos diferentes (**figura J.2**):

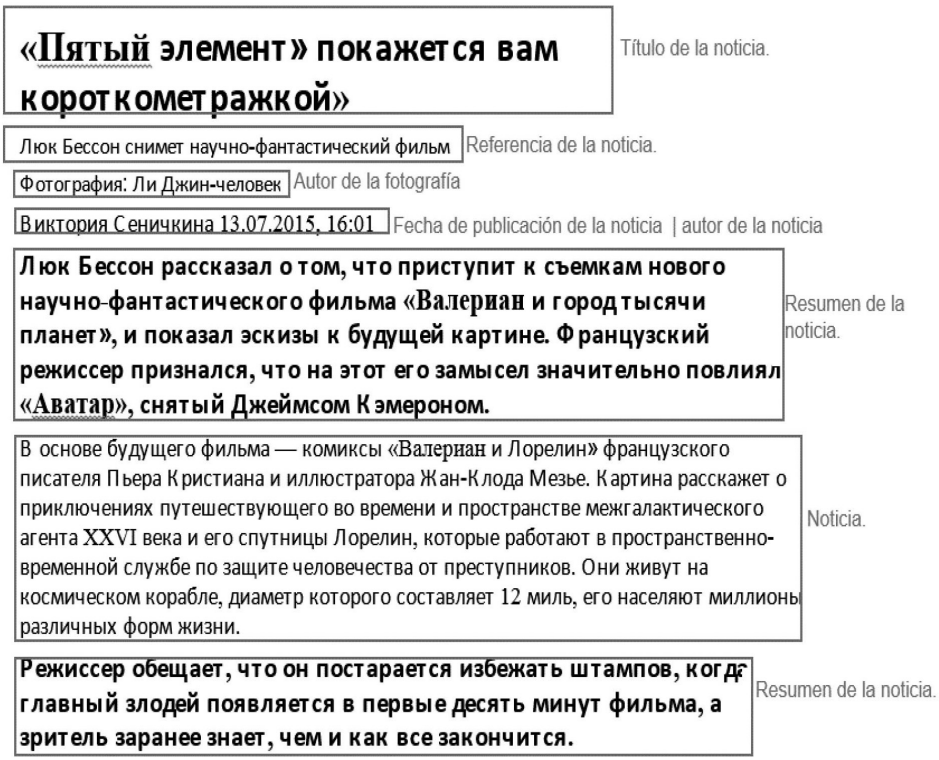


Figura J.1 Estructura del artículo 10CU140815_7654545.TXT de la colección del *corpus TEXTRUSS*

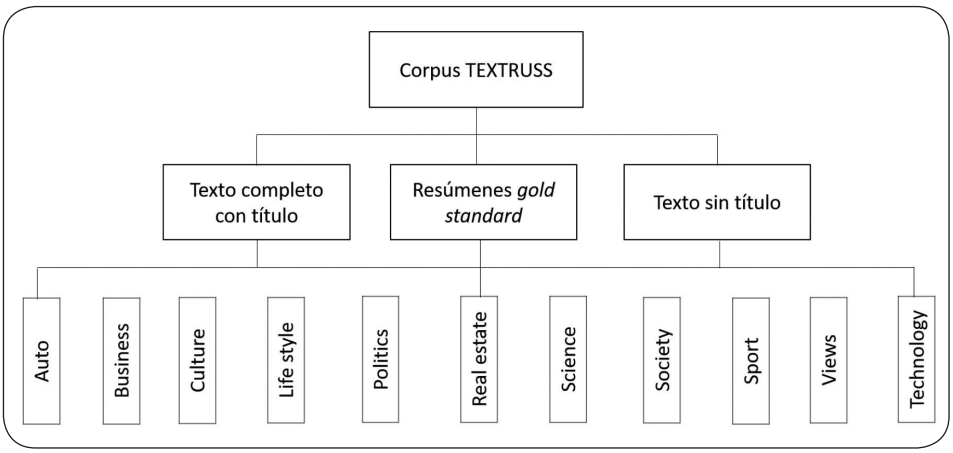


Figura J.2 Directorio del *corpus TEXTRUSS*

Detección de ideas principales y composición
de resúmenes en inglés, español, portugués y ruso.
60 años de investigación
de Griselda Areli Matias Mendoza,
Yulia Ledeneva y René Arnulfo García Hernández,
se terminó de editar el 28 de febrero de 2020
en Alfaomega Grupo Editor.

Revisión de pruebas (UAEM)
Osvaldo Renato Millán Zea

Este libro es una coedición entre la editorial Alfaomega Grupo
Editor y la Secretaría de Investigación y Estudios Avanzados
de la UAEMex, a través de la Dirección de Difusión y
Promoción de la Investigación y los Estudios Avanzados.

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso

60 años de investigación

Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación es un libro que aborda la tarea de generación automática de resúmenes desde la perspectiva cualitativa y cuantitativa. Primero se presentan los resultados de las pruebas de los test de *Turing* realizados a las máquinas que actualmente generan resúmenes de forma automática en los lenguajes más hablados y escritos: inglés, español, portugués y ruso, para saber si un resumen hecho por una máquina tiene la calidad para confundir a un humano y que no se dé cuenta que el resumen lo hizo una máquina. Posteriormente, se presenta la integración y el reporte cuantitativo de los métodos novedosos desarrollados hasta el momento y la comparación con los sistemas que generan resúmenes automáticos.

El libro está escrito en un lenguaje muy accesible por lo que cualquier persona puede leerlo, ya que a pesar de utilizar en algunas partes lenguaje técnico, éste se explica y se da el significado de cada término.

978-607-538-615-7

9 786075 386157

Computación
ÁREA

Nuevas tecnologías
SUBÁREA

www.alfaomega.com.mx

atencionalcliente@alfaomega.com.mx

Alfaomega Grupo Editor
Te acerca al conocimiento