



UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

**Generación automática de resúmenes en inglés y
español utilizando algoritmos genéticos**

Tesis

Para obtener el grado de

Doctora en Ciencias de la Computación

Que presenta:

M.C.C. Griselda Areli Matias Mendoza

Tutora académica:

Dra. Yulia Ledeneva

Tutores adjuntos:

Dr. René Arnulfo García Hernández

Dr. José Raymundo Marcial Romero

Resumen

Actualmente vivimos en una era digital, en donde la mayor parte de la información se encuentra en este formato, aunado a ello estamos en una época en donde el uso del internet y los documentos electrónicos es el día a día. Regularmente cuando se hace una consulta en internet sobre un tema en específico, se tiene como resultado una lista de páginas que hacen referencia al tema buscado. Sin embargo, para poder saber si la información es realmente de utilidad, se debe leer y analizar completamente. Por esto, surge la necesidad de contar con herramientas que permitan dar al humano la idea principal del texto, sin que este lo tenga que leer completamente, el cual ayude a contrarrestar los efectos negativos de la sobre carga de información, para que los humanos puedan obtener exclusivamente la información que realmente requiere, de forma más fácil y rápida, además que se encuentre en lenguaje que utilizan (español) o en el lenguaje universal (inglés).

La investigación sobre la generación automática de resúmenes en el lenguaje inglés tiene más de 60 años y los resultados obtenidos hasta el momento son aceptables, pero no lo mejores. Para la generación automática de resúmenes de tipo extractivo se consideran las características del texto. La característica de texto más estudiada hasta el momento es posición de las oraciones, existen muchos modelos propuesto en es el estado del arte para calcular esta característica. Sin embargo, no se sabe cuál es la mejor forma de calcularla.

Además, del lenguaje inglés hay otros lenguajes que están tomando relevancia en todos los ámbitos, sobre todo en internet, uno de ellos es lenguaje español. El español es la segunda lengua más hablada (por nativos) a nivel mundial y la segunda más utilizada en internet. El estudio de la generación automática de resúmenes en español no tiene un ámbito formal, por lo que no se puede determinar el grado de avance que se tiene en esta tarea.

Es por lo que en esta investigación se construye e implementa un método para la generación automática de resúmenes extractivos basado en un algoritmo genético, el cual es la base para hacer un estudio, para conocer la mejor forma de calcular la característica de texto, importancia de la posición de las oraciones. El método también se prueba en el lenguaje

español, lo cual contribuye a la actualización de la investigación sobre generación automática de resúmenes en este lenguaje.

La investigación de la generación automática de resúmenes se ha hecho de forma cuantitativa y existen muchos artículos y libros que muestra el avance que se tiene. Sin embargo, se ha dejado de lado el estudio cualitativo, es por lo que en esta investigación no solo se presentan los resultados cuantitativos, sino también los cualitativos. Para ellos se hace uso de la prueba del Test de Turing, el cual nos ayuda a conocer, si los resúmenes generados hasta el momento, por los métodos del estado del arte y los sistemas comerciales pueden confundir a un humano, al grado de pensar que el resumen es realizado por otro humano.

Índice General

Lista de tablas.....	14
CAPÍTULO 1. Introducción.....	16
Planteamiento del problema.....	19
Problema de investigación	20
Preguntas de apoyo	20
Objetivo.....	20
Objetivo General.....	20
Objetivos particulares	21
Hipótesis	21
Organización de la tesis	22
CAPÍTULO 2. Marco teórico	23
Inteligencia artificial.....	23
Generación automática de resúmenes de texto	24
Independencia del lenguaje	24
Independencia del dominio	25
Corpus.....	25
Evaluación de resúmenes automáticos.....	27
ROUGE.....	28
ROUGE-C.....	29
Jensen-Shannon divergence	30
Heurística en la generación automática de resúmenes.....	31
Baseline	31
Baseline:random.....	31
Topline	31
Concordancia.....	32
Preprocesamiento.....	32
Modelo de texto.....	33
Bolsa de palabras.....	33
N-gramas.....	34

Algoritmos genéticos	34
El algoritmo genético simple	34
Esquema general de un algoritmo genético.	35
Test de Turing.	36
Metodología	37
CAPÍTULO 3. Estado del Arte	38
Características del texto.....	38
Métodos en el lenguaje inglés	40
GA-4feature	40
MA-SingleDocSum.....	41
GA-Bag of words	41
GA-Multilanguage.....	42
TextRank	43
UnifiedRank.....	43
DE.....	44
FEOM	44
NetSum.....	44
CRF.....	44
Métodos del lenguaje español	45
Sistemas comerciales.....	49
Sistemas instalables	49
Sistemas en línea	49
CAPÍTULO 5. Método propuesto	51
Determinar la importancia de la posición de las oraciones	52
Entrada.....	52
Document Understanding Conference 2001 (DUC01)	52
Document Understanding Conference 2002 (DUC02)	52
Textos en Español para Resúmenes (TER).....	53
Preprocesamiento.....	53
Modelo de texto.....	54
Algoritmo genético	54
Modelos para determinar la importancia de las oraciones	57

Actualizar la investigación de resúmenes en español.....	59
Estudio de la generación automática de resúmenes para el lenguaje inglés y el español.....	59
Construcción del corpus "Textos en español para resúmenes (TER)"	62
Calcular las heurísticas al corpus TER	64
Probar los métodos del estado del arte y los sistemas comerciales para la generación automática de resúmenes.	64
Evaluación con ROUGE, ROUGE-C y JS divergence	65
Evaluación cualitativa.	65
Para el lenguaje español.....	65
Para el lenguaje inglés.....	68
CAPÍTULO 6. Experimentación.....	71
Experimentación y resultados del método para determinar la importancia de la posición de las oraciones	71
Modelo de texto.....	71
Probando los modelos para determinar la importancia de la posición de las oraciones.	74
Comparación con los métodos del estado del arte	75
Experimentación y resultados del método para la actualización de la investigación de la generación automática de resúmenes para el lenguaje español.....	76
Calcular el valor de las heurísticas	76
Con ROUGE	76
Con ROUGE-C	77
Con Jensen-Shannon divergence	77
Comparación con los métodos del estado del arte y los sistemas comerciales	77
Con ROUGE	78
Con ROUGE-C	78
Con Jensen-Shannon divergence	79
Matriz de ranqueo.....	80
Resultados del estudio cualitativo de la generación automática de resúmenes en inglés y español.	81
Para español.....	81

Para inglés.....	81
CAPÍTULO 5. Conclusiones.....	83
Conclusiones.....	83
Aportaciones	86
Publicaciones	87
Artículos.....	87
Libro	89
Referencias	90

Lista de Figuras

Figura 1. Representación gráfica de la heurística topline (Rojas, 2017).....	32
Figura 2. Diagrama de flujo del algoritmo genético.....	36
Figura 3. Diagrama de la metodología de investigación iterativa (Pratt, 2009).	37
Figura 4. Diagrama de la metodología propuesta	51
Figura 5. Representación gráfica del valor de la pendiente de la línea.	55
Figura 6. Método para la creación del corpus TER	62
Figura 7. Gráfica del mejor resultado por modelo de texto con DUC01.	72
Figura 8. Gráfica del mejor resultado por modelo de texto con DUC02.	73

Lista de tablas

Tabla 1. Lista de características del texto que se consideran para construir un resumen.....	39
Tabla 2. Descripción de los corpus de entrada.....	53
Tabla 3. Descripción de las etiquetas del texto.....	63
Tabla 4. Descripción de las etiquetas del resumen.....	64
Tabla 5. Resultados de f-measure para DUC01 con ROUGE-1.....	72
Tabla 6. Resultados de f-measure para DUC02 con ROUGE-1.....	73
Tabla 7. Comparación con diferentes modelos para determinar la importancia de oraciones con DUC01	74
Tabla 8. Comparación con diferentes modelos para determinar la importancia de las oraciones con DUC02.....	74
Tabla 9. Valor de F-measure en ROUGE-1 y ROUGE-2 de los métodos del estado del arte que prueban con DUC01	75
Tabla 10. Valor de F-measure en ROUGE-1 y ROUGE-2 de los métodos del estado del arte que prueban con DUC02.....	75
Tabla 11. Matriz de ranqueo global estimado para los métodos del estado del arte	76
Tabla 12. Resultados del cálculo de las heurísticas con ROUGE usando el corpus TER.	76
Tabla 13. Resultados del cálculo de las heurísticas con ROUGE-C usando el corpus TER.....	77
Tabla 14. Resultados del cálculo de las heurísticas con Jensen-Shannon divergence usando el corpus TER.	77
Tabla 15. Resultados evaluados con ROUGE para los métodos, sistemas y heurísticas usando TER.	78
Tabla 16. Resultados evaluados con ROUGE-C para los métodos, sistemas y heurísticas usando TER.	79
Tabla 17. Resultados evaluados con JS divergence para los métodos, sistemas y heurísticas usando TER.	79
Tabla 18. Ranking de los métodos y sistemas para la generación automática de resúmenes en español.	80

Tabla 19. Resultados del test de Turing con respecto al humano para el lenguaje español.....	81
Tabla 20. Resultados del test de Turing con respecto al humano para el lenguaje español.....	82



CAPÍTULO 1.

Introducción

Desde la antigüedad el hombre ha tenido la necesidad de comunicarse, esto lo podemos ver a través de las pinturas rupestres. En la época antigua no existía una forma de comunicación formal ni la tecnología que actualmente tenemos, por lo que el aprendizaje y las experiencias no podían ser transmitidas a nuevas generaciones; entonces, ¿qué pasaba con aquellos primeros cazadores, probablemente morían por falta del conocimiento previo. Una de las soluciones que encontraron los hombres de aquellas épocas fue pintar su conocimiento en paredes (pinturas rupestres), este conocimiento servía para enseñar y evitar posibles muertes. Pero ¿qué era lo que en realidad estaban haciendo?, aunque no lo sabían estaban haciendo un resumen de sus experiencias y vivencias. Como podemos observar el resumen surge como una necesidad de comunicación del ser humano. A lo largo de los años el conocimiento ha ido creciendo de tal manera que actualmente lo podemos ver plasmado en diferentes escritos, los cuales se encuentran en papel o en formato digital.

Ante el incremento acelerado de la información, ha surgido un gran interés por parte de los investigadores del Procesamiento del Lenguaje Natural (PLN), por investigar el área de generación automática de resúmenes con el objetivo de minimizar los problemas causados por el exceso de información. Sabemos que el objetivo de un resumen es transmitir

la información más importante de un documento (Ledeneva and García-Hernández, 2017). Actualmente el uso de los resúmenes se puede ver empleado en diferentes áreas. Por ejemplo, en noticias periodísticas, artículos científicos, videos, audios y en los últimos días para analizar la información de las redes sociales, entre otras.

Los resúmenes para su estudio se pueden clasificar: por su propósito (1), de acuerdo con su entrada (2) y por su estrategia de condensación (3).

1. Por su propósito los resúmenes se clasifican en: genéricos y orientados al usuario. Los genéricos están hechos para su uso en general, no tienen algún propósito específico. Los orientados al usuario, se deben adaptar a las necesidades del interesado, este tipo se puede generar de temas específicos.
2. Los resúmenes de acuerdo con su entrada se clasifican para, uno y múltiples documentos. Para un solo documento tienen como entrada un único texto (Ledeneva, 2008). Para múltiples documentos son aquellos que tienen como entrada más de un texto (García-Hernández et al., 2010) (Ledeneva, 2008).
3. Por su estrategia de condensación se clasifican, en abstractivo y extractivos (Berker, 2011a). Los abstractivos son aquellos generados a partir de la comprensión del documento y describen el contenido con palabras u oraciones que en algunas ocasiones no son encontradas en el texto original (da Cunha Fanego, 2005), (Qazvinian et al., 2008), (Ledeneva, 2008), (Montiel, R., 2009), (García-Hernández et al., 2009), (Plaza, 2011), (Suanmali et al., 2011a). Los resúmenes extractivos son generados a partir de la selección de frases clave, oraciones o párrafos considerados importantes del texto original; por lo que no requieren de la comprensión del documento (da Cunha Fanego, 2005), (Patel et al., 2007), (Qazvinian et al., 2008), (Ledeneva, 2008), (Montiel, R., 2009), (García-Hernández et al., 2009), (Plaza, 2011), (Suanmali et al., 2011a). En su mayoría los métodos del estado del arte y los sistemas comerciales que generan resúmenes utilizan los de tipo extractivos debido a la eficiencia y facilidad de uso que se ha obtenido de ellos.

Además de considerar el tipo de resumen a realizar según sus diferentes clasificaciones, se deben tomar en cuenta las características del

texto que se empleará para su realización. Las características del texto son el fragmento de una oración que se considera relevante para formar parte del documento original. Algunas de las características son: tener un alto parecido con el título, una gran longitud, contener datos numéricos o nombres propios, contener términos frecuentes, la posición de la oración en el texto, la cual es una de las más utilizadas, entre otras. Se sabe que las primeras oraciones del texto son más relevantes que el resto del texto (*baseline:first*), por eso se utiliza la posición de la oración (en orden inverso) para determinar su relevancia, lo que significa que la última frase prácticamente no tiene posibilidad de ser seleccionada. Existen diferentes formas de suavizar la importancia de las oraciones según la posición.

Otro de los aspectos importantes a considerar, es el lenguaje. Hasta hace algunos años la investigación de la generación automática de resúmenes estaba centrada en el lenguaje inglés, dejando de lado lenguajes igual de importantes.

El estudio de la generación automática de resúmenes comienza en 1968 con Lunh, y a lo largo de los años su estudio se ha enfocado al lenguaje inglés. La mayor parte de los recursos como son corpus, métricas de evaluación, concursos, entre otros, se encuentran en inglés. El lenguaje inglés es el tercero hablado a nivel mundial. Sin embargo, es el lenguaje que más países utilizan para comunicarse por lo que se le ha denominado el lenguaje universal. A pesar de que la generación automática de resúmenes se ha estudiado mucho para el lenguaje inglés, aun no se tienen resúmenes que iguallen a los hechos por los humanos en este lenguaje.

Otro de los lenguajes que actualmente ha tenido un realce es el lenguaje español. El lenguaje español es la segunda lengua nativa más hablada en el mundo y la segunda lengua de comunicación a nivel internacional. Sin embargo, a pesar de ser una lengua muy utilizada para la comunicación, no se han realizado estudios profundos de la generación automática de resúmenes para este lenguaje.

El estudio formal de la generación automática de resúmenes tiene más de 60 años de investigación y el avance que actualmente se tiene puede servir de base para el estudio en otros lenguajes. Por esto en esta tesis se hace un estudio profundo de la generación automática de resúmenes en inglés, para posteriormente realizar una actualización de la generación automática de resúmenes para el lenguaje español.

La investigación de la generación automática de resúmenes se ha enfocado en el estudio cuantitativo de los métodos del estado del arte y los sistemas comerciales que generan resúmenes. Sin embargo, se ha dejado de lado el estudio cualitativo, por lo que en esta tesis también se hace un estudio cualitativo de la generación automática de resúmenes para los lenguajes inglés y español. Para poder hacer el estudio cualitativo en la tarea de generación automática de resúmenes se puede hacer uso del test de Turing. El Test de Turing nos permite conocer si un resumen hecho por una máquina (computadora) puede pasar por uno hecho por un humano.

El estudio de la generación automática de resúmenes se ha abordado desde distintas técnicas, como, por ejemplo: estadísticas, grafos, reducción algebraica, redes neuronales, algoritmos genéticos, entre otros. Actualmente los métodos del estado del arte basados en algoritmos genéticos han tenido los mejores resultados. En la actualidad este tipo de métodos basados en estas técnicas buscan no solo generar resúmenes de manera más sencilla sino pretenden que sean útiles para los humanos.

Una de las necesidades de los humanos es contar con métodos que permitan generar resúmenes de documentos que están en el lenguaje que dominan (para nuestro caso el español) y en el lenguaje universal (inglés).

En esta tesis se aborda la tarea de generación automática de resúmenes para los lenguajes inglés y español. Para el lenguaje inglés se busca determinar la mejor forma de calcular la importancia de la posición de las oraciones, mientras que para el lenguaje español se busca hacer una actualización de la tarea de generación automática de resúmenes.

Planteamiento del problema

Ante el incremento acelerado de la información y la necesidad que tiene el ser humano por acceder a ella, surge la necesidad de contar con métodos que permitan mostrar el resumen con la información más importante de un documento. En la actualidad el lenguaje inglés es el tercero más hablado a nivel mundial, sin embargo, a pesar de estar en la tercera posición, es la lengua que más países utilizan (106 países). Actualmente el lenguaje inglés ocupa el primer lugar de los más utilizados en internet. Como se observa el lenguaje inglés es muy utilizado y con respecto al estudio de la generación automática de resúmenes es el más investigado. Sin embargo, los resultados obtenidos hasta el momento no

alcanzan el máximo resultado (topline) esperado, por lo que se busca proponer un método de generación automática de resúmenes que genere resúmenes más parecidos a los que realiza el humano. Además de que las investigaciones realizadas hasta el momento se han hecho de forma cuantitativa, dejando de lado la parte cualitativa, la cual nos permite ver si los resúmenes generados de forma automática en realidad son entendidos por los humanos. Otro de los lenguajes que han tomado relevancia en los últimos tiempos es el español. El lenguaje español es la segunda lengua nativa más hablada en el mundo y también tiene una importante participación en internet y las redes sociales; en la actualidad es la tercera más empleada en internet. A pesar de la importancia del lenguaje español a nivel mundial el estudio de la generación automática de resúmenes para este lenguaje es muy poca y no se ha hecho de forma formal, por lo que se busca realizar una actualización del estudio de la generación automática de resúmenes considerando las bases del estudio en el lenguaje inglés.

Problema de investigación

Construir un método que permita generar resúmenes automáticos extractivos independientes del lenguaje, que considere las características textuales del documento para uno documento usando algoritmos genéticos.

Preguntas de apoyo

¿Cómo determinar la importancia de la posición de las oraciones en la generación automática de resúmenes para el lenguaje inglés?

¿Cuál es el contexto de la generación automática de resúmenes para el lenguaje español?

¿Cómo medir cualitativamente un resumen generado de forma automática?

Objetivo

Objetivo General

Construir un método que permita generar resúmenes automáticos extractivos independientes del lenguaje, probado en los lenguajes inglés y español, considerando las características textuales del documento para uno

documento usando algoritmos genéticos. Además, hacer un estudio cualitativo de los resúmenes generados.

Objetivos particulares

- Determinar la importancia de la posición de las oraciones en la generación automática de resúmenes para el lenguaje inglés.
- Conocer la realidad del estudio de la generación automática de resúmenes para el lenguaje español.
- Medir cualitativamente los resúmenes generados automáticamente.

Hipótesis

Las hipótesis de investigación formuladas a partir de la pregunta de investigación y las preguntas de apoyo se listan a continuación.

H₁: Considerando que la característica del texto posición de las oraciones ha sido la más estudiada y se han propuesto diferentes maneras de calcularla. Si se genera un método que permita probar las diferentes formas propuestas entonces se podrá determinar la mejor forma de calcular la importancia de las oraciones.

H₂: Además, un enfoque evolutivo puede ayudar a encontrar una combinación adecuada de las oraciones que deben pertenecer al resumen.

H₃. Tomando en cuenta que el estudio de la generación automática de resúmenes tiene más de 60 años en el lenguaje inglés y hasta el momento se han alcanzado resultados aceptables, se puede tomar como base esta investigación para hacer la actualización de la investigación en el lenguaje español y así poder conocer la realidad del estudio de la generación automática de resúmenes en español.

H₄: El test de Turing fue creado para responder a la pregunta de ¿una máquina puede ser inteligente?, por lo que se puede usar para poder saber si un resumen hecho por una máquina puede pasar por hecho por un humano. De esta forma se puede obtener una forma cualitativa de medir los resúmenes.

Organización de la tesis

En este capítulo se presenta una introducción a la tarea de generación automática de resúmenes y se presenta la problemática sobre la misma. El resto de la tesis está organizado de la siguiente forma. En el capítulo dos, se presentan los conceptos usados en la generación automática de resúmenes, como son, independencia de lenguaje, preprocesamiento, modelo de texto, corpus, evaluación, heurísticas, algoritmo genético, entre otras. En el capítulo tres se presenta una tabla en donde se listan las principales características del texto, ocupadas para la generación automática de resúmenes, además se presentan los principales métodos del estado del arte para inglés y español, así como los sistemas comerciales de generación automática de resúmenes. En el capítulo cuatro se da a conocer el método propuesto para determinar la importancia de las oraciones, el método propuesto para la actualización de la investigación de la generación automática de resúmenes para español y los pasos para el estudio cualitativo. En el capítulo cinco se presentan los resultados experimentales del método propuesto para determinar la importancia de las oraciones, del método propuesto para la actualización de la investigación de la generación automática de resúmenes para español y del estudio cualitativo. En el capítulo seis se presentan las conclusiones, aportaciones y publicaciones.



CAPÍTULO 2.

Marco teórico

En el capítulo anterior se introduce al problema de generación automática de resúmenes. En este capítulo, se presentan los conceptos principales utilizados en este trabajo. Se da la definición de generación automática de resúmenes, independencia del lenguaje, se describe en que consiste el preprocesamiento y el modelado de texto, parámetros que se utilizaran en el método. También se definen los métodos de evaluación que posteriormente se ocuparan para validar la generación de los resúmenes.

Inteligencia artificial

Según Benítez (Benítez et al., 2014), la inteligencia artificial tiene como objetivo emular algunas de las facultades intelectuales humanas en sistemas artificiales. Refiriéndose a que, aplicando modelos teóricos a una maquina (computadora) esta sea capaz de ejecutarlos. Las áreas de estudio más frecuentes de la inteligencia artificial son: la robótica, sistemas expertos, problemas de percepción, aprendizaje, Procesamiento del Lenguaje Natural (PLN), entre otras (Coarite Choque, 2008).

En los últimos años el campo del PLN ha tenido un gran crecimiento debido al incremento de la información. Entre sus áreas de estudio están:

recuperación y extracción de información, minería de datos, traducción automática, sistemas de búsqueda de respuestas, generación de resúmenes automáticos, análisis de sentimientos, entre otros (Hernández and Gómez, 2013).

Generación automática de resúmenes de texto

Las primeras investigaciones sobre la generación de resúmenes automáticos datan de finales de los años 50 y comienzos de los 60 (Luhn, 1958a) (Edmundson and Wyllis, 1961). Sin embargo, a pesar de los años de estudio, aún hay mucho por hacer, debido al crecimiento de la información. Actualmente gracias al avance que se tiene en el ámbito computacional es más fácil utilizar algoritmos más sofisticados.

La generación automática de resúmenes se define como el proceso de destilar la información más importante de una fuente (o varias fuentes) para producir una versión abreviada destinada a un usuario (o conjunto de usuarios) determinado y para una tarea (o conjunto de tareas) determinada (Mani and Maybury, 1999).

Según Mendoza (2014) existen diferentes taxonomías para generar un resumen. De acuerdo a la forma en que se genera, puede ser de tipo abstractivo, (Lloret, 2011), (Nenkova and McKeown, 2012) o extractivo (Lloret, 2011), (Nenkova and McKeown, 2012). De acuerdo al número de documentos a procesar pueden ser para un documento (Ledeneva, 2008) (García-Hernández and Ledeneva, 2013) (Mendoza et al., 2014) o para múltiples documentos (Mendoza et al., 2014) (Wan, 2010) (Radev et al., 2004) (Villatoro E., 2007). Además, por lo que respecta al idioma del documento puede ser monolingüe (Ledeneva, 2008) (Mendoza et al., 2014) (García-Hernández and Ledeneva, 2013) o multilingüe (Patel et al., 2007), (Mihalcea and Tarau, 2005), (Last and Litvak, 2010), (Saggion, 2011).

Independencia del lenguaje

Antes de comenzar a hablar sobre las características o las técnicas que se utilizan en los métodos para la generación de resúmenes independientes del lenguaje se debe tomar en cuenta la definición que se le da en el estado del arte a independencia del lenguaje, multilingüe o plurilingüe.

Algunos sistemas como SUMMARIST (Hovy, E., 1999) generan resúmenes multilingüe, extrayendo las oraciones de un documento en diferentes lenguajes y traduciendo el resumen a distintos idiomas. Otros sistemas como NewsBlaster (Blair-Goldensohn et al., 2004) o los presentados por Bouayad-Agha (2009) y Saggion(2008) , realizan la traducción antes de efectuar la extracción de las oraciones; es decir, realizan un paso previo para traducir a un idioma común todos los documentos a resumir. Sin embargo, la mayoría de estos sistemas presentan un resultado pobre en cuanto a legibilidad y a calidad gramatical que refiere, debido principalmente al software utilizado para realizar la traducción automática.

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

Independencia del dominio

Cuando se habla de independencia de dominio, nos referimos al dominio temático y la estructura que maneja cada documento, por ejemplo, artículos periodísticos, artículos científicos, páginas web, twitters, cuentos, novelas, etc.

En su mayoría los métodos del estado del arte trabajan con dominios específicos (Mendoza et al., 2014) (Ledeneva, 2008) (García-Hernández and Ledeneva, 2013), en su mayoría trabajan con noticias. Sin embargo, hay algunos otros como el de Plaza (2011) que trabaja con tres dominios: Biomedicina, periodismo y turismo.

Corpus

Un corpus es un conjunto extenso de textos de diversas clases, ordenados y clasificados que sirven como base de una investigación (Corpus, 2014). En el área de generación automática de resúmenes los corpus pueden estar constituidos por textos de diferentes dominios como, por ejemplo, noticias, artículos científicos, textos literarios, de redes sociales, textos de cocina, entre otros y los resúmenes generados por el humano llamados *gold estándar* (resumen oro).

Para el lenguaje inglés los primeros corpus estándar creados fueron los de la conferencia *Document Understanding Conferences* (DUC). El objetivo de DUC es aumentar los avances de la investigación en la tarea de generación automática de resúmenes en el lenguaje inglés, realizando experimentos a gran escala ofreciendo utilizar los corpus y sus *gold standard*, para ellos se crearon 7 conferencias (DUC01, DUC02, DUC03, DUC04, DUC05, DUC06 y DUC07), cada una de ellas compuesta por varias tareas con sus respectivos corpus y *gold standard* (Over et al., 2007).

Como continuación de las conferencias DUC, en el año 2008, surgió la conferencia de análisis de textos llamada *Text Analysis Conference* (TAC). Esta conferencia está constituida por talleres de evaluación organizados para motivar la investigación en el PLN y las aplicaciones relacionadas. El objetivo 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 generación automática de resúmenes, siendo su principal área de estudio los resúmenes para multi-documentos enfocados al usuario.

En 2011, se creó la tarea MultiLing para evaluar métodos de generación de resúmenes independientes del lenguaje. Se crearon varios corpus MultiLing en 2011, 2013, 2015 y 2017 para la generación automática de resúmenes en varios lenguajes. La tarea de MultiLing ya trabaja con diferentes lenguajes. Sin embargo, los textos originales se recopilan en inglés y se traducen a diferentes idiomas, por lo que no existe un corpus real para cada idioma.

Por tanto, la creación de recursos para el lenguaje inglés es basta. Sin embargo, a pesar de los esfuerzos realizados para crear corpus estándar para la generación automática de resúmenes, el corpus más utilizado para el estudio de esta tarea en inglés es DUC02 (Matias Mendoza et al., 2020).

DUC02¹ fue construido con características específicas que lo hacen robusto y utilizable. Es un corpus de noticias en inglés sobre diferentes temas de tecnología, alimentación, política, finanzas, entre otros. Este corpus fue diseñado para la generación automática de resúmenes para dos tareas: múltiples documentos y un solo documento. Está compuesta por 567

¹ Para el acceso a los datos del corpus DUC02 se pueden obtener en la página <https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html>

documentos. Para cada documento del corpus se le crearon dos resúmenes por dos humanos expertos con una longitud mínima de 100 palabras. Está etiquetado lo que permite tener una separación clara de las oraciones.

Además de DUC02, el corpus DUC01² ha servido como referencia en los estudios de la generación automática de resúmenes en el lenguaje inglés. DUC01 es un corpus de noticias sobre desastres naturales en inglés. Este corpus fue diseñado para la generación de resúmenes múltiples documentos y de un solo documento. Está compuesto por 30 conjuntos de referencia y 30 de prueba, los archivos de prueba comprenden 309 documentos. Cada conjunto contiene los documentos originales, así como los resúmenes para un solo documento y múltiples documentos, generados manualmente (*gold standard*). Este corpus está etiquetado lo que permite tener una separación clara de las oraciones y por ende mejor manejo de la información con la que está constituido.

Para el lenguaje español, se han realizado investigaciones, pero ninguna de ellas utiliza corpus estándar o especializado para la tarea de GART. En el 2001 Acero, presenta un trabajo sobre generación de resúmenes personalizados para español utilizando su propio corpus (Acero et al., 2001). En el 2009 Téllez propone un corpus para la extracción de información para el lenguaje español, el cual ha sido adaptado para ser utilizado en la generación automática de resúmenes. (Téllez et al., 2009). También existen otras investigaciones sobre la generación automática de resúmenes en español (Da Cunha et al., 2009), (Toledo-Báez, 2010), (Plaza, 2011), (Molina, 2013), (Cabral et al., 2014), entre otros. Sin embargo, en cada una de las investigaciones se hace uso de diferentes recursos (corpus y métricas de evaluación), 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.

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. Originalmente, los métodos de evaluación para la generación automática

² Para el acceso a los datos del corpus DUC01 se pueden obtener en la página <https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html>

de resúmenes se procesaban manualmente, es decir, eran evaluados por humanos. Sin embargo, estos procesos consumían mucho tiempo y eran muy costosos, por lo que se desarrollaron métodos de evaluación automática para contrarrestar estas problemáticas.

Los métodos de evaluación se clasifican en intrínsecos y extrínsecos (Sparck Jones and Galliers, 1995). Los métodos intrínsecos 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 generados automáticamente contra los generados por los expertos, sin embargo, también se puede comparar con otro texto o el mismo texto original (Steinberger and Ježek, 2012). Los métodos de evaluación extrínseca estudian el resumen en el contexto de la tarea para la que fue generada (por ejemplo, evaluación de la relevancia), tratando de determinar el efecto del resumen en alguna otra tarea (Berker, 2011b).

ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) fue propuesto 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 resúmenes 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 generado por 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:

$$ROUGE - N = \frac{\sum_{O \in \{\text{Resúmenes De Referencia}\}} \sum_{grama_n \in O} cuenta_{coincidencia}(grama_n)}{\sum_{O \in \{\text{Resúmenes De Referencia}\}} \sum_{grama_n \in O} cuenta(grama_n)} \quad (1)$$

Donde n es la longitud del $n - grama$, y $cuenta_{coincidencia}(grama_n)$ 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 $S = (s_1, s_2, \dots, s_n)$ es una subsecuencia de otra secuencia $X = (x_1, x_2, \dots, x_m)$, si existe una estricta secuencia en aumento (i_1, i_2, \dots, i_k) de los índices de X tal que para todo $j = 1, 2, \dots, k$, existe $x_{i_j} = s_j$. Dadas dos secuencias X e Y , la Subsecuencia Común más Larga (SCL) de X e Y es la subsecuencia común con longitud máxima. 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-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 pueden aplicar para la evaluación de la calidad de los resúmenes generados automáticamente, ya que lograron el 95% de correlación entre juicios humanos. Para cada métrica del sistema ROUGE, se obtienen indicadores de Precisión, Recuerdo y F-measure.

ROUGE-C

ROUGE-C se presenta como una herramienta para evaluar resúmenes sin el resumen de referencia (*gold standard*) realizado por el humano. El método ROUGE-C alternativamente reemplaza los resúmenes de referencia por el documento fuente, así como con información centrada en consultas (en dado de caso de haber), por lo tanto, permite una forma totalmente manual e independiente de evaluar el resumen de varios documentos. ROUGE-C para evaluar los resúmenes de un documento se utilizan puntuaciones similares a las de ROUGE, por ejemplo, ROUGE-C-N, que es bastante similar con ROUGE-N, se define como se muestra a continuación.

$$ROUGE - C - N = \frac{\sum_{set\{PeerSummary\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{SourceDocument\}} \sum_{gram_n \in S} Count(gram_n)} \quad (2)$$

Donde n es la longitud del $n - gram$, $Count_{match}(gram_n)$ es el número máximo de $n - grams$ que co-ocurren en el resumen de pares y el documento fuente. ROUGE-C-N es la proporción de gramas superpuestos en el total de

n-gramas del documento fuente. ROUGE-C-N es una medida relacionada con la precisión, el denominador de la ecuación ocurre en el lado de la prueba (He et al., 2008).

Jensen-Shannon divergence

Jensen-Shannon *divergence* es un método que evalúa el contenido de un resumen y no requiere de los resúmenes modelos hechos por los humanos (*gold standard*). Supone que la distribución de las palabras en el documento fuente y el resumen generado deben ser similares entre sí. Jensen-Shannon *divergence* es una medida que compara dos distribuciones de probabilidad de palabras: el texto del documento original (P) y el texto del resumen a evaluar (Q). La baja divergencia de los documentos de entrada en el resumen elaborado se toma como señal de un buen resumen. Dadas dos distribuciones de probabilidad sobre palabras ((P Y Q), la divergencia de Jensen-Shannon se define como:

$$D_{JS}(P||Q) = \frac{1}{2} \sum_w P_w \log_2 \frac{2P_w}{P_w+Q_w} + Q_w \log_2 \frac{2Q_w}{P_w+Q_w} \quad (3)$$

La medida se puede aplicar a la distribución de unidades en los resúmenes del sistema P y los resúmenes del sistema Q. El valor obtenido se puede utilizar como una puntuación ara el resumen del sistema.

La fórmula de divergencia Jensen-Shannon data la ecuación 3, se implementa con la siguiente especificación (ver Ec. 4) para la distribución de probabilidad de las palabras w.

$$Q_w \begin{cases} \frac{C_w^S}{N_S} & \text{if } w \in S \\ \frac{C_w^T + \delta}{N + \delta * B} & \text{otherwise} \end{cases} \quad (4)$$

Donde P es la distribución de probabilidad de las palabras w en el texto T y Q es la distribución de probabilidad de las palabras w en el resumen S; N es el número de palabras en el texto y el resumen $N = N_T + N_S$, $B = 1.5|V|$, donde V es el tamaño del vocabulario de los documentos C_w^T es el número de palabras en el texto y C_w^S es el número de palabras en el resumen. Para suavizar las probabilidades del resumen, se recomienda usar $\delta = 0.005$ (Torres-Moreno et al., 2010).

Heurística en la generación automática de resúmenes

La palabra “heurística” se traduce del lenguaje griego como hallar o inventar. La heurística propone reglas, procedimientos o técnicas de solución de problemas de como lo haría un humano para solucionar un problema. En la metodología científica, la heurística se aplica para resolver tareas para las que no se cuenta con un procedimiento algorítmico de solución (Polya and Zugazagoitia, 1965). A continuación, se explican las principales heurísticas usadas en la generación automática de resúmenes.

Baseline

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

Baseline:random

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

Topline

El *topline* consiste en obtener la mejor combinación de oraciones de todas las posibles. Lo que nos permite saber cuál es el máximo resultado al que podemos llegar al evaluar los resúmenes generados con la colección (Rojas J., 2017).

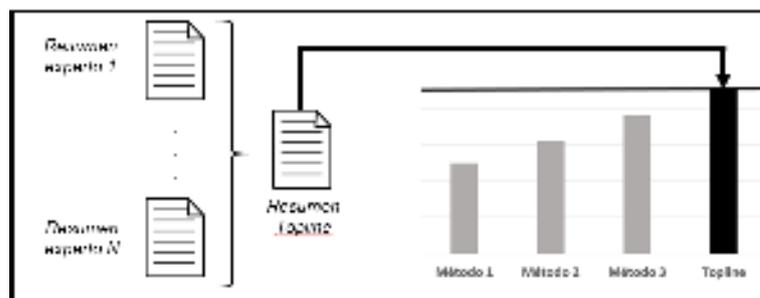


Figura 1. Representación gráfica de la heurística *topline* (Rojas, 2017)

Concordancia

La concordancia consiste en obtener la correspondencia o conformidad que existen entre los resúmenes realizados por los humanos (Mitrat et al., 1997), por lo cual esta heurística es solo informativa y no una referencia de evaluación.

Preprocesamiento

El preprocesamiento es la etapa donde se trata el texto de entrada para producir el texto de salida que se utilizarán en el método generador de resúmenes, el texto se transforma a una forma estructurada o semiestructurada de su contenido.

Al preprocesar los textos se pueden obtener representaciones sencillas que faciliten el análisis del texto. Para poder preprocesar el texto se pueden utilizar las siguientes etapas, limpieza, segmentación de oraciones, normalización, eliminación de palabras vacías (stop words), aplicación de stemming, etiquetado de texto, entre otras.

- Limpieza. Es la etapa en la que se limpia el texto de aquellas etiquetas que no son relevantes, pueden ser etiquetas HTML, fórmulas matemáticas, espacios, caracteres especiales, entre otros (Neri Mendoza, 2019).
- segmentación de oraciones. El texto es segmentado en oraciones (Alguliev et al., 2012) (Alguliyev et al., 2019).
- Normalización. El texto se normaliza eliminando los acentos de las palabras para su mejor manejo, se convierten las palabras en mayúsculas o minúsculas, según se requiera (Alcón and Lloret, 2015) (Ledeneva and García-Hernández, 2017) (Sidorov, 2019).

- Eliminación de palabras vacías (stop words). Eliminación de palabras muy repetitivas que no proporcionan información relevante, palabras que aparecen en más del 80% del documento no son consideradas y se llaman stopwords. Generalmente las candidatas son los artículos, preposiciones, conjunciones y pronombres. Para cada idioma se puede tener una lista de estas palabras consideradas vacías (Alcón and Lloret, 2015) (Mendoza et al., 2014) (Mosa et al., 2019).
- Aplicación de stemming. Stemming, es un método para reducir una palabra a su raíz (en inglés) a stem o lema, con el objetivo de eliminar prefijos, sufijos y de permitir la recuperación de los documentos que tienen variaciones sintácticas de los términos que se están recuperando. Los primeros algoritmos de stemming se desarrollaron para el idioma inglés, uno de los principales algoritmos para stemming es el algoritmo de Porter (Porter, 1980). Sin embargo, esta técnica puede ser adaptada a diferentes idiomas. Estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común (Goularte et al., 2019) (Mosa et al., 2019).
- Etiquetado de texto. El objetivo de esta etapa es etiquetar palabras o caracteres que tengan alguna importancia o significado especial y que sean de ayuda a la tarea, por ejemplo, cantidades, correos, etc (Matias Mendoza, 2013).

Modelo de texto

El modelado de texto consiste en seleccionar los términos que serán extraídos y convertirlos en un patrón que pueda ser analizado posteriormente. La diferencia entre modelos es el tipo de término que se extrae del documento. Para poder utilizar un modelo de texto primero se debe pasar por la etapa de preprocesamiento (Ledeneva and García-Hernández, 2017). Los modelos más utilizados son bolsa de palabras (Salton et al., 1975), n-gramas (Villatoro-Tello et al., 2006) y secuencias frecuentes maximales (SFM) (Ahonen-Myka, 1999), debido a su casi nula dependencia de lenguaje y dominio.

Bolsa de palabras

El modelo bolsa de palabras consiste en extraer todas las palabras diferentes de un texto. Se le conoce como bolsa de palabras debido a que se extraen las palabras y éstas no se encuentran ordenadas. Posteriormente, cada oración del documento es indexada, es decir representada por un

vector de términos, donde cada término corresponde a una palabra de la bolsa de palabras (Ledeneva and García-Hernández, 2017).

N-gramas

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

Sea una secuencia S de elementos ordenados $S_1 S_2 S_3 \dots S_K$ se denomina n-grama a cualquier subsecuencia $A = S_{i+1} S_{i+2} \dots S_{i+n}$ donde i es un valor entre 0 y $|S| - n$ para garantizar que la longitud de A sea siempre n o lo que es lo mismo $|A| = n; n > 1$ (Matias Mendoza, 2016).

Algoritmos genéticos

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

El algoritmo genético simple

Los pasos del algoritmo simple según (Kuri and Galaviz, 2002) son:

1. . Decidir cómo codificar el dominio del problema
2. Generar un conjunto aleatorio (población inicial) de N posibles soluciones codificadas al problema. A ésta se le llamará población actual.
3. Calificar cada posible solución (individuo) de la población actual
4. Seleccionar dos individuos de la población actual con una probabilidad proporcional a su calificación.
5. Lanza una moneda al aire* (con probabilidad p_c cae cara)

6. Si cayó cara mezclar los códigos de los dos individuos seleccionados para formar dos híbridos, a los que llamaremos a los individuos *nuevos individuos*
7. Si cayó cruz llamamos a los individuos seleccionados *nuevos individuos*
8. Por cada bit de cada nuevo individuo lanzar otra moneda al aire (con probabilidad p_m cae cara)
9. Si cae cara cambiar el bit en turno por su complemento
10. Si cae cruz el bit permanece inalterado
11. Incluir a los dos nuevos individuos en una *nueva población*
12. Si la población tiene ya N individuos, llamar la *población actual* y regresar al paso 3, a menos que se cumpla alguna condición de terminación.
13. Si no regresar al paso 4

Esquema general de un algoritmo genético.

Los algoritmos genéticos permiten procesar simultáneamente un conjunto de posibles soluciones de un problema dado. Los algoritmos genéticos trabajan con una representación de la solución llamados **individuos** (los cuales están compuestos por genes) y el conjunto de ellos forman una **población**, que es con la que trabaja el algoritmo genético. La población se va modificando a lo largo de las iteraciones del algoritmo que se denominan **generaciones**. A lo largo de cada una de las generaciones se crean nuevos individuos mediante los **operadores genéticos** de **crusa** y **mutación**. Cada generación incluye un proceso de **selección**, que da mayor probabilidad de permanecer en la población y participar en las operaciones de reproducción a los mejores individuos. Esta selección debe ser aleatoria para poder dar oportunidad a los individuos con menor adaptación de poder ser elegidos, aunque con menor probabilidad. Los mejores individuos son aquellos que tienen el mejor valor de la **función de aptitud** o **función de evaluación**, por esto son los que tienen la mayor probabilidad de sobrevivir y reproducirse.

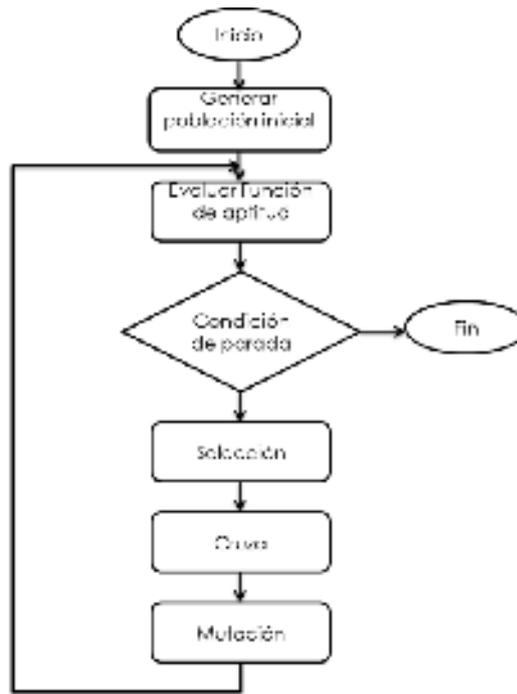


Figura 2. Diagrama de flujo del algoritmo genético.

Test de Turing.

El Test de Turing consiste en un juego de imitación, lo juegan 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”. El interrogador puede formular preguntas para A y B, todas las respuestas se dan de forma escrita para que la voz no ayude al interrogador a emitir una respuesta. La variante introducida por Turing consiste en substituir 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 él sepa que hay una máquina substituyendo a uno de los interrogados. La máquina podría pasar el test de Turing cuando el interrogador no logrará 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 le conoce como el padre de la inteligencia artificial.

Metodología

Pratt en el 2009 propone la metodología de investigación iterativa, diseñada especialmente para impulsar la generación de investigaciones en el campo de la inteligencia artificial (Pratt, 2009).

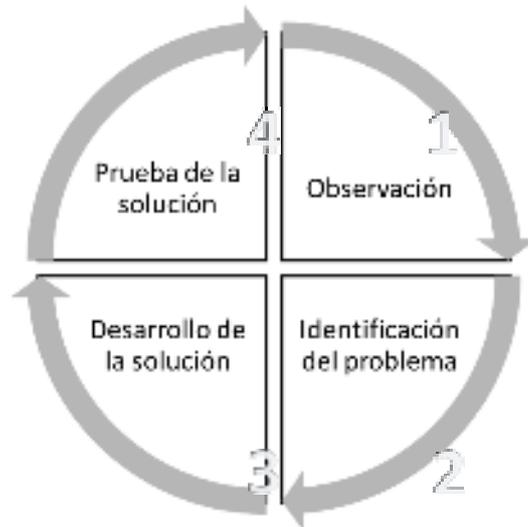


Figura 3. Diagrama de la metodología de investigación iterativa (Pratt, 2009).

La etapa de observación ayuda al siguiente paso de la metodología que es la identificación del problema de investigación, posterior a esto se resuelve el problema y finalmente se prueba la solución propuesta.



CAPÍTULO 3.

Estado del Arte

En el capítulo anterior se presentan los conceptos principales que nos permiten entender la tarea de generación automática de resúmenes. En este capítulo se presenta un estudio sobre las principales características del texto que se utilizan en la generación automática de resúmenes. Además, se enlistan los principales métodos de generación automática de resúmenes para el lenguaje inglés y para el lenguaje español.

Características del texto

La característica del texto es la parte de una oración que se considera relevante para formar parte del documento original. En la tabla 1, se muestra un listado de las principales características consideradas por los métodos del estado del arte para la generación automática de resúmenes.

Tabla 1. Lista de características del texto que se consideran para construir un resumen

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)	(Orasan, 2003)	(Berker, 2011b)	(Alfonseca and Rodríguez, 2003)	(Suanmali et al., 2011b)	(Qazvinian et al., 2008)	(Mateo et al., 2003)	(Babar and Patil, 2015)	(Kiyomarsi, 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
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

Como se puede observar una de las características del texto más utilizada es la posición de las oraciones, a pesar de que se ha estudiado por muchos métodos del estado del arte en el lenguaje inglés aún no se tiene la mejor forma de calcular esta característica, por lo que para este trabajo nos dimos a la tarea de hacer un estudio de las principales formas de calcular la importancia de la posición de las oraciones.

Métodos en el lenguaje inglés

En esta sección se presentan los métodos del estado del arte para la generación automática de resúmenes para el lenguaje inglés. Cabe mencionar que los trabajos descritos a continuación son con los que en el capítulo cinco se realiza la comparación.

GA-4feature

Vázquez (2018) presenta un método para optimizar la combinación de cuatro características. Las características consideradas son: similitud con el título (δ), longitud de las oraciones (γ) propuestas en (Mendoza et al., 2014), posición de las oraciones (β) y cobertura (α) propuestas por (García-Hernández and Ledeneva, 2013), el método está basado en un algoritmo genético, el cual considera el mejor peso que deben tener las características propuestas (Vazquez Vazquez et al., 2018). La función de aptitud usada se presenta a continuación.

$$fitness = w_1\alpha + w_2\beta + w_3\gamma + w_4\delta \quad (5)$$

El método es probado con las colecciones de documentos DUC01 y DUC02, las cuales son evaluadas con el método ROUGE.

MA-SingleDocSum

El método Ma-SingleDocSum propuesto por Mendoza (Mendoza Becerra, 2015) está basado en un algoritmo memético, enfocado en la generación de resúmenes para un solo documento. Además de utilizar operadores genéticos para la generación de los resúmenes utiliza la búsqueda local. Los parámetros que considera para la función de aptitud son: posición de las oraciones, relación de la oración con el título, longitud de la oración, cohesión y la convergencia (conocida como temática del texto).

Los pesos encontrados para la función objetivo son: $\alpha = 0.35$, $\beta = 0.35$, $\gamma = 0.29$, $\delta = 0.005$, $\rho = 0.005$; los cuales corresponden a las características de, posición (P), relación con el título (RT), longitud de las oraciones (L), cohesión (CoH) y cobertura (Cov), respectivamente. Para evaluar la calidad de un resumen representado por la representación de solución X_k , se requiere de una función objetivo, que se maximizara de acuerdo con la siguiente ecuación.

$$Max(f(X_k)) = \alpha P(X_k) + \beta RT(X_k) + \gamma L(X_k) + \delta CoH(X_k) + \rho Cov(X_k) \quad (6)$$

La función de aptitud propuesta en el trabajo de Mendoza se calculó como:

$$\alpha + \beta + \gamma + \delta + \rho = 1 \quad (7)$$

GA-Bag of words

El método propuesto por (García-Hernández and Ledeneva, 2013), es uno de los que han obtenido los mejores resultados en la generación automática de resúmenes para un solo documento en el lenguaje inglés. Está basado en un algoritmo genético y utiliza el modelo de texto bolsa de palabras. La función de aptitud utilizada en el trabajo de García toma dos características principales, la posición de las oraciones y la cobertura.

La característica de posición de las oraciones está determinada bajo la hipótesis de que las primeras oraciones son más importantes. Sin embargo, lo que busca este método es suavizar la selección de las oraciones, considerando que las primeras oraciones tienen más probabilidad de ser seleccionadas, sin dejar de lado las siguientes. 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 $t(i - x) + x$, donde $x = 1 + \frac{(n-1)}{2}$ 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:

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

La cobertura es considerada por el método de García considerando, que un resumen debe tener ideas diferentes y no debe ser repetitivo, pero a la vez contenga las palabras importantes. 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 número máximo de palabras. Por lo tanto, el resumen 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.

$$\beta = \frac{\sum_{p=\{word \in S\}}^m frequency(p,T)}{\sum_{q=\{word \in T\}}^m frequency(q,T)} \quad (9)$$

GA-Multilanguage

El método GA-Multilanguage propuesto por (Matias Mendoza, 2016) es independiente del lenguaje, está basado en un algoritmo genético, el cual utiliza el modelo de texto n -gramas (donde $n = 1, 2, 3, 4$ y 5). Utiliza dos características textuales principalmente, frecuencia de términos (δ) basada en el trabajo de (García-Hernández and Ledeneva, 2013) y la importancia de la posición de las oraciones (β), la cual se calcula mediante una regresión simbólica propuesta por (Vazquez Vazquez et al., 2019).

Para el lenguaje inglés se propone la siguiente fórmula para calcular la importancia de las oraciones.

$$\delta = \frac{(-49.8563 + 2 * X + 0.5 (-90.4102 + X) * X^{-X})}{\left(\left(\left(\frac{X}{14.7} \right)^{14.7} \right) - 92.5 \right)} \quad (10)$$

Para el lenguaje español se propone la siguiente fórmula para calcular la importancia de las oraciones.

$$\delta = \frac{(-28.7 - N)}{-57.4} \quad (11)$$

Para la función de aptitud aplica la siguiente fórmula

$$fitness = \beta * \delta \quad (12)$$

TextRank

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

UnifiedRank

Es un método que propone un enfoque novedoso unificado para el resumen simultáneo de un solo documento y de múltiples documentos. Las influencias mutuas entre las dos tareas se incorporan en un modelo de grafos

y las puntuaciones de ranqueo de una oración para las dos tareas se pueden obtener en un proceso de ranqueo unificado. Wan utiliza los corpus DUC para realizar su experimentación (Wan, 2010).

DE

Es un método de generación de resúmenes basado en la agrupación de oraciones. Utiliza un algoritmo de evolución diferencia discreto para optimizar la función de aptitud, seleccionando frases representativas de cada grupo. La selección de las oraciones del resumen se realiza bajo un esquema recursivo, que toma en cuenta el grado de pertenencia de cada oración al grupo correspondiente, midiendo la centralidad de cada oración al grupo al que pertenece, esto lo realiza en base a la distancia de Google normalizada (Alguliyev, 2009).

FEOM

Es un método que propone un modelo de optimización evolutiva difusa. En este enfoque, las oraciones se categorizan en términos de su contenido, y luego se selecciona la oración más importante para cada grupo. FEOM utiliza algoritmos genéticos para la generación de los vectores de solución con los grupos y aplica tres parámetros de control para regular la probabilidad de cruce y mutación de cada solución (Song et al., 2011).

NetSum

Es un método que utiliza el algoritmo de aprendizaje RankNet para entrenar un clasificador de oraciones, basado en pares y puntuar cada oración en el documento y así identificar las oraciones más importantes considerando un conjunto de características de cada oración. Este método realiza un resumen automático basado en redes neuronales. (Svore et al., 2007).

CRF

El método denominado CRF es un marco basado en campos aleatorios condicionales, toma el resultado de métodos anteriores como características y los integra a la perfección. Trata la tarea de generación de resúmenes como un problema de etiquetado de secuencia, entonces se considera que cada documento es una secuencia de oraciones y el procedimiento de resumen etiqueta las oraciones con 1 y 0. La etiqueta de una oración depende de la asignación de otros.

Métodos del lenguaje español

En esta sección se presentan los métodos del estado del arte que estudian la generación automática de resúmenes para el lenguaje español. Debido a las pocas investigaciones, y a la falta de estandarización en cuestión de recursos (corpus y métricas de evaluación), esta sección es solamente descriptiva.

- Generación automática de resúmenes personalizados (Acero et al., 2001). Es un trabajo que se encuentra enmarcado dentro del sistema Hermes³, un enviador personalizado de noticias que maneja información en inglés y en español. Para la realización del resumen su sistema ocupa tres heurísticas para la selección de frases.
 - a) Heurística de posición. Consiste en dar mayor puntuación a las cinco primeras frases de un texto.
 - b) Heurística de palabras clave. Consiste en extraer las M palabras más significativas de cada texto y comprobar a continuación, cuantas de estas palabras clave se encuentran en cada frase. De esta forma se asigna el mayor peso a las frases que contengan mayor número de palabras clave.
 - c) Heurística de personalización. Consiste en potenciar aquellas frases que tengan mayor relevancia para un modelo del usuario, con el fin de personalizar el resumen.

El corpus utilizado consta de 109 noticias obtenidas en la edición electrónica del diario ABC.

- Hacia un modelo lingüístico de resumen automático de artículos médicos en español (da Cunha Fanego, 2005). Se centra en el resumen automático de textos especializados en español, específicamente en el ámbito de la medicina. El corpus que utiliza consta de 20 artículos médicos en español que forman parte del Corpus Técnico del Instituto Universitario de Lingüística Aplicada (IULA) de la universidad de Fabra de Barcelona. El método que utiliza consta de cuatro etapas.
 - a) Selección de corpus de trabajo. El corpus seleccionado se divide en dos subcorpus, uno de referencia y uno de contraste.

³ Hermes aims at smart personalization in news delivery services by integrating techniques of automatic textual content analysis and user modeling with bilingual capabilities. This project has been partially funded by the Ministry of Science and Technology (PROFIT, 2000/020)

b) Análisis de los textos del subcorpus de referencia. Se analiza la estructura textual del artículo médico, sus unidades léxicas representativas y su estructura discursiva, sintáctica y comunicativa.

c) Desarrollo del modelo.

- Definición del modelo de resumen
- Desarrollo de las reglas lingüísticas
- Validación manual del funcionamiento de las reglas
- Implementación de las reglas
- Aplicación de las reglas sobre textos del subcorpus de contraste

d) Evaluación del modelo.

- Aproximación al resumen automático como herramienta de ayuda a la traducción jurídica en el ámbito del Derecho turístico (Toledo-Báez, 2010). Esta investigación está hecha para documentos en español del ámbito del derecho turístico. Sin embargo, no presenta ningún método para la generación automática de resúmenes, ya que solamente aplica la herramienta Copernic Summarizer ("Copernic Summarization-Technologies White Paper," 2003) para generar los resúmenes que posteriormente sirven para realizar una traducción.
- Plataforma para el resumen independiente del lenguaje (Cabral et al., 2014). Se propone una plataforma de generación de resúmenes independiente del lenguaje que proporciona adquisición de corpus, clasificación de idiomas, traducción y resumen de texto para 25 idiomas diferentes. Si el texto está en inglés, se envía al módulo de resumen extractivo, que selecciona las oraciones más significativas del texto original después del preprocesamiento utilizando los diversos métodos de puntuación de oraciones descritos en la literatura reconocidos como los más eficientes para el resumen extractivo. De lo contrario, el texto se somete a algunos algoritmos de resumen independientes del idioma y a varias herramientas que traducirán cada una de las oraciones del texto original al inglés. Dado que es probable que el proceso de traducción automática introduzca pérdidas semánticas en el texto original, el uso de más de una herramienta de traducción puede compensar dichas pérdidas. Las diversas versiones del texto traducido se envían al módulo de resumen extractivo, obteniendo para cada entrada un conjunto de números, cada uno de ellos relacionado con las oraciones del texto original. Los diferentes conjuntos de oraciones elegidos son analizados

por el Módulo de Puntuación y Selección de Oraciones, que producirá un nuevo conjunto de índices que corresponden al resumen, abarcando las oraciones elegidas en el texto original. El corpus que utiliza es CNN-Español, la versión actual de este corpus presenta 400 textos asignados a 08 categorías: deportes, entretenimiento, mundo, nacional, opinión, tecnología, viajes y noticias de salud.

- Generación automática 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 se le da al algoritmo de aprendizaje son pares (documento, resumen). De tal forma que el problema de generación resúmenes se convierte en una actividad de aprendizaje supervisado. Para la experimentación con el lenguaje en español se utiliza el corpus Desastres (Téllez et al., 2009). Cabe mencionar, que el corpus está diseñado para clasificación y fue adaptado para la generación de resúmenes. El conjunto de datos de Desastres consta de 300 noticias recolectadas de diferentes periódicos publicados en México. Cada una de las oraciones fue etiquetada utilizando dos etiquetas básicas: Relevante y No-Relevante. De forma que fuera posible evitar la subjetividad en el proceso de etiquetado, los expertos fueron instruidos para marcar como oraciones "Relevantes" solo aquellas que contengan al menos un hecho concreto sobre el evento sucedido. Por ejemplo, la fecha o el lugar en el que el desastre natural ocurrió, o el número de personas o casas afectadas, danos económicos, magnitud o escala del evento.
- Generación automática de resúmenes (Cardoso and Abelleira, 2013). Se propone un método basado en técnicas de aprendizaje supervisado, de forma específica en clasificación. El corpus que utiliza está compuesto por más de 8000 documentos que contienen 9 años de resoluciones rectorales de la Universidad Católica de Salta. El método usa un proceso de etiquetado para determinar si las oraciones son o no relevantes. Además, cada oración debe tener una etiqueta de si debe o no pertenecer al resumen. Para los experimentos utilizaron la herramienta de software Weka, que incluye una amplia colección de técnicas de clasificación. Entre los clasificadores utilizados por este método están: ADTree, ID3, C4.5 con poda, C4.5 sin poda, Tabla de decisión, Ripper and

NaïveBayes. La construcción de árboles de decisión obtiene resúmenes de calidad adecuada, que sirven como resúmenes indicativos para el usuario de un buscador semántico en el corpus propuesto.

- Un nuevo enfoque de resumen automático entre idiomas basado en la energía textual (Careaga-Moya et al., 2012). Este método propone un sistema de resumen en varios idiomas que utiliza la energía textual y la medición del tiempo de traducción. Esto hace que los resúmenes finales de las noticias sean más fiables. El método de resumen automático que utiliza energía textual implica que toma un método inspirado en la física estadística y lo combina con un modelo de espacio vectorial VSM y redes neuronales. El método ENERTEX (Fernández et al., 2007) considera las palabras de un texto como conjuntos de unidades, que interactúan entre sí y se ven afectadas por el campo que genera cada una de ellas. Así, cada palabra obtiene una puntuación según su energía textual. Este enfoque considera el tiempo de traducción de cada enunciado. Luego genera una matriz energética textual que contribuirá a la generación del resumen. La evaluación del sistema se realizó con el marco FRESA, considerando resúmenes de línea base creados automáticamente para diferentes porcentajes de los textos originales.
- Uso de grafos semánticos en la generación automática de resúmenes y estudio de su aplicación en distintos dominios: Biomedicina, Periodismo y Turismo. El trabajo de (Plaza, 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 diferentes dominios y con unas características de estructura y estilo muy dispares: artículos científicos de biomedicina, noticias periodísticas y páginas web de información turística en el lenguaje español. El método que utiliza está basado en el uso de grafos semánticos, el cual está constituido por las siguientes etapas: pre-procesamiento, traducción de las oraciones a conceptos, representación de las oraciones como grados de conceptos, construcción del grafo del documento, clustering de conceptos, asignación de oraciones a clusters, selección de oraciones para el resumen y finalmente la construcción del resumen.
- Compresión automática de fases: un estudio hacia la generación de resúmenes en español. El trabajo de (Molina, 2013) propone la generación de resúmenes automáticos para el lenguaje español considerando las siguientes características del texto. La segmentación discursiva, la cual consiste en representar el documento a través de un

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

Sistemas comerciales

En esta sección se listan los principales sistemas comerciales que generan resúmenes automáticos. Para el lenguaje español se prueban estos sistemas para saber cómo funcionan para este lenguaje en comparación con los métodos del estado del arte.

Sistemas instalables

- **Copernic Summarizer** (Copernic, 2015) es un software que fue desarrollado exclusivamente para la generación automática de resúmenes, lo cual hace que sea una herramienta flexible y adecuada porque ofrece las opciones que el resumen resultante sea del 5%, 10%, 25% o 50% de palabras del texto original; o resúmenes de 100, 250 y 1000 palabras sin importar el tamaño del texto original.
- **Microsoft Office Word** es una herramienta ofimática la cual podemos encontrar en las versiones de *Microsoft Office Word 2003* (Word, 2003) y *Microsoft Office Word 2007* (Word, 2007). Esta herramienta permite generar resúmenes de 10 o 20 oraciones; 100 o 500 palabras (o menos); o bien en porcentajes de 10%, 25%, 50% y 75% de palabras del documento original. Si algunos de los porcentajes no son adecuados, el usuario lo puede cambiar según sus necesidades.

Sistemas en línea

- **Summarizing** es una herramienta para la generación automática de resúmenes de manera rápida y eficaz. Su mayor uso es comercial, permite resumir un texto con tan solo colocarlo en el recuadro disponible en la página y seleccionar el número de palabras deseadas. La herramienta tiene las opciones de 100, 150, 200 y 300 palabras.
- **Text Compactor** es una herramienta gratuita en línea para la generación automática de resúmenes de grandes cantidades de información, fue creada por Keith Edyburn y está basada en la herramienta *Open Text*

Summarizer. Text Compactor permite generar resúmenes desde 0% hasta 100% del texto original

- **Open Text Summarizer (OTS)** es una aplicación libre de código abierto para resumir textos, que puede ser descargada de internet de forma gratuita (OTS, 2017), sin embargo, también puede encontrarse la interfaz de ésta en línea (Gohr, 2013). OTS genera resúmenes automáticos en diferentes porcentajes y es multilingüe.



CAPÍTULO 5.

Método propuesto

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

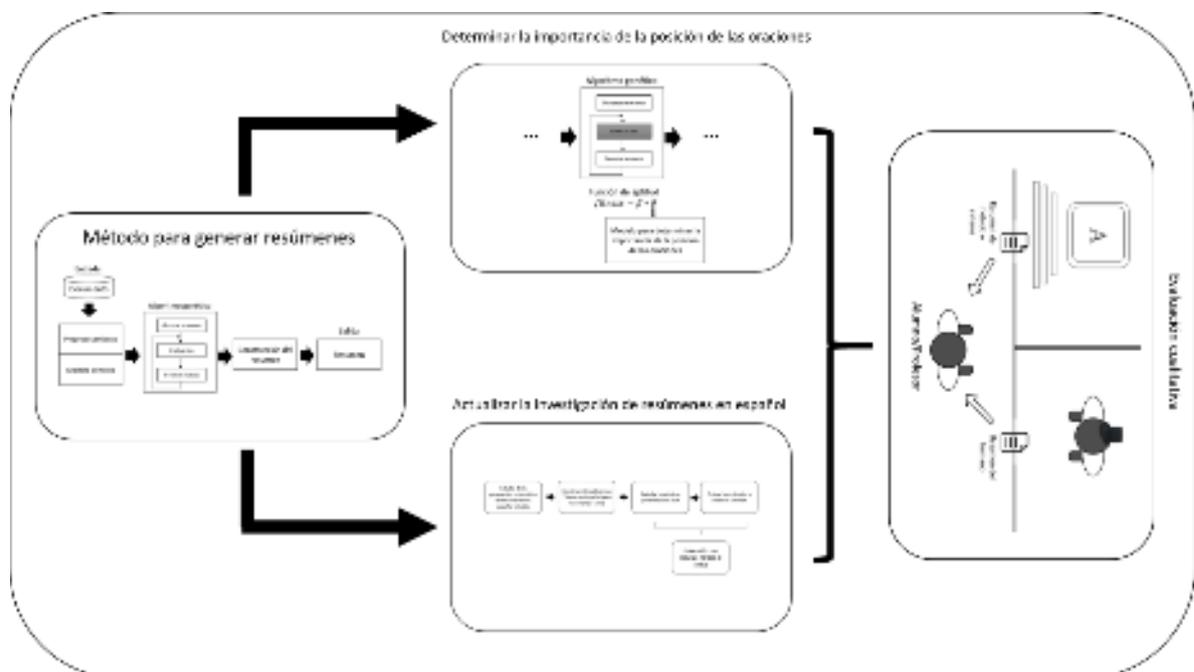


Figura 4. Diagrama de la metodología propuesta

Determinar la importancia de la posición de las oraciones

Entrada.

Como primera etapa del método propuesto se tiene la entrada, en esta etapa se obtienen los corpus de documentos con los que se va a trabajar. Para probar el método se utilizaron tres corpus, dos para el lenguaje inglés (DUC01 y DUC02) y uno para el lenguaje español (TER). A continuación, se describen.

Document Understanding Conference 2001 (DUC01)

DUC01⁴: Es un corpus de noticias sobre desastres naturales en inglés. Este corpus fue diseñado para la generación de resúmenes múltiples documentos y de un solo documento. Está compuesto por 30 conjuntos de referencia y 30 de prueba, los archivos de prueba comprenden 309 documentos. Cada conjunto contiene los documentos originales, así como los resúmenes para un solo documento y múltiples documentos, generados manualmente. Este corpus está etiquetado lo que permite tener una separación clara de las oraciones y por ende mejor manejo de la información con la que está constituido. Además, se cuenta con los resultados de diferentes medidas de baseline.

Document Understanding Conference 2002 (DUC02)

DUC02⁵: Es un corpus de noticias en inglés sobre diferentes temas de tecnología, alimentación, política, finanzas, entre otros. Este corpus fue diseñado para la GART para dos tareas: múltiples documentos y un solo documento. Está compuesta por 567 documentos. Para cada documento del corpus se le crearon dos resúmenes por dos humanos expertos con una longitud mínima de 100 palabras. Además, cuenta con los resultados de diferentes medidas de baseline. DUC02 es uno de los corpus más utilizados por los investigadores en el área de GART. Está etiquetado lo que permite tener una separación clara de las oraciones.

A continuación, en la tabla 2 se hace una descripción breve de los corpus DUC01 y DUC02.

⁴ Para el acceso a los datos del corpus DUC01 se pueden obtener en la página <https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html>

⁵ Para el acceso a los datos del corpus DUC02 se pueden obtener en la página <https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html>

Tabla 2. Descripción de los corpus de entrada

	DUC01	DUC02
Grupos	30	59
Documentos	309	567
Tarea	Tarea 1	Tarea 1
Longitud del resumen	100 palabras	100 palabras

Textos en Español para Resúmenes (TER)

TER⁶, es una colección de documentos compuesta por 240 noticias en el lenguaje español. El corpus TER está compuesto por noticias periodísticas adquiridos del periódico mexicano Crónica⁷, sobre 12 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:

- El corpus TER es creado a partir de noticias.
- Es para el lenguaje español.
- Tiene el fin de ser utilizado para la GART extractivos.
- Los resúmenes son para un sólo documento.
- Las noticias están en formato digital.
- La longitud de los resúmenes debe ser igual o mayor a 100 palabras.

Para la construcción de corpus se seleccionaron 20 noticias de las siguientes categorías: Academia, Bienestar, Ciudad, Cultura, Deportes, Espectáculos, Estados, Mundo, Nacional, Negocios, Opinión y Sociedad. Dando un total de 240 textos. Una de las consideraciones más importantes para la selección de las noticias fue que que tuvieran diferentes longitudes, pero siempre más de 100 palabras.

Preprocesamiento

Antes de utilizar el texto en el algoritmo genético, es necesario pasar por un preprocesamiento, en el que se le da al texto el formato adecuado. El preprocesamiento aplicado consiste en: (1) Limpieza, en esta etapa se limpia el texto de etiquetas, símbolos o caracteres especiales. (2) Segmentación de oraciones, cabe mencionar que esta etapa fue muy fácil ya que los corpus con los que se trabajó están etiquetados para identificar

⁶ <https://github.com/gmatiasm/Corpus-TER>

⁷ <https://www.cronica.com.mx/>

el principio y fin de las oraciones. (3) Normalización, en esta etapa se eliminaron los acentos de las palabras, se convirtieron a mayúsculas y se separaron por comas. (4) Eliminación de palabras vacías, en esta etapa se eliminaron las palabras que no brindan información y que no son relevantes (stopwords). (5) Aplicación de stemming, se aplicó el conocido algoritmo de Porter para colocar las palabras en su raíz. Finalmente (6) Etiquetado de texto, en esta etapa se etiquetaron las palabras que, si tienen significado especial y que se consideraran en algún momento de la tarea, como son cantidades y correos electrónicos.

Modelo de texto

Los modelos de representación de texto son una técnica que se basa e la extracción de los términos de un texto o documento. El modelo de texto consiste en extraer el termino (la unidad) a extraer y convertirlos en un patrón que luego se pueda analizar. La diferencia entre los modelos es el tipo de término que se extrae del documento. En el método propuesto se utiliza bolsa de palabras y n-gramas con $n = 1,2,3,4$ y 5.

Algoritmo genético

El método propuesto está basado en un algoritmo genético, a continuación, se explica cada una de sus etapas.

- Codificación de cromosoma. La codificación que se da al individuo para este problema de generación automática de resúmenes es binaria. Cada gen representa una oración del texto original. Si el gen es 1, significa que la oración es parte del resumen, de lo contrario, la oración no ha sido seleccionada para formar parte del resumen.

Si gen=1, oración seleccionada y si gen = 0, oración no seleccionada

S1 S2 S3 S4 S5 S6 S7

1	1	0	1	1	0	0
---	---	---	---	---	---	---

Oraciones 1,2,4 and 5 son seleccionadas

- Población inicial. Una vez determinada la representación de los individuos, se debe generar la población inicial, cabe mencionar que para la tarea de generación automática de resúmenes se debe considerar el número de palabras que debe tener el resumen para poder cumplir con este parámetro a la hora de generar la población inicial.
- Tamaño de la población inicial. El tamaño de la población inicial se determinó de acuerdo con el tamaño del documento de entrada. Se sabe que un individuo tiene tantos genes como frases tiene el texto. Para automatizar el cálculo de tamaño de la población inicial, se definió que el número de individuos (I) en una generación es el doble del número de oraciones (S).

$$I = \sum_{i=1}^n S_i(2) \quad (13)$$

- Función de aptitud. La función de aptitud implementada en este trabajo se basa en la función de aptitud propuesta en (García-Hernández and Ledeneva, 2013), donde se propone suavizar la posición de las oraciones calculando la pendiente de la recta (Ver. Fig. 3)

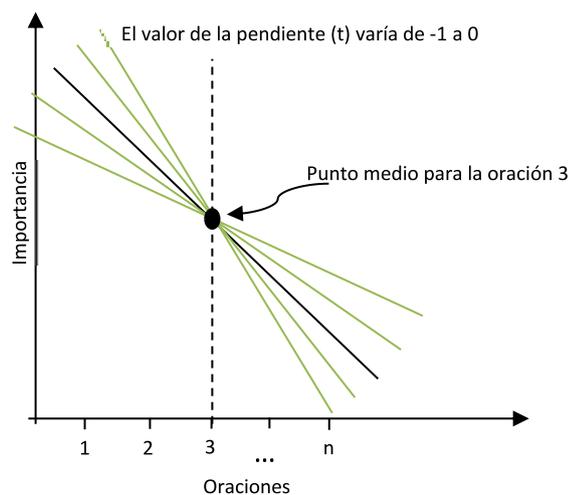


Figura 5. Representación gráfica del valor de la pendiente de la línea.

Los valores de la pendiente considerados para calcular la importancia de las oraciones fueron $t = -0.25, t = -0.3, t = -0.375, t = -0.45, t = -0.5, t = 0.55, t = -0.6, t = -0.625, t = -0.65, t = -0.7, t = -0.75, t = -0.8, t = -0.85, t = -0.9$. Cabe mencionar que estos valores se tomaron de forma aleatoria. Además de considera la característica de posición de las oraciones, también se considera la cobertura.

- Operador para la selección de padres. En este operador, los individuos de la población se seleccionan según su aptitud. El tipo de selección implementado en esta investigación es el operador ruleta. Con la selección por ruleta se pretende dar más posibilidades de ser seleccionados los individuos más fuertes (con mayor función de aptitud) y menos a los débiles.
- Operador de cruce. El operador de cruce que se utiliza es el cruce con prioridad de genes, el cual fue propuesto por (García-Hernández and Ledeneva, 2013). Este operador fue creado para la tarea de generación automática de resúmenes ya que esta tarea suele tener como parámetro cumplir con un cierto número de palabras, por lo que no se puede aplicar operadores de cruce comunes. Para crear un nuevo hijo se eligen los genes de ambos padres al azar, pero considerando solo aquellos con valor 1. De esta forma, si un gen tiene un valor de 1 en ambos padres, tiene más probabilidad de ser seleccionado para el cromosoma del hijo. Cada vez que se selecciona un gen en el cromosoma del hijo se revisa el número mínimo de palabras para el resumen.
- Operador de mutación. El operador de mutación que se utiliza, es el de mutación por doble inversión, el cual está basado en el trabajo de (García-Hernández and Ledeneva, 2013), así como la cruce para el operador de mutación se debe considerar el parámetro de número de palabras que contendrá el resumen. La mutación por doble inversión consiste en que el operador invertido se aplica dos veces al cromosoma hijo, pero la primera vez solo se consideran los genes con valor 1 para invertir el valor; en el segundo tiempo, se consideran para invertir solamente los genes con valor 0. Posteriormente se revisa el número de palabras en el resumen, si el número de palabras no corresponde al número de palabras especificado por el usuario, se invierte otro gen con valor 0, este proceso continúa hasta que el número de palabras especificadas por el usuario estén completas. La probabilidad de mutación utilizada es de 0.1%.
- Condición de parada. La condición de parada se refiere a la condición que se debe cumplir para que el algoritmo deje de evolucionar para posteriormente presentar la mejor solución encontrada. En esta tesis se considera la ecuación 14, propuesta por (Vazquez Vazquez et al., 2018).

$$maxGenerations = \sqrt{4 * NG * NS} \quad (14)$$

Donde *maxGenerations* es el número máximo de generaciones que se ejecutara el algoritmo para cada documento, que depende de su número de oraciones *NS* y la base numérica *NG*.

Modelos para determinar la importancia de las oraciones

El método propuesto fue probado con los diferentes modelos propuestos en el estado del arte para calcular la importancia de las oraciones. A continuación, se describen.

- **MSA1.** Propuesto en (Belkebir and Guessoum, 2015), utiliza una función discreta que da el valor cero (0) en caso de que la oración no sea la primera en el texto y uno (1) en caso contrario.
- **MSA2.** Propuesto en los trabajos (Bossard et al., 2008) y (Mendoza et al., 2014)

$$P = \sum_{\forall S_i \in \text{Summary}} \sqrt{\frac{1}{q_i}} \quad (15)$$

Donde q_i indica la posición de la oración S_i en el documento, y P es el resultado del calculo para todas las oraciones del resumen. En esta ecuación, P tiene valores altos cuando las oraciones en el resumen pertenecen a las primeras oraciones del documento y P tiene valores bajos cuando las oraciones pertenecen a las últimas oraciones del documento.

- **MSA3.** En el trabajo de (Ouyang et al., 2010), se proponen dos formas de calcular la importancia de las posición de las oraciones (ver ecuación 16 y 17).

$$(N - i + 1)/N \quad (16)$$

$$\frac{1}{i} \quad (17)$$

- **MSA4.** En el trabajo de (Lin, 1999) asigna puntuaciones más altas a las oraciones de los primeros cuatro párrafos.
- **MSA5.** En los trabajos de (Berker, 2011b), (Babar and Patil, 2015) y (Suanmali et al., 2011b) se consideran hasta 5 posiciones desde la parte superior del documento. Por ejemplo, la primera oración de un párrafo tiene una puntuación de 5/5; la segunda oración tiene una puntuación de 4/5, y así sucesivamente.

$$S = \frac{5}{5} \text{ for } 1^{\text{st}}, \frac{4}{5} \text{ for } 2^{\text{nd}}, \frac{3}{5} \text{ for } 3^{\text{rd}},$$

$$\frac{2}{5} \text{ for } 4^{\text{th}}, \frac{1}{5} \text{ for } 5^{\text{th}},$$

$$\frac{0}{5} \text{ for the other sentences}$$
(18)

- **MSA6.** Se proponen tres modelos para calcular la importancia de las oraciones (ver ecuación 19, ecuación 20 y ecuación 21). Los resultados mostrados en este trabajo están considerando la ecuación 20.

$$(S_i) = i$$
(19)

La puntuación de la oración es proporcional a su proximidad al final del documento; donde i es el número secuencial de la oración en el documento.

$$(S_i) = \frac{1}{i}$$
(20)

Como en (Ouyang et al., 2010) la puntuación de la oración es proporcional a su cercanía al comienzo de documento.

$$(S_i) = \max\left(\frac{1}{i}, \frac{1}{n-i+1}\right)$$
(21)

Finalmente, la puntuación de la oración es proporcional a su cercanía a los bordes del documento, donde n es el número total de oraciones.

- **MSA7.** En (Vazquez Vazquez et al., 2019) se realiza un estudio especializado para los corpus DUC01 y DUC02, para determinar la mejor forma de calcular la importancia de las oraciones de este conjunto de documentos. Para DUC01 la ecuación 22 y para DUC02 la ecuación 23.

$$f(i) = \frac{114X - 233X^2}{166 - 192X \left(\sqrt{X + \frac{1}{10325^X}} \right)}$$
(22)

$$f(i) = \frac{41.3294}{15.53 + X^2 - 2X + \frac{704.55}{X}} + \frac{X^2 + 4X - 663.61}{-2X^3 - 665.61X - \frac{31.06}{X}}$$
(23)

Actualizar la investigación de resúmenes en español.

La investigación de la generación automática de resúmenes hasta el año 2000 se centró en el idioma inglés, porque los recursos (corpus y medidas de evaluación estándar) estaban disponibles para este lenguaje. Sin embargo, otros idiomas más hablados tienen un crecimiento acelerado según (Fernández, 2018), por ejemplo, el español es el segundo idioma más hablado del mundo y el tercero más utilizado en Internet.

El problema es que no existe un corpus estándar con resúmenes generados por humanos y métodos de evaluación que estén altamente correlacionados con juicios humanos; por lo tanto, no existe un estudio formal de la generación automática de resúmenes en español.

En consecuencia, para poder actualizar el estudio de la generación automática de resúmenes de texto en español, es necesario conocer cómo ha avanzado el estudio de esta tarea para el idioma inglés a lo largo de los 60 años de investigación

Estudio de la generación automática de resúmenes para el lenguaje inglés y el español

El estudio de la generación automática de resúmenes tiene más de 60 años de investigación y en su mayoría se ha realizado en lenguaje inglés, los recursos (corpus, competencias y métricas de evaluación) creados para su estudio están en inglés.

La mayoría de los métodos del estado del arte han basado su estudio en los resúmenes extractivos, esto debido a su fácil implementación y sus resultados competentes. Los métodos de generación automática de resúmenes extractivos extraen las partes esenciales de un texto (oraciones, párrafos o frases clave) que se consideran importantes; por lo tanto, no requieren de métodos complejos o sofisticados (lingüísticos). El estudio de la generación automática de resúmenes tiene sus inicios a finales de los años 50s con el trabajo de (Luhn, 1958b). Luhn fue el primero en realizar un estudio sobre la generación automática de resúmenes. Posteriormente, la investigación siguió con (Edmundson, 1969), (Paice, 1990), (Kupiec et al., 1995), (Benbrahim and Ahmad, 1995), (McKeown and Radev, 1995), (Minel et al., 1997), (Marcu, 1997), (Jing et al., 1998), (Carbonell and Goldstein, 1998), (Barzilay and Elhadad, 1999), (Mani et al., 1999) entre otros. Hasta el año 2000, toda la investigación se enfocaba en el idioma inglés y se

realizaba sin tener un corpus estándar ni medida de evaluación, por lo que no se podía realizar una comparación de los métodos propuestos y por lo tanto no se sabía el grado de avance.

En el 2001, se crearon las conferencias Document Understanding Conference (DUC) con el objetivo de seguir avanzando en el estudio de la generación automática de resúmenes para el lenguaje inglés y permitir que los investigadores participaran a gran escala. Se crearon varios corpus DUC durante los años 2001 – 2007. DUC01 y DUC02 se centran en el resumen automático de texto para un solo documento y múltiples documentos; DUC03 a DUC07 enfocados a múltiples documentos con diferentes tareas. Como continuación de las conferencias DUC en 2008 se crea Text Analysis Conference (TAC), son una serie de talleres de evaluación creados para mejorar la evaluación de los sistemas. Los corpus TAC fueron creados en los años 2008 – 2011 y 2014, siendo su principal objetivo el estudio de los resúmenes para múltiples documentos enfocados al usuario final. En el año 2011, se creó la tarea Multiling enfocado en la evaluación de los algoritmos de resumen independientes del lenguaje. Se crearon varios corpus Multiling en 2011, 2013, 2015 y 2017 para la generación automática de resúmenes en varios lenguajes. Multiling considera varios lenguajes, los textos originales se recopilan en inglés y se traducen a diferentes lenguajes, por lo que no existe un corpus real para cada lenguaje. A pesar de que se han generado diversos recursos para el lenguaje inglés el corpus DUC02 es el más utilizado para el estudio de la generación automática de resúmenes (Matias Mendoza et al., 2020). DUC02 está construido con características específicas (dominio de noticias, etiquetado, tiene resúmenes modelo, maneja una longitud específica, se tienen el cálculo de las heurísticas baseline) que lo hacen robusto y utilizable.

Otro factor esencial para el estudio de la generación automática de resúmenes es el método de evaluación. Inicialmente, los métodos de evaluación se procesaban manualmente, es decir, eran evaluados por humanos. Sin embargo, estos procesos manuales eran costosos y consumían mucho tiempo. Posteriormente, se desarrollaron métodos de evaluación automática para reducir los costos presentados por los métodos manuales.

Los métodos de evaluación de los resúmenes se clasifican en dos categorías: intrínsecos y extrínsecos (Da Cunha et al., 2009). Para los métodos intrínsecos, tiene un texto de referencia, generalmente un resumen

creado por un humano (patrón oro). Sin embargo, también se puede utilizar otro texto o el mismo documento original (Steinberger and Ježek, 2012). Los métodos de la evaluación extrínseca determinan el efecto del resumen en otras tareas (por ejemplo, evaluación de la relevancia) (Berker, 2011b).

Actualmente, el método de evaluación intrínseca más utilizado es ROUGE. El método de evaluación ROUGE compara el resumen que se va a evaluar (resumen del candidato) con el resumen creado por el ser humano (resumen del modelo o resumen de referencia) (Lin, 2004). Debido a que ROUGE utiliza como referencia al resumen creado por el humano, la evaluación se realiza en relación con los criterios que el humano utilizó para generar el resumen. Además de ROUGE, existen otros métodos de evaluación intrínseca que no consideran los resúmenes modelos, si no que consideran como referencia el texto original, para hacer una evaluación más objetiva de la generación automática de resúmenes, se proponen probar los métodos de evaluación ROUGE-C y Jensen Shannon divergence (JS).

Desde la creación del corpus estándar DUC y la creación de métodos de evaluación automáticos, se ha podido conocer el avance de la generación automática de resúmenes en el lenguaje inglés, además, se han calculado diferentes heurísticas, entre ellas *baseline:random*, *baseline:first*, *topline* y concordancia. Las heurísticas han servido de referencia para conocer el avance que tienen los métodos del estado del arte y los sistemas comerciales de generación automática de resúmenes. Para el idioma español, estas heurísticas no se han calculado por falta de recursos.

Como se mencionó, la mayor parte de la investigación de la generación automática de resúmenes se realiza para el lenguaje inglés. Sin embargo, los métodos realizados y probados en inglés no son exclusivos de este lenguaje. Muchos de los métodos del estado del arte mencionan ser independientes del lenguaje (Mihalcea and Tarau, 2005), (Patel et al., 2007), (Litvak et al., 2010), (Saggion, 2011) y algunos otros, a pesar de no decir que son independientes del lenguaje, funcionan dentro de estructuras (extractivas) que les permiten trabajar con diferentes lenguajes. Los mejores métodos que se han realizado son los basados en gráficos (Mihalcea, 2004) y los basados en algoritmos genéticos, (García-Hernández and Ledeneva, 2013), (Mendoza et al., 2014), (Matias Mendoza, 2016), (Vazquez Vazquez et al., 2019).

Además de los métodos del estado del arte, actualmente también se encuentran disponibles sistemas comerciales para la generación automática de resúmenes. Los sistemas de generación automática de resúmenes son métodos disponibles para el público ya sea de forma gratuita o con algún pago, entre ellos están, Open Text Summarization, Text Compactor, Copernic Summarizer, Microsoft Office Word (MOW), Summarizing, entre otros.

Para el idioma español, se han realizado pocos esfuerzos en la investigación de la generación automática de resúmenes. El estudio de la generación automática de resúmenes en español empieza en 2001 con Acero, prácticamente 40 años después que el inicio de los estudios de generación automática de resúmenes en inglés. Acero, presenta la generación automática de trabajo de resúmenes personalizados utilizando su propio corpus, construido con noticias del diario ABC (Acero et al., 2001). Villatoro (2007) utiliza un corpus creado para la tarea de extracción de información y lo adapta para aplicarlo a la generación de resúmenes automático de texto de múltiples documentos para el idioma español (Villatoro E., 2007), (Téllez et al., 2009). También existen otras investigaciones sobre la generación automática de resúmenes en español como: (Da Cunha et al., 2009), (Acero et al., 2001), (Toledo-Báez, 2010), (Plaza, 2011) (Cabral et al., 2014), (Molina, 2013), entre otros. Sin embargo, a pesar de la investigación realizada para la generación automática de resúmenes en español, se desconoce el avance actual debido a que se han utilizado corpus propios o adaptados, lo que no permite una comparación entre los métodos y, en segundo lugar, a la falta de corpus estándar.

Construcción del corpus “Textos en español para resúmenes (TER)”

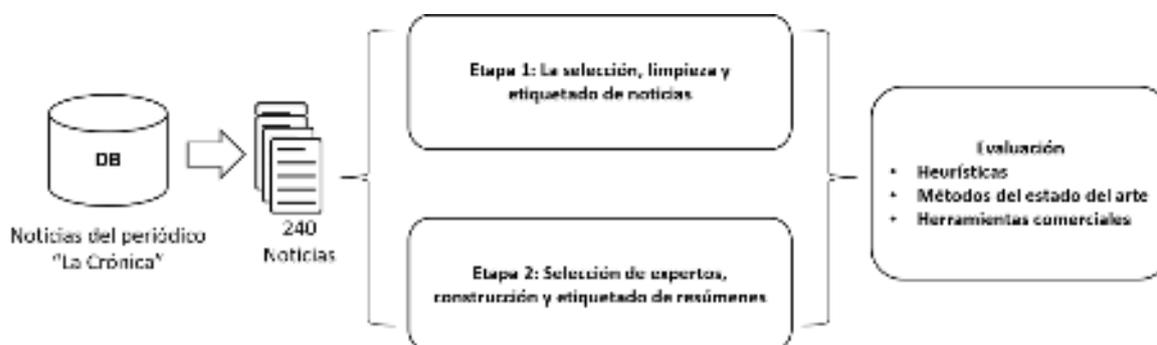


Figura 6. Método para la creación del corpus TER

TER es un corpus compuesto por noticias en español de México obtenidas del periódico “Crónica”. La construcción del corpus se divide en dos etapas, la primera para la selección, limpieza y etiquetado de noticias y la segunda para la selección de expertos, construcción y etiquetado de resúmenes.

En la primera etapa se seleccionaron al azar 20 noticias de las siguientes categorías: Academia, Bienestar, Ciudad, Cultura, Deportes, Entretenimiento, Estados, Mundo, Nacional, Negocios, Opinión y Sociedad, dando un total de 240 noticias. Los textos se limpiaron de etiquetas e imágenes extrayendo solo el título, la categoría, la fecha y el texto principal de la noticia. Posteriormente, se llevó a cabo una normalización de los textos, mediante el etiquetado de los textos. El etiquetado del texto ayuda principalmente a saber dónde comienza y termina una oración. De esta forma, se facilita su uso, y se garantiza que los métodos que lo utilicen utilizarán la misma separación de frases. Las etiquetas utilizadas se muestran en la Tabla

Tabla 3. Descripción de las etiquetas del texto

Tags	Description
<DOC></DOC>	Indica el inicio y final de documento
<DOCNO> </DOCNO>	Indica el nombre del documento
<FILEID></FILEID>	Indica un número único del documento
<TITLE></TITLE>	Indica el título del documento
<CATEGORY></ CATEGORY >	Indica la categoría a la que pertenece el documento
<DATE></DATE>	Indica la fecha de expedición del documento
<TEXT></TEXT>	Indica el texto de la noticia
<s></s>	Indica el inicio y fin de una oración

En la segunda etapa se crean los resúmenes hechos por los humanos (patrón oro). Para ello se seleccionó a un grupo de humanos de nacionalidad mexicana y educación mínima universitaria. Al humano se le entregó el texto separado por oraciones con el número de palabras que le correspondía a cada una de ellas para que solo leyera el texto y seleccionaran las frases que consideraban importantes. De las oraciones elegidas, se le pidió que creara un resumen extractivo de no menos de 100 palabras, esto lo hicieron dos humanos. Los resúmenes también fueron etiquetados para su mejor uso. A continuación, se describen las etiquetas utilizadas para los resúmenes.

Tabla 4. Descripción de las etiquetas del resumen

Tags	Description
<SUM></SUM>	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.

Finalmente se formó un corpus de 240 noticias en el lenguaje español mexicano, con dos resúmenes hechos por humanos para cada noticia. Cabe mencionar que el corpus fue construido considerando las características de DUC02.

Calcular las heurísticas al corpus TER

Calcular el valor de las diferentes heurísticas como *baseline:random*, *baseline:first*, *topline* y concordancia, permite tener una referencia del avance que tiene los métodos del estado del arte y los sistemas comerciales en la generación automática de resúmenes. Cabe mencionar que para esta actualización de la generación automática de resúmenes en español se está tomando como base lo ya investigado en el lenguaje inglés. Se sabe que la heurística *baseline:first* es un punto de referencia muy importante para el lenguaje inglés y que fue en el 2008 con el trabajo de Ledeneva que se logra superar esta heurística para este lenguaje, entonces tiene un poco más de 10 años que se logró superar (Ledeneva, 2008). Considerando lo anterior lo que se pretende para el lenguaje español es calcular esta heurística y superarla utilizando los métodos del estado del arte que pueden trabajar con el lenguaje español

Probar los métodos del estado del arte y los sistemas comerciales para la generación automática de resúmenes.

Actualmente, se sabe cuáles son los mejores métodos del estado del arte y los sistemas comerciales para la generación automática de resúmenes para inglés. Entonces, si se prueban con un corpus estándar de español y se mide su desempeño con diferentes métodos de evaluación, la investigación en español se puede actualizar 60 años después del inicio de la tarea de la generación automática de resúmenes. Además, se pueden

calcular las heurísticas que se consideran de referencia para comparar los métodos y los sistemas comerciales.

Evaluación con ROUGE, ROUGE-C y JS divergence

En la actualidad, el método de evaluación intrínseca más utilizado es ROUGE. Para hacer una evaluación más objetiva de la generación automática de resúmenes en el lenguaje español, se proponen otros métodos intrínsecos: ROUGE-C y Jensen Shannon divergence (JS). Estos dos evaluadores, a diferencia de ROUGE, utilizan el documento original como texto de referencia en lugar del resumen realizado por el humano, lo que les permite evaluar el desempeño de los métodos en todo el contenido del documento.

Evaluación cualitativa.

El estudio de la generación automática de resúmenes es una tarea que cumple 60 años de investigación, desde la publicación de (Luhn, 1958b). Existe un gran avance en las investigaciones sobre la generación automática de resúmenes en inglés, reflejado en libros especializados y artículos científicos, que se muestra a través de la calidad de los métodos y las técnicas de forma cuantitativa 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 una evaluación cualitativa que permita conocer el grado de avance de esta tarea a nivel de comprensión y lectura del humano. Para ello se aplica el test de Turing a las máquinas (métodos del estado del arte y sistemas comerciales) que actualmente generan resúmenes de forma automática, en los lenguajes inglés y español.

Para el lenguaje español

A partir de una noticia en español (del periódico mexicano “La Crónica”⁸) del corpus TER, se pidió a un conjunto de personas que

⁸ Periódico de alto renombre en México. Se puede acceder a el de manera electrónica por medio de la página web <http://www.cronica.com.mx/noticias.php>

identificaran cuáles de los resúmenes fueron realizados por humanos y cuáles por máquinas.

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 elaborar un diseño económico que cumpliera requisitos de la región, como la nieve, por ejemplo, desarrolló una idea sencilla pero práctica. “Si una 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 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 se le pidió tanto a los humanos como a las máquinas (métodos y sistemas) que generaran resúmenes con una longitud de 100 palabras, por lo cual, los resúmenes generados se recortaron con el objetivo de tener una evaluación objetiva. En específico hay que identificar de los resúmenes que se presentan cuáles son los 2 resúmenes 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...

Para esta prueba en lenguaje español del test de Turing los dos resúmenes que corresponden a los hechos por los humanos son los resúmenes 2 y 4.

Para el lenguaje inglés

A partir de una noticia en inglés del corpus DUC02, se pidió a un conjunto de personas que identificaran cuáles de los resúmenes fueron realizados por humanos y cuáles por máquinas.

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. Cabe mencionar que todos los resúmenes en inglés tienen la misma longitud de 100 palabras igual que en los resúmenes en español.

- 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...

Para esta prueba en lenguaje inglés del test de Turing los dos resúmenes que corresponden a los hechos por los humanos son los resúmenes 4 y 6.



CAPÍTULO 6.

Experimentación

En este capítulo, se muestran los resultados obtenidos de los experimentos realizados. Primero se presentan los resultados del método que determina la importancia de las oraciones, posteriormente se presentan los resultados de los experimentos hechos con el corpus TER a los principales métodos de estado del arte y los sistemas comerciales.

Experimentación y resultados del método para determinar la importancia de la posición de las oraciones

Para la evaluación se utiliza el método de evaluación ROUGE.

Modelo de texto

Los modelos de texto que se utilizaron para llevar a cabo la experimentación fueron bolsa de palabras y n-gramas con $n = 2, 3, 4$ y 5 . Los resultados obtenidos para cada uno de ellos se describen a continuación.

La tabla 5 muestra los resultados para diferentes valores de la pendiente t para los modelos de texto bolsa de palabras y n-gramas. De los resultados obtenidos para DUC01, el mejor resultado es con el modelo bolsa

de palabras con 0.45253. Con base en los resultados obtenidos, se puede concluir que para este conjunto de datos las primeras oraciones son más importantes que las del resto del documento, ya que el valor de la pendiente que se obtuvieron los mejores resultados oscila entre -0.7 y -0.9.

Tabla 5. Resultados de f-measure para DUC01 con ROUGE-1

Valor de (t)	Bolsa de palabras	n-grams			
		n=2	n=3	n=4	n=5
-0.25	0.43008	0.44135	0.44156	0.44430	0.44155
-0.3	0.43608	0.44388	0.43936	0.44308	0.44255
-0.375	0.43615	0.44176	0.43974	0.44240	0.44287
-0.45	0.43719	0.44494	0.44137	0.44374	0.44366
-0.5	0.43721	0.44452	0.44542	0.44448	0.44161
-0.55	0.43821	0.44738	0.44549	0.44285	0.44313
-0.6	0.43826	0.44813	0.44464	0.44338	0.44217
-0.625	0.43901	0.44728	0.44501	0.44329	0.44265
-0.65	0.43961	0.44840	0.44548	0.44328	0.44351
-0.7	0.44067	0.45032	0.44555	0.44400	0.44241
-0.75	0.44211	0.44653	0.44506	0.44562	0.44310
-0.8	0.44474	0.44757	0.44354	0.44272	0.44860
-0.85	0.44481	0.44823	0.44393	0.44440	0.44256
-0.9	0.45253	0.45095	0.44415	0.44361	0.44307

En la Fig. 5, se muestran gráficamente los resultados obtenidos para el corpus DUC01, en relación con el modelo de texto. Los mejores resultados se obtienen con bolsa de palabras y n-gramas con $n = 2$. También se puede ver que la línea de tendencia muestra que cuanto mayor es el valor de n , menor es el valor del *f-measure*.

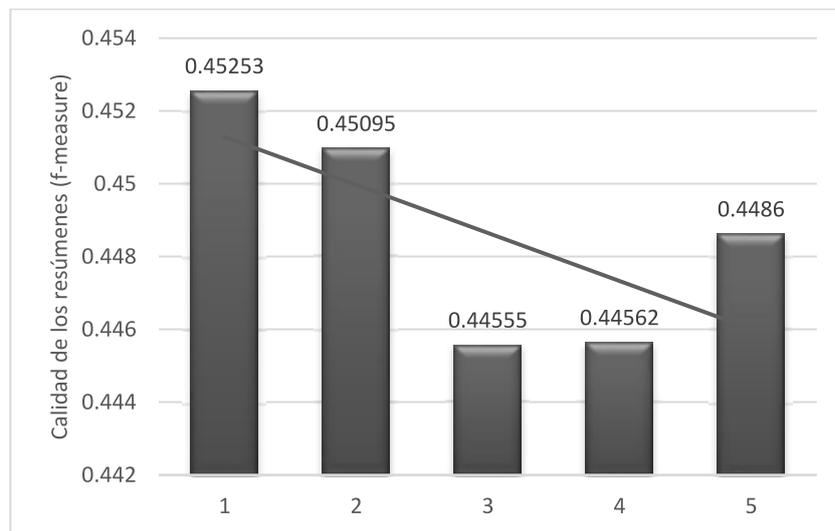


Figura 7. Gráfica del mejor resultado por modelo de texto con DUC01.

De la misma forma, se realizó el experimento para el conjunto de datos DUC02. Los resultados se muestran en la Tabla 6. Para DUC02, el mejor modelo de texto son palabras de bolsa con una medida f de 0.48183. A partir de los resultados obtenidos se puede concluir que para este conjunto de datos las primeras oraciones son más importantes que las del resto de documentos ya que el valor de la pendiente en la que se obtuvieron los mejores resultados oscila entre -0,75 y 0,9.

Tabla 6. Resultados de f-measure para DUC02 con ROUGE-1

Valor de (t)	Bolsa de palabras	n-grams			
		n=2	n=3	n=4	n=5
-0.25	0.47648	0.47101	0.47527	0.47677	0.47675
-0.3	0.47689	0.47246	0.47637	0.47641	0.47820
-0.375	0.47828	0.47669	0.47612	0.47663	0.47833
-0.45	0.47755	0.47681	0.47664	0.47704	0.47716
-0.5	0.47912	0.47804	0.47756	0.47601	0.47851
-0.55	0.47980	0.47693	0.47782	0.47659	0.47661
-0.6	0.47846	0.47721	0.47748	0.47683	0.47868
-0.625	0.47838	0.47556	0.47771	0.47612	0.47805
-0.65	0.47818	0.47670	0.47818	0.47717	0.47745
-0.7	0.47858	0.47772	0.47823	0.47641	0.47779
-0.75	0.48183	0.47853	0.47787	0.47628	0.47732
-0.8	0.48032	0.47894	0.47814	0.47721	0.47910
-0.85	0.48029	0.47731	0.47809	0.47786	0.47816
-0.9	0.47984	0.47844	0.47841	0.47767	0.47727

Para DUC02, es más claro ver que la tendencia hacia los mejores resultados se encuentra entre los modelos de texto bolsa de palabras y n-grama con n = 2 (ver Fig. 6).

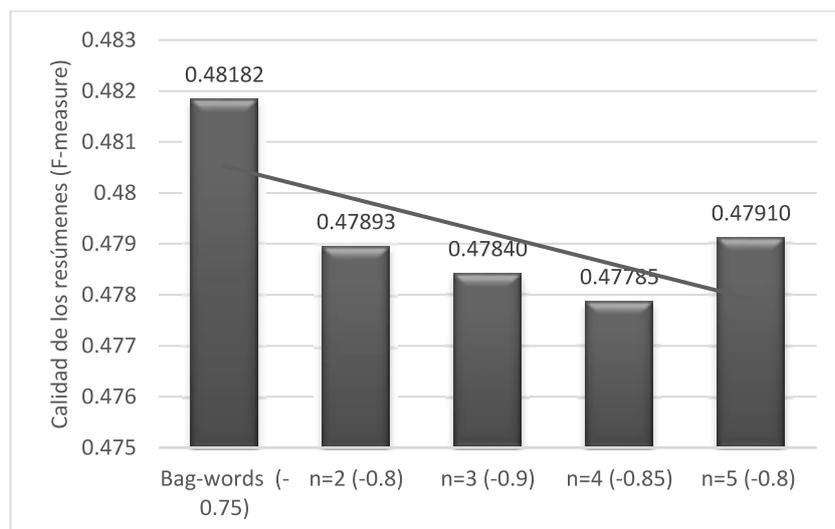


Figura 8. Gráfica del mejor resultado por modelo de texto con DUC02.

Probando los modelos para determinar la importancia de la posición de las oraciones.

En la tabla 7 se muestra una comparación de los diferentes modelos propuestos en el estado del arte para determinar la importancia de las oraciones en el corpus DUC01.

Tabla 7. Comparación con diferentes modelos para determinar la importancia de oraciones con DUC01.

Modelo	ROUGE-1	ROUGE-2
Proposed	0.45095	0.19763
MSA7	0.44940	0.19450
MSA3	0.44870	0.19358
MSA3	0.44654	0.18997
MSA4	0.44370	0.19279
MSA5	0.44309	0.19261
Baseline:first	0.44272	0.19701
MSA2	0.41280	0.14079
MSA1	0.41053	0.14078
MSA6	0.35803	0.09545

Para este corpus se puede observar que el método que considera suavizar la importancia de las oraciones es el que obtiene mejores resultados. Cabe mencionar que el modelo propuesto en EA7 fue construido especialmente para DUC01 y, aun así, el método propuesto obtiene mejores resultados. También se puede observar que tres de los modelos propuestos en el estado del arte no superan la heurística *baseline:first*.

En la tabla 8 se hace una comparación de los diferentes modelos del estado del arte para calcular la posición de las oraciones, para el corpus DUC02. El método propuesto es el segundo para este conjunto de datos. Sin embargo, el modelo propuesto en EA7 está especialmente construido para DUC02, además la diferencia entre el método propuesto y el primero no es relevante. Para DUC02, solo un modelo no logra superar la heurística *baseline:first*.

Tabla 8. Comparación con diferentes modelos para determinar la importancia de las oraciones con DUC02.

Model	ROUGE-1	ROUGE-2
MSA7	0.48470	0.22792
Proposed	0.48213	0.22333
MSA2	0.47877	0.22169
MSA3	0.47807	0.21983
MSA3	0.47788	0.21921
MSA6	0.47788	0.22208
MSA5	0.47773	0.22494
MSA4	0.47546	0.21889
Baseline:first	0.47294	0.22216
MSA1	0.44604	0.17432

Comparación con los métodos del estado del arte

A continuación, se muestra la comparación del método propuesto con los diferentes métodos del estado del arte que utilizan los corpus DUC01 y DUC02.

Tabla 9. Valor de F-measure en ROUGE-1 y ROUGE-2 de los métodos del estado del arte que prueban con DUC01

Método	ROUGE-1	ROUGE-2
Proposed	0.45095	0.19763
GA-4feature	0.45058	0.19619
FEOM	0.47728	0.18549
DE	0.47856	0.18528
NetSum	0.46427	0.17697
UnifiedRank	0.45377	0.17649
CRF	0.45512	0.17327

Tabla 10. Valor de F-measure en ROUGE-1 y ROUGE-2 de los métodos del estado del arte que prueban con DUC02

Method	ROUGE-1	ROUGE-2
UnifiedRank	0.48478	0.21462
GA-4feature	0.48423	0.22471
Proposed	0.48213	0.22333
DE	0.46694	0.12368
FEOM	0.46575	0.1249
NetSum	0.44963	0.11167
CRF	0.44006	0.10924

Considerando los resultados obtenidos parcialmente de las evaluaciones con ROUGE-1 y ROUGE-2 de los corpus DUC01 y DUC02, la tabla 11 muestra la posición de cada método obtenida por medio de la matriz de ranqueo, la cual fue calculada en base a lo propuesto en (Aliguliyev, 2009) (ver ecuación 24).

$$Ran = \sum_{r=1}^n \frac{(n - r + 1)R_r}{n} \quad (24)$$

Donde n es el número de métodos y sistemas involucrados en la comparación, y R_r se refiere al número de veces que el método o sistemas afecta la posición r -ésima.

Tabla 11. Matriz de ranqueo global estimado para los métodos del estado del arte

Método	R(r)							Global Ranking
	1	2	3	4	5	6	7	
UnifiedRank	1	0	1	0	1	1	0	2.4
GA-4feature	1	2	0	0	0	0	1	2.8
Proposed	1	1	1	0	0	1	0	2.8
DE	1	0	0	2	1	0	0	2.5
FEOM	0	1	1	1	1	0	0	2.5
NetSum	0	0	1	0	1	2	0	1.7
CRF	0	0	0	1	0	0	3	1

La matriz de ranqueo bis permite determinar cuál es el mejor método; en este caso tanto el método GA-4feature propuesto por (Vazquez Vazquez et al., 2018) y el método propuestos son los que obtienen mejores resultados.

Experimentación y resultados del método para la actualización de la investigación de la generación automática de resúmenes para el lenguaje español.

Calcular el valor de las heurísticas

Como ya se mencionó en el capítulo 5, para poder hacer la actualización de la investigación de generación automática de resúmenes en español, se consideran el cálculo de las heurísticas: *baseline:random*, *baseline:first*, *topline* y concordancia. A continuación, se muestran los resultados de cada una de ellas, obtenida con métodos de evaluación, ROUGE, ROUGE-C y Jensen-Shannon *divergence* usando el corpus estándar TER.

Con ROUGE

En la tabla 12 se muestran los resultados obtenidos con el método de evaluación ROUGE en sus configuraciones ROUGE-1, ROUGE-2 y ROUGE-SU4, para cada una de las heurísticas: *baseline:random*, *baseline:first*, *topline* y concordancia.

Tabla 12. Resultados del cálculo de las heurísticas con ROUGE usando el corpus TER.

Heurística	ROUGE - 1	ROUGE - 2	ROUGE - SU4
<i>Baseline:random</i>	0.4969	0.2933	0.3201
<i>Baseline:first</i>	0.7626	0.6229	0.6326
<i>Topline</i>	0.8344	0.7664	0.7649
<i>Concordancia</i>	0.6665	0.5432	0.5552

Con ROUGE-C

En la tabla 13 se muestran los resultados obtenidos con el método de evaluación ROUGE-C en sus configuraciones ROUGE-C1, ROUGE-C2, ROUGE-CL y ROUGE-CSU4, para cada una de las heurísticas: *baseline:random*, *baseline:first* y *concordancia*. Para este método no se calcula el valor del *topline*, ya que la comparación no se realiza con los resúmenes *gold-standard* sino con el texto original, por lo que el valor del *topline* se considera con valor a 1.

Tabla 13. Resultados del cálculo de las heurísticas con ROUGE-C usando el corpus TER.

Heurística	ROUGE - C1	ROUGE - C2	ROUGE - CL	ROUGE - CSU4
<i>Baseline:first</i>	0.2791	0.2756	0.2764	0.2699
<i>Baseline:random</i>	0.2538	0.2322	0.2475	0.2147
<i>Concordancia</i>	0.3709	0.3602	0.3684	0.3441

Con Jensen-Shannon divergence

En la tabla 15 se muestran los resultados obtenidos con el método de evaluación Jensen-Shannon *divergence* en sus configuraciones JS-SMT y JS-WSMT, para cada una de las heurísticas: *baseline:random*, *baseline:first* y *concordancia*. Al igual que para ROUGE-C, para este método no se calcula el valor del *topline*, ya que la comparación no se realiza con los resúmenes *gold-standard* sino con el texto original, por lo que el valor del *topline* se considera con valor a 1.

Tabla 14. Resultados del cálculo de las heurísticas con Jensen-Shannon divergence usando el corpus TER.

Heurística	JS-SMT	JS-WSMT
<i>Baseline:first</i>	0.7321	0.7107
<i>Baseline:random</i>	0.7105	0.6884
<i>Concordancia</i>	0.7886	0.7720

Comparación con los métodos del estado del arte y los sistemas comerciales

Para conocer el avance que hay en la investigación de la generación automática de resúmenes en el lenguaje español, se probaron los métodos del estado del arte que han dado buenos resultados para el lenguaje inglés y que dicen ser independientes del lenguaje. Además, se probaron diferentes sistemas comerciales que trabajan con español. Lo anterior para

poder comparar el avance que hay en los métodos de estado del arte con respecto a los sistemas.

Con ROUGE

En la tabla 15 se muestran los resultados obtenidos con el método de evaluación ROUGE, en donde se puede observar que todo los métodos del estado del arte y los sistemas comerciales, superan a la heurística base *baseline:random*. Sin embargo, solo un método supera a la heurística *baseline:first*. Como se mencionó anteriormente para el lenguaje inglés esta heurística tardó 40 años para poder superarse. Teniendo el lenguaje español 20 años de investigación se ha logrado superar la heurística *baseline:first*.

Tabla 15. Resultados evaluados con ROUGE para los métodos, sistemas y heurísticas usando TER.

Método/Sistema	ROUGE - 1	ROUGE - 2	ROUGE - SU4
<i>Topline</i>	0.8344	0.7664	0.7649
<i>GA-Multilenguaje</i>	0.7274	0.6289	0.6378
<i>Baseline:first</i>	0.7626	0.6229	0.6326
<i>GA-4feature</i>	0.7131	0.6072	0.6180
<i>GA-Bag of words</i>	0.6989	0.5852	0.5972
<i>MA-SingleDocSum</i>	0.6883	0.5706	0.5842
OTS	0.6761	0.5562	0.5698
Text Compactor	0.6749	0.5537	0.5678
<i>TextRank</i>	0.6606	0.5390	0.5532
Copernic	0.6187	0.4711	0.4898
MOW2007	0.6178	0.4691	0.4854
MOW 2003	0.6160	0.4649	0.4819
Summarizing	0.5775	0.4098	0.4290
<i>Baseline:random</i>	0.4969	0.2933	0.3201

Con ROUGE-C

Los resultados de las evaluaciones de los métodos y sistemas comerciales de generación automática de resúmenes con ROUGE-C y JS *divergence* son muy similares en el ranking con respecto a la posición de los métodos y sistemas. Para ROUGE-C y JS *divergence*, *baseline:first* no tiene mucha relevancia porque la referencia de evaluación es el documento completo, por lo que su evaluación es muy baja. De acuerdo con los resultados presentados en la Tabla 16 para ROUGE-C (R-C), todos los métodos y sistemas superan *baseline:random* y para *baseline:first* solamente un sistema no supera esta heurística.

Tabla 16. Resultados evaluados con ROUGE-C para los métodos, sistemas y heurísticas usando TER.

Método/Sistema	ROUGE - C1	ROUGE - C2	ROUGE - CL	ROUGE - CSU4
<i>GA-4feature</i>	0.5041	0.4968	0.5041	0.4864
<i>MA-SingleDocSum</i>	0.5044	0.4945	0.5044	0.4803
<i>TextRank</i>	0.4402	0.4290	0.4402	0.4128
<i>GA-Multilenguaje</i>	0.3915	0.3867	0.3915	0.3793
MOW2007	0.3688	0.3567	0.3654	0.3395
MOW 2003	0.3559	0.3438	0.3527	0.3266
<i>GA-Bag of words</i>	0.3477	0.3411	0.3477	0.3309
OTS	0.3509	0.3413	0.3490	0.3272
Text Compactor	0.3406	0.3315	0.3389	0.3177
Summarizing	0.2754	0.2636	0.2726	0.2466
<i>Baseline:first</i>	0.2791	0.2756	0.2764	0.2699
Copernic	0.2971	0.2852	0.2952	0.2733
<i>Baseline:random</i>	0.2538	0.2322	0.2475	0.2147

Con Jensen-Shannon divergence

Los resultados de los métodos del estado del arte y los sistemas comerciales para la generación automática de resúmenes son muy similares a los resultados obtenidos con ROUGE-C. Esto se debe a la similitud que tienen los métodos de evaluación, al considerar al documento original como referencia.

Tabla 17. Resultados evaluados con JS divergence para los métodos, sistemas y heurísticas usando TER.

Método/Sistema	JS-SMT	JS-WSMT
<i>GA-4feature</i>	0.8524	0.8436
<i>MA-SingleDocSum</i>	0.8452	0.8362
<i>TextRank</i>	0.8223	0.8120
<i>GA-Multilenguaje</i>	0.7950	0.7812
MOW2007	0.7920	0.7773
MOW 2003	0.7858	0.7702
<i>GA-Bag of words</i>	0.7796	0.7634
OTS	0.7745	0.7592
Text Compactor	0.7690	0.7526
Summarizing	0.7343	0.7124
<i>Baseline:first</i>	0.7321	0.7107
Copernic	0.7250	0.7061
<i>Baseline:random</i>	0.7105	0.6884

Matriz de ranqueo

El objetivo del método es realizar una actualización de la investigación de generación automática de resúmenes para el lenguaje español. Sin embargo, a partir de los resultados obtenidos gracias a los métodos de evaluación se puede generar un ranqueo de los métodos y sistemas de generación automática de resúmenes, lo que permite conocer, cuál de ellos es el mejor hasta el momento. Se utilizaron tres métodos de evaluación (ROUGE, ROUGE-C y JS divergence), y cada uno usa una forma diferente de calcular su resultado de salida (Ver tabla 15-17) por lo que no es posible determinar cuál es mejor. Por lo tanto, se propone la unificación de los métodos y sistemas considerando la posición que ocupa cada método y sistema según la medida de evaluación. La tabla 18, muestra la posición de los métodos y sistemas una vez unificados en base a la ecuación 24 propuesta por (Aliguliyev, 2009). Según los resultados de la tabla 18 el mejor método del estado del arte para la generación automática de resúmenes en español es GA-4feature propuesto por (Vazquez Vazquez et al., 2019).

Tabla 18. Ranking de los métodos y sistemas para la generación automática de resúmenes en español.

Método/Sistema	R(r)											R
	1	2	3	4	5	6	7	8	9	10	11	
<i>GA-4feature</i>	4	5	0	0	0	0	0	0	0	0	0	8.5
<i>MA-SingleDocSum</i>	2	4	0	3	0	0	0	0	0	0	0	7.8
<i>GA-Multilanguage</i>	3	0	0	6	0	0	0	0	0	0	0	7.3
<i>TextRank</i>	0	0	6	0	0	0	3	0	0	0	0	6.2
<i>GA-Bag of words</i>	0	0	3	0	0	1	4	1	0	0	0	5.1
MOW 2007	0	0	0	0	6	0	0	0	3	0	0	4.6
OTS	0	0	0	0	3	0	2	4	0	0	0	4.2
MOW 2003	0	0	0	0	0	5	0	1	0	3	0	3.6
Text Compactor	0	0	0	0	0	3	0	0	6	0	0	3.2
Copernic	0	0	0	0	0	0	0	3	0	4	2	2.0
Summarizing	0	0	0	0	0	0	0	0	0	2	7	1.0

Resultados del estudio cualitativo de la generación automática de resúmenes en inglés y español.

Para español

Para el lenguaje español se hicieron tres pruebas de Turing a 73 estudiantes y profesores de nivel superior y de posgrado. Los cuales tienen como lengua nativa el español. Los resultados son mostrados en la siguiente tabla.

Tabla 19. Resultados del *test de Turing* con 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 19, se muestra como en solo 8% de las ocasiones el humano acertó 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 hecho por el humano y uno por un sistema. Sin embargo, es interesante ver que el 36% de las personas pensaron que los resúmenes elaborados automáticamente eran los creados por humanos. Con estos resultados es posible ver que no sólo ha pasado el test de Turing, sino que ha superado al humano.

Para inglés

Para el lenguaje inglés se hicieron tres pruebas de Turing a 68 estudiantes y profesores de nivel superior y de posgrado. Los cuales tienen dominio en lectura del lenguaje inglés. Los resultados son mostrados en la siguiente tabla.

Tabla 20. Resultados del *test de Turing* con respecto al humano para el lenguaje español.

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

Para el lenguaje inglés solamente se pudieron identificar 13% de los casos de manera correcta, es decir, Humano-Humano. La mayoría de la confusión estuvo en selecciones Humano-Máquina con 46%. Sin embargo, en 41% de los casos se prefirió a los resúmenes hechos por el sistema, es decir, de cada 4 pruebas, 3 fueron claramente confundidas y se prefirió el resumen de la máquina. Nuevamente, el test de Turing fue superado, al menos en el dominio⁹ de las noticias.

⁹ El dominio hace referencia al ámbito en el que están escritos los documentos, por ejemplo, noticias, artículos científicos, poemas, tweets, correos electrónicos, entre otros.



CAPÍTULO 5.

Conclusiones

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

Conclusiones

A partir de la investigación se llegó a las siguientes conclusiones.

- Fue construido un método basado en algoritmos genéticos, para la generación automática de resúmenes de tipo extractivo, el cual es independiente del lenguaje, fue probado en inglés y español. El método considera las características del texto, posición de las oraciones y cobertura.
 - Los algoritmos genéticos facilitaron el proceso de generación automática de resúmenes, al hacer uso de operadores adaptados a la tarea.
 - Para facilitar el proceso de la generación automática de resúmenes se preprocesaron los textos y se hizo uso de los modelos de texto bolsa de palabras y n-gramas.
 - El método se probó con los corpus DUC01 (inglés), DUC02 (inglés) y TER (español).

- La tendencia al evaluar los modelos de texto, en especial n-gramas se observó que cuanto mayor es el valor de n , menor es el valor del *f-measure*.
- La función de aptitud considera dos características, posición de las oraciones y cobertura. En cuanto a la posición de las oraciones el modelo propuesto considera la pendiente de una recta para poder suavizar esta importancia. Para este trabajo la mejor pendiente para el lenguaje inglés está en el rango de -0.7 a -0.9. Para el lenguaje español está en -0.75 a -0.9.
- Una vez construido el método se utilizó para poder determinar la importancia de la posición de las oraciones en el lenguaje inglés, utilizando los diferentes modelos del estado del arte.
 - La característica posición de las oraciones es la más estudiada y aplicada al crear un método para la generación automática de resúmenes.
 - Se sabe que en el dominio de noticias la heurística *baseline:first* es muy elevada, lo que indica que las primeras oraciones son muy importantes en un texto. Sin embargo, en un buen resumen se deben considerar todas las oraciones, por lo que todas deben tener oportunidad de ser seleccionadas.
 - Se probaron siete formas de calcular la importancia de la posición de las oraciones. Para el corpus DUC01, la mejor forma fue la propuesta en el método. Para el corpus DUC02 la mejor forma fue la propuesta por (Vazquez Vazquez et al., 2019).
 - Se realizó la comparación de los principales métodos del estado del arte con el método propuesto aplicando el mejor modelo para calcular la posición de las oraciones.
- Se determino en que contexto esta la investigación de generación automática de resúmenes para el lenguaje español, en donde también se probó el método propuesto.
 - Se realizó el estudio de los 60 años de investigación de la generación automática de resúmenes en inglés, y se pudo conocer la forma y cuáles son los recursos necesarios para hacer una investigación formal.

- Tomando como base la investigación de la generación automática de resúmenes en inglés, se determinó la necesidad de un corpus estándar, el cálculo de las principales heurísticas y el uso de métricas de evaluación que permitieron la comparación con los métodos y sistemas.
- Se usó el corpus TER, el cual tiene características similares al corpus DUC02 (el más usado para el lenguaje inglés).
- Se probaron y evaluaron 4 métodos del estado del arte y 6 sistemas comerciales para la generación automática de resúmenes en español.
- Se utilizaron tres métodos de evaluación, ROUGE, ROUGE-C y JS divergence (los dos últimos no consideran resumen modelo para su evaluación)
- Se observó que la heurística *baseline:first* evaluada con ROUGE obtiene resultados muy elevados en comparación al lenguaje inglés.
- En los métodos de evaluación ROUGE-C y JS divergence la heurística *baseline:first* es muy baja, esto se atribuye a que los métodos utilizan como referencia de comparación al texto completo.
- Los métodos del estado del arte son mejores que los sistemas comerciales.
- Finalmente se comprobó que utilizando la prueba del test de Turing se puede medir la calidad de los resúmenes generados automáticamente, de tal forma que se miden de forma cualitativa.
 - Se aplicaron tres test de Turing a textos en inglés en español. Con el objetivo de que un conjunto de personas pudiera identificar de seis resúmenes presentados, los dos resúmenes hechos por el humano y los cuatro generados por la máquina (método o sistema).
 - Para el lenguaje español solamente el 8% de las personas pudieron identificar los dos resúmenes hechos por los humanos, mientras que el 92% restante resultó confundido por una máquina al seleccionar por lo menos uno generado por un método o sistema.
 - Para el lenguaje inglés solamente el 13% de las personas pudieron identificar los dos resúmenes hechos por los humanos, mientras que el 87% restante resultó confundido

por una máquina al seleccionar por lo menos uno generado por un método o sistema.

- Podemos decir que los resúmenes actualmente generados automáticamente por un método o sistema son de buena calidad, tanto que han logrado que un humano piense que el resumen es generado por otro humano.

Aportaciones

- Se propuso un método para la generación automática de resúmenes, extractivos, independientes del lenguaje.
- Se determinó la mejor forma de calcular la importancia de la posición de las oraciones para el lenguaje inglés.
- Se hizo una actualización de la investigación de la generación automática de resúmenes para el lenguaje español.
- Se presentó una evaluación cualitativa de los resúmenes generados automáticamente.

Artículos

Griselda Areli Matias Mendoza, Yulia Ledeneva, René Arnulfo García Hernández, Mikhail Alexandrov, Ángel Hernández Castañeda. Ground Truth Spanish Automatic Extractive Text Summarization Bounds. Vol. 24, No 3, pp. 1241–1256, 2020. Computación y Sistemas (incluida en el índice de Revistas Mexicanas de Investigación Científica y Tecnológica de CONACyT) doi: 10.13053/CyS-24-3-3484 <https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3484> <http://hdl.handle.net/20.500.11799/109468>

Determining the Importance of Sentence Position for Automatic Text Summarization

Griselda Areli Matias Mendoza*, Yulia Ledeneva* and René Arnulfo García-Hernández*
Universidad Autónoma del Estado de México, Unidad Académica Profesional Tlanguistenco, Instituto Literario, Toluca, Edo. Mex, México

Abstract. The methods of Automatic Extractive Summarization (AES) uses the features of the sentences of the original text to extract the most important information that will be considered in summary. It is known that the first sentences of the text are more relevant than the rest of the text (this heuristic is called baseline), so the position of the sentence (in reverse order) is used to determine its relevance, which means that the last sentences have practically no possibility of being selected. In this paper, we present a way to soften the importance of sentences according to the position. The comprehensive tests were done on one of the best AES methods using the bag of words and n-grams models with the with DUC02 and DUC01 data sets to determine the importance of sentences.

Keywords: Automatic Text Summarization, n-gram Model, Bag of Words Model, Slope Calculation, Genetic Algorithm

1. Introduction

Currently, information is exponentially growing and thus, the necessary time available for processing. Therefore, it is essential to have methods that allow Automatic Extractive Summarization (AES). The purpose of the methods AES is to generate summaries more similar to those generated by the human. Presently, summaries can be used in different areas. There are employed to summarize information, for example, for videos [1], newspapers [2] [3] [4], scientific papers [5] and social networks as Twitter [6] [7] or blog [8], where information rapidly changes and technologies are required to access real-time information represented in reduced form.

According to Ladda Saunmali [9], the purpose of the text summary is to present the most important information in a shorter version of the original text, maintaining its main content and helping the user to quickly understand the large volume of information. According to Alfonso, Berker, Da Cunha Fanego among others [9] [10] [11] [12] [13] [14] [15] [16] [17], the summaries are classified according to their

strategy of condensation in abstractive and extractive summaries. The abstractive summaries are those summaries generated from understanding the document and describe the content with words or sentences that sometimes are not in the original text. Instead, extractive summaries are generated from the selection of key phrases, sentences, or paragraphs considered essential for the original text; so, they do not require the understanding of the document.

Among the methods proposed for AES are those that need a large number of language resources [18] [19] [20] [21] [22] [23], so they have a high dependence on language or require sophisticated processes to generate a summary. There are also methods that only use the structure and distribution of the original text, so they are less dependent on language [2] [3] [24] [25] [26] [27] [28] [29]. The language-dependent methods may show better results than language-independent ones. However, research in language-independent methods has grown because of its possibility to apply to a wide range of languages. In this study, only language-independent extractive methods are considered.

* Corresponding authors: Griselda Areli Matias Mendoza, Yulia Ledeneva and René Arnulfo García-Hernández, Universidad Autónoma del Estado de México, Unidad Académica Profesional Tlanguistenco, Instituto Literario No. 100, CP 50000, Toluca, Edo. Mex, México. E-mails: gris_9123@hotmail.com (G.M. Matias), yledeneva@yahoo.com (Y. Ledeneva), renearnulfo@hotmail.com (R.A. García-Hernández)

- Griselda Matias Mendoza, Yulia Ledeneva, Rene Arnulfo García-Hernandez. Determining the Importance of Sentence Position for Automatic Text Summarization. Journal of Intelligent & Fuzzy Systems. ISSN print 1064-1246, ISSN online 1875-8967. Vol. 39, no.2, pp. 2421-2431. Ed. IOS Press (indexed by Journal Citation Report - IMPACT FACTOR 2020 1.851), 31 August 2020. DOI: 10.3233/JIFS-179902 <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179902>.

ISSN 2007-9737

Ground Truth Spanish Automatic Extractive Text Summarization Bounds

Griselda Areli Matias Mendoza¹, Yulia Ledeneva^{1,2}, René Arnulfo García Hernández², Mikhail Alexandrov^{2,3}, Ángel Hernández Castañeda¹

¹ Universidad Autónoma del Estado de México, Unidad Académica Profesional Tianguistenco, Mexico

² Presidential Academy of National Economy and Public Administration, Department of system analysis and informatics, Russia

³ Autonomous University of Barcelona, Spain

{gris_9123, reneamulfo}@hotmail.com, yledeneva@yahoo.com, malexandrov@mail.ru, angelhc2305@gmail.com

Abstract. The textual information has accelerated growth in the most spoken languages by native Internet users, such as Chinese, Spanish, English, Arabic, Hindi, Portuguese, Bengali, Russian, among others. It is necessary to innovate the methods of Automatic Text Summarization (ATS) that can extract essential information without reading the entire text. The most competent methods are Extractive ATS (EATS) that extract essential parts of the document (sentences, phrases, or paragraphs) to compose a summary. During the last 60 years of research of EATS, the creation of standard corpus with human-generated summaries and evaluation methods which are highly correlated with human judgments help to increase the number of new state-of-the-art methods. However, these methods are mainly supported for the English language, leaving aside other equally important languages such as Spanish, which is the second most spoken language by natives and the third most used on the Internet. A standard corpus for Spanish EATS (SAETS) is created to evaluate the state-of-the-art methods and systems for the Spanish language. The main contribution consists of a proposal for configuration and evaluation of 5 state-of-the-art methods, five systems and four heuristics using three evaluation methods (ROUGE, ROUGE-C, and Jensen-Shannon divergence). It is the first time that Jensen-Shannon divergence is used to evaluate AETS. In this paper the ground truth bounds for the Spanish language are presented, which are the heuristics *baseline.first*, *baseline.random*, *topline* and

concordance. In addition, the ranking of 30 evaluation tests of the state-of-the-art methods and systems is calculated that forms a benchmark for SAETS.

Keywords. Spanish automatic text summarization, ROUGE, ROUGE-C, Jensen Shannon divergence, corpus TER.

1 Introduction

The information has become a necessary resource whose growth is increasing in different languages spoken in the world. Among the most spoken languages according to their number of native speakers are Chinese, Spanish, English, Arabic, Hindi, Portuguese, Bengali, and Russian, among others [1]. To have access to the information that is generated day by day, it is suggested to use the methods of Automatic Text Summarization (ATS). ATS aims to extract the most relevant information of a document [2].

The most of the state-of-the-art methods have been based on Automatic Extractive Text Summarization (AETS) because of its easy implementation and competent results. The methods of AETS extract essential parts of a text

Referencias

- Acero, I., Alcojor, M., Díaz Esteban, A., Gómez Hidalgo, J.M., Maña López, M.J., 2001. Generación automática de resúmenes personalizados. *Proces. Leng. Nat.* Nº 27 Septiembre 2001 Pp 281-290.
- Ahonen-Myka, H., 1999. Finding all maximal frequent sequences in text. Presented at the Proc. of the ICML99 workshop on machine learning in text data analysis, pp. 11–17.
- Alcón, O., Lloret, E., 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües.
- Alfonseca, E., Rodríguez, P., 2003. Generating extracts with genetic algorithms. Presented at the European Conference on Information Retrieval, Springer, pp. 511–519.
- Alguliev, R.M., Aliguliyev, R.M., Hajirahimova, M.S., 2012. GenDocSum+ MCLR: Generic document summarization based on maximum coverage and less redundancy. *Expert Syst. Appl.* 39, 12460–12473.
- Alguliyev, R., 2009. Evolutionary Algorithm for Extractive Text Summarization. <https://doi.org/10.4236/iim.2009.12019>
- Alguliyev, R.M., Aliguliyev, R.M., Isazade, N.R., Abdi, A., Idris, N., 2019. COSUM: Text summarization based on clustering and optimization. *Expert Syst.* 36, e12340.
- Aliguliyev, R.M., 2009. Performance evaluation of density-based clustering methods. *Inf. Sci.* 179, 3583–3602.
- Araujo, L., Cervigón, C., 2009. Algoritmos evolutivos: un enfoque práctico. Alfaomega.
- Babar, S., Patil, P.D., 2015. Improving Performance of Text Summarization. *Procedia Comput. Sci.* 46, 354–363.
- Barzilay, R., Elhadad, M., 1999. Using lexical chains for text summarization. *Adv. Autom. Text Summ.* 111–121.
- Belkebir, R., Guessoum, A., 2015. A supervised approach to arabic text summarization using adaboost, in: *New Contributions in Information Systems and Technologies*. Springer, pp. 227–236.
- Benbrahim, M., Ahmad, K., 1995. Text summarisation: The role of lexical cohesion analysis. *New Rev. Doc. Text Manag.* 321–335.

- Benítez, R., Escudero, G., Kanaan, S., Rodó, D.M., 2014. Inteligencia artificial avanzada. Editorial UOC.
- Berker, M., 2011a. Using genetic algorithms with lexical chains for automatic text summarization.
- Berker, M., 2011b. Using genetic algorithms with lexical chains for automatic text summarization.
- Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., McKeown, K., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Siddharthan, A., Siegelman, S., 2004. Columbia university at duc 2004. Presented at the Proceedings of the Document Understanding Conference, DUC-2004, Boston, USA.
- Bossard, A., Génèreux, M., Poibeau, T., 2008. Description of the LIPN System at TAC 2008: Summarizing Information and Opinions. Presented at the TAC 2008, pp. 282–291.
- 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, in: 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, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 335–336.
- Cardoso, A.C., Abelleira, M.A.P., 2013. Generación automática de resúmenes, in: De 1er Congreso Nacional de Ingeniería Informática/Sistemas de Información, CoNallSI.
- Careaga-Moya, J.-A., Medina-Urrea, A., Torres-Moreno, J.-M., 2012. A New Cross-Lingua Automatic Summerization Approach Based on Textual Energy. Journ. Int. D'Analyse Stat. Données Textuelles 247–255.
- Coarite Choque, R., 2008. Areas de aplicación de la Inteligencia Artificial. Rev. Inf. Technol. Soc. 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 Edupdiulairia Dacunha 0 202, 07–04.

- Da Cunha, I., Torres-Moreno, J.-M., Velázquez-Morales, P., Vivaldi, J., 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática* 1, 67–79.
- Darwin, C., Bynum, W.F., 2009. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt New York.
- Edmundson, H.P., 1969. New methods in automatic extracting. *J. ACM JACM* 16, 264–285.
- Edmundson, H.P., Wyllys, R.E., 1961. Automatic abstracting and indexing—survey and recommendations. *Commun. ACM* 4, 226–234.
- Fernández, D.V., 2018. *El español una lengua viva. Informe 2018*. Instituto Cervantes. Página Web. URL https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2018.pdf (accessed 5.15.18).
- Fernández, S., SanJuan, E., Torres-Moreno, J.M., 2007. Textual energy of associative memories: Performant applications of enertex algorithm in text summarization and topic segmentation, in: *Mexican International Conference on Artificial Intelligence*. Springer, pp. 861–871.
- 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.
- García-Hernández, R.A., Ledeneva, Y., Mendoza, G.M., Dominguez, Á.H., Chavez, J., Gelbukh, A., Fabela, J.L.T., 2009. Comparing commercial tools and state-of-the-art methods for generating text summaries. Presented at the *Artificial Intelligence, 2009. MICAI 2009. Eighth Mexican International Conference on*, IEEE, pp. 92–96.
- García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., 2010. Finding maximal sequential patterns in text document collections and single documents. *Inform. Int. J. Comput. Inform.* 34, 93–101.
- Goularte, F.B., Nassar, S.M., Fileto, R., Saggion, H., 2019. A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Syst. Appl.* 115, 264–275.
- He, T., Chen, J., Ma, L., Gui, Z., Li, F., Shao, W., Wang, Q., 2008. ROUGE-C: A fully automated evaluation method for multi-document summarization, in: *Granular Computing, 2008. GrC 2008. IEEE International Conference On*. IEEE, pp. 269–274.
- Hernández, M.B., Gómez, J.M., 2013. Aplicaciones de Procesamiento de Lenguaje Natural. *Rev. Politécnica* 32.

- 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.
- Hovy, E., C.L., 1999. Automated Text Summarization in SUMMARIST. En Mani M Maybury Eds Adv. Autom. Text Summ. 81–94.
- Jing, H., Barzilay, R., McKeown, K., Elhadad, M., 1998. Summarization evaluation methods: Experiments and analysis, in: AAI 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.
- Kiyomarsi, F., 2015. Evaluation Of Automatic Text Summarizations Based On Human Summaries. *Procedia-Soc. Behav. Sci.* 192, 83–91.
- Kupiec, J., Pedersen, J., Chen, F., 1995. A trainable document summarizer, in: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 68–73.
- Kuri, Á., Galaviz, J., 2002. Algoritmos genéticos. IPN.
- Last, M., Litvak, M., 2010. Language-independent Techniques for Automated Text Summarization. 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.
- 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, in: 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, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, p. 605.
- Litvak, M., Lipman, H., Ben Gur, A., Last, M., Kisilevich, S., Keim, D., 2010. Towards multi-lingual summarization: A comparative analysis of sentence extraction methods on English and Hebrew corpora.
- Lloret, E., 2011. Text summarisation based on human language technologies and its applications. Universidad de Alicante.
- Luhn, H.P., 1958a. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165.
- Luhn, H.P., 1958b. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165.
- 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, in: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 77–85.
- Mani, I., Maybury, M.T., 1999. Advances in automatic text summarization. MIT Press.
- Marcu, D., 1997. From discourse structures to text summaries. Intell. Scalable Text Summ.
- 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, 29–36.
- Matias Mendoza, G., 2013. Generación automática de resúmenes usando algoritmos genéticos. Universidad Autónoma del Estado de México.
- Matias Mendoza, G.A., 2016. Generación Automática de Resúmenes Independientes del Lenguaje. Universidad Autónoma del Estado de México.
- Matias Mendoza, G.A., Ledeneva, Y., García Hernández, R.A., Alexandrov, M., Hernández Castañeda, Á., 2020. Ground Truth Spanish Automatic Extractive Text Summarization Bounds. Comput. Sist. 24.

- McKeown, K., Radev, D.R., 1995. Generating summaries of multiple news articles, in: 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, 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. 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.
- Mitrat, M., Singhal, A., Buckleytt, C., 1997. Automatic text summarization by paragraph extraction. *Intell. Scalable Text Summ.*
- Molina, A., 2013. Compresión automática de frases: un estudio hacia la generación de resúmenes en español. *Intel. Artif.* 16, 41–62.
- Montiel, R., 2009. Generación automática de resúmenes mediante aprendizaje no supervisado. Instituto Tecnológico de Toluca, Edo. México.
- Mosa, M.A., Anwar, A.S., Hamouda, A., 2019. A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowl.-Based Syst.* 163, 518–532.
- Nandhini, K., Balasundaram, S.R., 2014. Extracting easy to understand summary using differential evolution algorithm. *Swarm Evol. Comput.* 16, 19–27.
- Nenkova, A., McKeown, K., 2012. A survey of text summarization techniques, in: *Mining Text Data*. Springer, pp. 43–76.
- Neri Mendoza, V., 2019. Modelado de Posición de Oraciones y Cobertura Mediante un Algoritmo Genético para la Generación Automática de Resúmenes de Múltiples Documentos. Universidad Autónoma del Estado de México.

- Orăsan, C., 2003. An evolutionary approach for improving the quality of automatic summaries, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering-Volume 12. Association for Computational Linguistics, pp. 37–45.
- 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, 1506–1520.
- Paice, C.D., 1990. Constructing literature abstracts by computer: techniques and prospects. *Inf. Process. Manag.* 26, 171–186.
- 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. Universidad Complutense de Madrid, Madrid.
- Polya, G., Zugazagoitia, J., 1965. *Cómo plantear y resolver problemas*. Trillas.
- Porter, M.F., 1980. An algorithm for suffix stripping. *Program* 14, 130–137.
- Pratt, K.S., 2009. *Design Patterns for Research Methods: Iterative Field Research*. Presented at the AAAI Spring Symposium: Experimental Design for Real-World Systems.
- Qazvinian, V., Hassanabadi, L.S., Halavati, R., 2008. Summarising text with a genetic algorithm-based sentence extraction. *Int. J. Knowl. Manag. Stud.* 2, 426–444.
- Radev, D.R., Jing, H., Styś, M., Tam, D., 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.* 40, 919–938.
- Rojas J., 2017. *Cálculo de Topline para la generación automática de resúmenes usando algoritmos genéticos*. Universidad Autónoma del Estado de México, Edo. México.
- Saggion, H., 2011. Using SUMMA for Language Independent Summarization at TAC 2011. Presented at the TAC.
- Salton, G., Wong, A., Yang, C.-S., 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 613–620.
- Sidorov, G., 2019. *Syntactic n-grams in computational linguistics*. Springer.

- Song, W., Choi, L.C., Park, S.C., Ding, X.F., 2011. Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. *Expert Syst. Appl.* 38, 9112–9121.
- Sparck Jones, K., Galliers, J.R., 1995. Evaluating natural language processing systems: An analysis and review. Springer Science & Business Media.
- Steinberger, J., Ježek, K., 2012. Evaluation measures for text summarization. *Comput. Inform.* 28, 251–275.
- Suanmali, L., Salim, N., Binwahlan, M.S., 2011a. Genetic algorithm based sentence extraction for text summarization. *Int. J. Innov. Comput.* 1, 2180–4370.
- Suanmali, L., Salim, N., Binwahlan, M.S., 2011b. Genetic Algorithm Based Sentence Extraction for Text Summarization. *Int. J. Innov. Comput.* 1.
- Svore, K., Vanderwende, L., Burges, C., 2007. Enhancing single-document summarization by combining RankNet and third-party sources, in: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
- Téllez, A., Montes, M., Villaseñor-Pineda, L., 2009. Using Machine Learning for Extracting Information from Natural Disaster News Reports. *Comput. Sist. Instituto Politécnico Nacional*, 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, in: *El Español, Lenguaje de Traducción Para La Cooperación y El Dialogo*. Madrid.
- Torres-Moreno, J.-M., 2014. Automatic text summarization. John Wiley & Sons.
- Torres-Moreno, J.-M., Saggion, H., Cunha, I. da, SanJuan, E., Velázquez-Morales, P., 2010. Summary Evaluation with and without References. *Polibits* 13–20.
- Turing, A.M., 1950. Computing Machinery and Intelligence. *Computing Machinery and Intelligence* 433–460.
- Uddin, M.N., Khan, S.A., 2007. A study on text summarization techniques and implement few of them for Bangla language, in: *Computer and Information Technology, 2007. Iccit 2007. 10th International Conference On. IEEE*, pp. 1–4.
- Vazquez Vazquez, E., García-Hernández René, R.A., Ledeneva, Y., 2018. Sentence Features Relevance for Extractive Text Summarization using Genetic Algorithms. *J. Intell. Fuzzy Syst. Appl. Eng. Technol.* 35, 353–365.

- Vazquez Vazquez, E., Ledeneva, Y., García Hernández, R.A., 2019. Learning relevant models using symbolic regression for automatic text summarization. *Comput. Sist.* 23, 127.
- Villatoro E., 2007. Generación automática de resúmenes de múltiples documentos. Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla.
- Villatoro-Tello, E., Villaseñor-Pineda, L., Montes-y-Gómez, M., 2006. Using word sequences for text summarization. Presented at the International Conference on Text, Speech and Dialogue, Springer, pp. 293–300.
- 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.