UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

MODELADO DE CARACTERÍSTICAS PARA LA GENERACIÓN AUTOMÁTICA DE RESÚMENES EXTRACTIVOS DE MÚLTIPLES DOCUMENTOS

T E S I S

PARA OBTENER EL GRADO DE
DOCTORA EN CIENCIAS DE LA COMPUTACIÓN

QUE PRESENTA:
VERÓNICA NERI MENDOZA

TUTORA ACADÉMICA:
DRA. YULIA NIKOLAEVNA LEDENEVA

TUTORES ADJUNTOS:
DR. RENÉ ARNULFO GARCÍA HERNÁNDEZ
DR. ÁNGEL HERNÁNDEZ CASTAÑEDA

TIANGUISTENCO, ESTADO DE MÉXICO ENERO, 2024

Resumen

Actualmente, un gran número de documentos electrónicos están disponibles en línea, por lo que los lectores enfrentan dificultades para encontrar información relevante. Los lectores se cansan leyendo una gran cantidad de textos que pueden omitir la lectura de documentos importantes. Por lo tanto, se necesitan métodos para la Generación Automática de Resúmenes (GAR) que permitan resumir un conjunto de documentos.

De acuerdo con el enfoque de condensación, un resumen puede ser generado de forma extractiva (seleccionando oraciones del documento fuente), abstractiva (reinterpretando el documento), e híbrida (combina las dos anteriores). Por otra parte, según el número de documentos a resumir, se puede generar resúmenes de Documentos Individuales (GARDI) y de Múltiples Documentos (GARMD). La GARMD es más difícil que la GARDI porque requiere identificar una forma de ordenar los documentos de entrada. Asimismo, los documentos a resumir pudieron ser escritos en diferentes momentos. En el Estado del Arte (EA) se han establecido tres principales requerimientos en la GARMD: (1) Las oraciones deben ser relevantes, (2) deben cubrir todos los temas tratados en los documentos (cobertura) y (3) se debe reducir la redundancia de información.

Los métodos de la GARMD con enfoque extractivo se centran en modelar características textuales, que pueden ser estadísticas, lingüísticas e híbridas. De esta manera, cada característica pondera cada oración de los documentos fuente y después se seleccionan oraciones para construir el resumen. En consecuencia, se pueden distinguir dos aspectos: selección de oraciones y modelado de características. Cada uno de estos aspectos presenta diferentes dificultades.

En la selección de oraciones se debe obtener un subconjunto de oraciones del texto fuente para generar el resumen. Esto ha sido abordado en el EA como problema de clasificación, determinando si una oración será incluida en el resumen

o no. En cambio, otras investigaciones han realizado aproximaciones mediante métodos de optimización combinatoria, tal como el Algoritmo Genético (AG).

Por otro lado, el problema al que se enfrenta en el modelado de características es determinar el nivel de importancia de cada una de ellas (normalmente, determinado por un coeficiente de relevancia). Para establecer esta importancia, es necesario analizar la contribución que tiene cada una en los requerimientos de la GAR, en los niveles textuales (palabra, oración, párrafo y grafo) y de estructura lingüística (léxico, sintáctico y semántico). En el EA se ha propuesto y estudiado ampliamente un conjunto de características híbridas. Sin embargo, cada investigación presenta diferentes maneras de modelar características y de ponderar sus niveles de importancia. Comúnmente, esta importancia se obtiene mediante ajustes manuales o métodos de optimización, siendo alternativas subjetivas y costosas. Por tanto, se plantea el siguiente problema: se desconoce un modelo de características híbridas que pueda mejorar la selección de oraciones para la GARMD extractivos.

A lo largo de la investigación de la GAR, se han creado diversos conjuntos de datos, que incluyen resúmenes de referencia escritos por humanos, con la finalidad de comparar la capacidad de los métodos propuestos, con las habilidades humanas.

Hipótesis. Partiendo de lo anterior mencionado, en esta investigación se planteó la siguiente hipótesis: Si los resúmenes de referencia son el modelo objetivo de la tarea de la GARMD extractivos, entonces la obtención de sus características híbridas en un modelo permitirá mejorar la selección de oraciones en dicha tarea. Esta hipótesis parte del hecho que los resúmenes de referencia son generados con pocos o nulos errores, por lo que el modelado de características de estos documentos podrá mejorar la GARMD extractivos.

Experimentación. Para evaluar el desempeño del modelado de características propuesto, fueron considerados dos conjuntos de datos para la GAR: genéricos (DUC01, con 4 diferentes longitudes de resumen) y de actualización (TAC08), bajo

el dominio de noticias, con el objetivo de tener una perspectiva más amplia de la calidad de las oraciones seleccionadas en el resumen final.

Resultados y conclusiones. A partir del modelado de características propuesto, se seleccionaron oraciones a través del AG. Los resultados obtenidos en ambas colecciones de documentos superan a otros métodos del EA, así como heurísticas de referencia. Por lo tanto, se comprueba la hipótesis previa, logrando mejoras en la selección de oraciones.

Contenido

Figuras

Tablas

CAPÍTULO 1
Introducción

1.1 Antecedentes

La Inteligencia Artificial (IA), se ha definido como la capacidad de las computadoras para realizar actividades propias del ser humano (Russell & Norvig, 2014; Zini & Awad, 2022). Actualmente, las aplicaciones desarrolladas a través de la IA dan una aproximación de soluciones a problemas muy complejos para el ser humano (Dubey et al., 2022). Esto se ha logrado, debido a la constante investigación en el Procesamiento de Lenguaje Natural (PLN).

El PLN es el campo de la IA y la lingüística computacional que tiene como objetivo la construcción de técnicas que permitan la comunicación entre el ser humano y las computadoras, a través de sistemas inteligentes capaces de comprender, identificar, y extraer significado de texto y habla (Hirschberg & Manning, 2019; Khurana et al., 2022; Meera & Geerthik, 2022). Estos sistemas han otorgado a las computadoras la capacidad de emular actividades tales como análisis de sentimientos, comprensión del lenguaje,

corrección de texto, traducción automática, reconocimiento de voz, extracción de información y Generación Automática de Resúmenes (GAR) (Dubey et al., 2022; Zini & Awad, 2022).

Para comprender la idea de la GAR, es necesario entender el concepto fundamental del resumen. Un resumen es un documento breve que contiene los aspectos principales de uno o más textos (Brown et al., 1983; Matias et al., 2020; Spirgel et al., 2016). El proceso que realiza el ser humano para generar un resumen consta de en las siguientes etapas:

1. Leer el texto que se va a resumir.
2. Identificar la idea general del texto y luego las ideas más importantes de cada uno de los párrafos, omitiendo la información poco relevante.
3. Organizar las ideas del texto leído.
4. Escribir las ideas identificadas y organizadas, siguiendo una estructura de párrafos.

En general, este proceso resulta tedioso, agotador, además si se requiere resumir varios textos, esta actividad requiere de bastante tiempo. Por lo que es necesario automatizar este proceso debido a la situación anterior, a la explosión de información en la Internet y las redes sociales. De esta situación emerge la GAR, que es una tarea cuyo objetivo es crear resúmenes de uno o varios documentos por medio de un software. De esta manera, los resúmenes generados ayudan transmitir la idea principal al lector reconociendo y comprendiendo la información más significante.

Los resúmenes generados a través de una computadora ahorran tiempo y ayudan a seleccionar oraciones de manera eficiente. En comparación con los resúmenes generados por humanos, los sistemas que generan resúmenes automáticos son menos sesgados (Abualigah et al., 2020; Mojrian & Mirroshandel, 2021; Sanchez-Gomez et al., 2022). Debido a que las computadoras carecen de conocimiento humano y capacidad de comprender el lenguaje, la GAR es una tarea difícil y no trivial (Allahyari et al., 2017; El-Kassas et al., 2021).

Dependiendo de la cantidad de documentos de entrada, la GAR puede dividirse en las siguientes tareas: GAR de Documentos Individuales (GARDI) y GAR de Múltiples Documentos (GARMD). La historia de la GAR comenzó hace aproximadamente 60 años

(Matias et al., 2020), centrándose en la GARDI. Sin embargo, el renacimiento de este campo sucedió en los 90s, donde también comenzó la investigación de la GARMD, con el objetivo de proporcionar al usuario un resumen breve e informativo de un conjunto de documentos relacionados a un tema en particular (Abualigah et al., 2020; Hark et al., 2022; Roul & Sahoo, 2022).

La diferencia entre la GARDI y la GARMD no solo depende del número de documentos de entrada. La GARMD es más complicada, ya que es necesario identificar una forma de ordenar los documentos en cuestión, la información se encuentra diversificada temáticamente y contradictoria (es decir, se presentan hechos contrarios u opuestos). Además, los documentos a resumir pudieron ser escritos en diferentes momentos. Por lo tanto, el resumen debe considerar diferentes fuentes de información, así como la tasa de compresión (longitud del resumen). En relación a la longitud del resumen, la mayoría de las herramientas desarrolladas en el Estado del Arte (EA) para la GARDI han sido probadas bajo una evaluación estándar a 100 palabras, mientras que en la GARMD se consideran resúmenes de 50, 100, 200 y 400 palabras (Asawa & Balaji, 2020; Sanchez-Gomez et al., 2020). Por otro lado, en la GARMD la complejidad algorítmica es mayor.

De acuerdo con El-Kassas et al. (2021) y Hou et al. (2021), en la GAR se pueden distinguir tres enfoques que se describen a continuación:

- **Extractivo:** Los métodos basados en este enfoque asignan pesos a las oraciones de acuerdo con características lingüísticas o estadísticas. Posteriormente, se seleccionan las oraciones con mayor peso.
- **Abstractivo:** Este enfoque permite que los métodos generen resúmenes incorporando nuevas palabras y oraciones provenientes de un corpus o diccionarios. Además, requiere técnicas más sofisticadas como paráfrasis y fusión de oraciones.
- **Híbrido:** En el EA se han propuesto modelos híbridos que combinan las ventajas de los modelos extractivos y abstractivos. En este enfoque, el texto se procesa en dos pasos: primero se genera un resumen selectivo o extractivo. Posteriormente, el resumen generado es utilizado por un método abstractivo como entrada para construir el resumen final.

Particularmente, en el enfoque extractivo se distinguen dos principales aspectos:

1. Selección de oraciones
2. Modelado de características

En la selección de oraciones se debe obtener un subconjunto de oraciones de los documentos fuente que represente la idea general de los documentos de entrada. Por ejemplo, en la colección DUC01 el promedio de oraciones por documento para la GARMD es de 356. Por lo tanto, el número de posibles resúmenes extractivos que se pueden generar a partir de 356 oraciones es de $14{,}678 \times 10^{107}$. En consecuencia, esta situación se ha abordado en el EA como un problema de optimización combinatoria a través del Algoritmo Genético.

Por otra parte, el modelado de características juega un papel trascendental para la selección de oraciones del texto fuente deberán ser incluidas en el resumen. Entre las características que han sido propuestas en el estado del arte se destacan las siguientes:

- Posición de las oraciones dentro del texto
- Cobertura
- Frecuencia de palabras
- Similitud con el título
- Longitud de las oraciones
- Reducción de redundancia
- Coherencia
- Inclusión de nombre propios
- Inclusión de palabras temáticas
- Aparición de verbos
- Aparición de adverbios
- Entidades Nombradas (NER)
- Inclusión de números
- Frecuencia inversa de las oraciones
- Palabras clave positivas
- Palabras clave negativas
- Similitud entre las oraciones
- Posición de oración relativa al párrafo

Las características anteriormente mencionadas se clasifican en estadísticas y lingüísticas. Dentro del primer grupo, no se emplea conocimiento del lenguaje y se analiza la frecuencia y distribución de los términos empleados sin necesidad de entender el documento. Para el segundo grupo, se refiere al uso de conocimiento lingüístico o de dominio para analizar oraciones. Además, de ambos grupos se suelen emplear características obteniendo modelos híbridos de características.

Uno de los problemas comunes en relación con el modelado de características ha sido determinar: ¿cuáles de ellas se deben considerar en el proceso de la generación del resumen? Sin embargo, incluir todas puede considerarse una solución inadecuada o poco útil en la selección de oraciones (Yao et al., 2017). Por lo tanto, se han desarrollado diversas investigaciones sobre el modelado de características debido a su importancia. Algunos trabajos han tratado este problema analizando el impacto de las características individuales y algunas combinaciones de ellas, tal y como se propuso en (Sanchez-Gomez et al., 2021; S. Verma & Vagisha Nidhi, 2019).

Otro de los grandes problemas en el modelado de características ha sido la asignación del nivel de importancia en las características. En general, no todas se tratan con la misma importancia, ya que algunas aportan mejor información que otras. Para realizar una asignación de importancia, se requiere que a cada característica se le proporcione un coeficiente de relevancia cuya precisión puede ser un parámetro variable. En algunas investigaciones del EA se ha utilizado una precisión de cinco decimales para obtener mejores resultados. Por lo tanto, si se considera un grupo de 60 características, se tendrían 7.27×10^{134} posibles formas de combinar la importancia de ellas, lo que implica que sea una tarea costosa. En otros casos se han realizado ajustes manuales de coeficientes de relevancia, lo que tiende a sugerir valores subjetivos.

A lo largo de más de 60 años de investigación de la GAR, se han creado diversos conjuntos de datos, que incluyen resúmenes de referencia escritos por humanos, con la finalidad de comparar la capacidad de los métodos propuestos, con las habilidades humanas. Estos conjuntos de datos abarcan numerosos dominios. Uno de ellos es el de noticias, donde es necesario producir descripciones condensadas de temas de interés público a partir de múltiples fuentes de información. Entre este conjunto de datos están DUC01 y TAC008. Los tipos de resúmenes que se pueden generar a partir de estos conjuntos de datos son:

- **Genéricos:** Los resúmenes genéricos son documentos que contienen la mayor cantidad de información relevante sobre un hecho. La longitud de estos resúmenes suele ser de 50, 100, 200 y 400 palabras.

- **De actualización:** El resumen de actualización tiene el propósito de informar mientras sucede la noticia. Para esto, se generan dos resúmenes de 100 palabras: el inicial (Resumen A) que contiene una introducción sobre la noticia, y el de actualización (Resumen B).

1.2 Planteamiento del Problema

Se desconoce qué tan útil puede ser el modelado de características híbridas obtenidas de los resúmenes de referencia que pueda mejorar la selección de oraciones para la GARMD extractivos.

1.3 Hipótesis

Si los resúmenes de referencia son el modelo objetivo de la tarea de la GARMD extractivos, entonces la obtención de sus características híbridas permitirá mejorar la selección de oraciones en dicha tarea.

1.4 Objetivo General

Obtener un modelo de características híbridas de los resúmenes de referencia escritos por humanos que mejore la selección de oraciones en la GARMD extractivos.

1.4.1 Objetivos Específicos

1. Analizar los conjuntos de datos.
2. Seleccionar un enfoque de concatenación de documentos de entrada.
3. Analizar las características que se han propuesto en el estado del arte para la GARMD.
4. Modelar las características de los documentos de referencia escritos por humanos.
5. Optimizar la selección de oraciones a través del AG, con base en el modelo de características obtenido.
6. Evaluar y analizar el desempeño de los resúmenes generados a través de ROUGE.

7. Comparar los resultados obtenidos del modelo propuesto con los trabajos del estado del arte

1.5 Estructura de la tesis

El resto de la tesis está organizado de la siguiente manera: Dentro del Capítulo 2 se presenta el Marco Teórico que presenta los fundamentos en los que se basa esta investigación. Posteriormente, el Capítulo 3 presenta el trabajo relacionado que ha sido propuesto en el estado del arte y que da pauta para el método propuesto. Por otro lado, en el Capítulo 4 se presenta el método propuesto. Mientras que en el Capítulo 5 se muestran las configuraciones y las pruebas realizadas del método propuesto. Finalmente, en el Capítulo 6 se presentan las conclusiones, así como las aportaciones y el trabajo futuro derivado de esta investigación.

CAPÍTULO 2
Marco teórico

En este capítulo se presenta el Marco Teórico que presenta los fundamentos en los que se basa esta investigación. Se describe el área de investigación y los conceptos que ayudan a contextualizar el tema de estudio.

2.1 Procesamiento de Lenguaje Natural (PLN)

Es un área de investigación de la lingüística computacional y aprendizaje automático, el cual explora técnicas y mecanismos para que las computadoras puedan comprender y manipular texto o habla en lenguaje natural, con el propósito de realizar tareas útiles (Chowdhary, 2020; Dubey et al., 2022). En cuanto a texto se refiere, las aplicaciones que se han abordado son: Clasificación de texto, Indexación textos largos, Traducción automática, Extracción de información de currículos, Generación de texto y diálogos, y Generación automática de resúmenes (Chowdhary, 2020).

El PLN es un tema ampliamente discutido e investigado en la actualidad, debido a que ha tenido un gran avance en diversas áreas de conocimiento por medio de diversos algoritmos. No obstante, aún no se ha alcanzado la perfección, pero continúa mejorando constantemente (Dubey et al., 2022).

2.1.1 Recuperación y extracción de información

Dentro de cada tarea del PLN, la recuperación de información (RI) desempeña una función importante, ya que ayuda a encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto), o semiestructurada (páginas Web), a partir de repositorios grandes de datos, por medio de equipos locales o en Internet (Croft & Bruce, 2019; Hernández & Gómez, 2013; Manning et al., 2010). Generalmente, la RI supera a la búsqueda de bases de datos tradicionales, convirtiéndose en la forma dominante de acceso a la información que satisface una necesidad de información (Ledeneva & García-Hernández, 2017). A su vez, las áreas de la RI dan soporte a un campo relativamente creciente y con un gran valor comercial denominado Minería de texto (Cardoso & Pérez-Abelleira, 2013; Goularte et al., 2019; Kiefer et al., 2019; Montes-y-Goméz, 2014; Noshi & Schubert, 2019).

2.2 Generación Automática de Resúmenes (GAR)

Con el rápido crecimiento de la información en la actualidad, las personas pueden obtener y compartir información casi instantáneamente, de una gran cantidad de datos fuente. Como resultado de esto, causa una sobrecarga de información, por lo que se necesitan de herramientas que brinden acceso oportuno a la vasta información que hay en la Internet (S. Verma & Vagisha Nidhi, 2019; Widyassari et al., 2020). El uso de estas herramientas que brindan acceso oportuno a la información y generan resúmenes para aliviar la sobrecarga de información que enfrentan las personas. Por lo tanto, estos problemas han impulsado el interés de desarrollar sistemas que generen resúmenes automáticos (S. Verma & Vagisha Nidhi, 2019; Widyassari et al., 2020; Yao et al., 2017).

El proceso por el cual el ser humano realiza un resumen consta de las siguientes etapas (Alarico, 1996; Alonso Arévalo, 1976; Cardoso & Pérez-Abelleira, 2013; Cortés, 2011; Nazari & Mahdavi, 2019):

- **Lectura:** El documento en cuestión requiere ser leído completamente.
- **Análisis**: En este paso se identifican y se combinan las ideas centrales del texto, (este paso suele requerir un conocimiento previo).
- **Escritura:** Finalmente, se escribe toda la información relevante, de manera que se obtenga un texto de menor tamaño que el original.

Las etapas anteriormente mencionadas son consideradas como estándar para generar resúmenes manuales. Sin embargo, no son inmediatamente realizadas, por lo que requiere esfuerzo y tiempo por parte de las personas que generan el resumen. En cambio, la obtención de dichos resúmenes de forma automática implica el uso de un software, el cual toma un texto de entrada, lo condensa y presenta una versión reducida de grandes cantidades de información (Cardoso & Pérez-Abelleira, 2013; Gelbukh, 2010; Lloret; Elena, 2011; Matías, 2013; Nazari & Mahdavi, 2019).

2.2.1 Importancia de los resúmenes

La GAR aporta de manera eficiente y oportuna información relevante sobre un tema en particular. Entre las principales ventajas de la GAR se encuentran:

- Los resúmenes generados reducen el tiempo de lectura.
- El proceso de selección de información útil y relevante se convierte en una tarea más eficiente.
- Mejora la efectividad de la indexación de documentos.
- Los algoritmos utilizados para generar resúmenes automáticos son más imparciales y objetivos que los resúmenes realizados por personas.
- Facilitan la retención del material estudiado, ya que se asimila una síntesis de los aspectos esenciales de cada tema.
- Permiten estructurar las ideas del texto y establecer las relaciones entre ellas. Por tanto, facilitan el estudio y repaso posterior.

2.2.2 Clasificación de los resúmenes

Existen diferentes tipos de resúmenes debido a dos razones: la primera depende del tipo y fuente del documento de entrada/salida, mientras la segunda se relaciona al

propósito o el uso por el cual el resumen es generado. Actualmente existen muchas maneras de realizar resúmenes, según los criterios de búsqueda se tienen las siguientes clasificaciones (Cardoso & Pérez-Abelleira, 2013; Nazari & Mahdavi, 2019; Zainal Arifin et al., 2018):

- **De acuerdo con la salida:** El resumen se puede generar de manera abstractiva, extractiva, o híbrida. A continuación, se proporciona una descripción detallada de cada tipo de resumen, así como el proceso que conlleva su generación.
    - **Resumen abstractivo:** Para el resumen abstractivo, este debe ser generado interpretando el contenido extraído del texto. Los métodos o modelos que generan resúmenes abstractivos usualmente emplean nuevas palabras y oraciones obtenidas de otros corpus o diccionarios para presentar la información principal de los documentos de entrada. En comparación con los métodos que generan resúmenes extractivos, el proceso de la obtención del resumen abstractivo es más similar al de los escritos por los humanos. Sin embargo, la GAR abstractivos requiere técnicas de generación y comprensión de lenguaje natural, cómo paráfrasis y fusión de oraciones.
    - **Resumen extractivo:** Los resúmenes extractivos son generados mediante una selección de oraciones. Cada oración extraída se caracteriza según el grado de información que aporta. Por lo tanto, los métodos y modelos que generan resúmenes extractivos clasifican las oraciones de acuerdo con su valor de importancia y se seleccionan a las mejores, conteniendo la información principal y reduciendo la redundancia (Roul, 2020). Por otro lado, en el resumen extractivo no se introduce vocabulario nuevo (Sanchez-Gomez et al., 2021).
    - **Resumen híbrido:** Los resúmenes híbridos son generados a través de una combinación de técnicas de abstracción y selección de oraciones. Por lo general, los modelos que generan resúmenes híbridos procesan el texto en dos etapas: *extracción-abstracción* y *abstracción-abstracción*. De esta manera, se intenta recopilar información relevante de los documentos fuente con métodos extractivos o abstractivos en la primera etapa, lo que puede reducir significativamente la longitud de los documentos de entrada. Posteriormente,

los textos procesados del paso anterior se introducen en un modelo abstractivo para formar resúmenes (Ma et al., 2020).

- **Según su función:** Los resúmenes pueden ser indicativos o informativos. Los indicativos son aquellos que proporcionan información de un documento de forma general (es similar a una tabla de contenidos). Además, es útil cuando el usuario no está seguro si es necesario leer el documento fuente o no. Mientras que el resumen informativo es una versión corta que refleja el contenido general de uno o varios documentos, sin necesidad de incluir información específica o detallada (El-Kassas et al., 2021).
- **De acuerdo con su entrada:** Un resumen puede ser construido a partir de un solo documento de texto (GARDI), o un conjunto de documentos (GARMD). La GARMD es más compleja que la GARDI ya que suele abordar la reducción de redundancia, aumento de cobertura, la relación temporal del contenido y la tasa de reducción de información que el resumen debe cumplir (El-Kassas et al., 2021).

2.2.3 Generación Automática de Resúmenes de Múltiples Documentos

La GARMD es una tarea que consiste en generar un resumen a partir de un grupo de dos o más documentos, el cual contiene y representa la información más relevante sobre un tema en particular (Ledeneva & García-Hernández, 2017; Saggion & Poibeau, 2013). Desde un punto de vista formal, la GARMD consiste en generar un resumen conciso e informativo a partir de un conjunto de documentos $D = \{d_i | \, i \in [1, N]\}$, los cuales están relacionados a un tema en general, donde N es el número de documentos. Cada documento d_i consta de M oraciones $\{s_{i,j} | \, j \in [1, M]\}$, donde $s_{i,j}$ se representa a la j-ésima oración en el i-ésimo documento (Ma et al., 2020).

En los inicios de la GAR, los resúmenes se generaban a partir de un solo documento, ya sea de una noticia, un artículo científico, un programa de difusión o una conferencia. A medida que avanza la investigación e incrementa el volumen de información de la internet, surgió la GARMD y se aplicó a grupos de artículos de noticias sobre el mismo evento, con el objetivo de producir un resumen breve de varios documentos. Gran parte del trabajo hasta la fecha se ha realizado en el contexto de resumen genérico,

haciendo pocas suposiciones sobre la audiencia o el objetivo para generar el resumen. Además, la GARMD se ha enfocado en resumir información de diferentes medios digitales, tales como sitios web, revistas electrónicas, periódicos, libros de texto, etc. (Haque et al., 2013).

Para que un grupo de documentos se resuma, es más completo y preciso generar un resumen a partir de múltiples documentos escritos en diferentes momentos, cubriendo diferentes perspectivas (Ma et al., 2020). Sin embargo, desde un punto de vista técnico, la GARMD es más complicada y difícil de abordar que la GARDI ya que las dos cuentan con diferencias significativas que se describen a continuación.

2.2.4 Diferencias entre la GARDI y la GARMD

Aunque la mayoría de los métodos de la GARMD han sido utilizados en la GARDI, se identifican las siguientes diferencias (Alguliev et al., 2013; Baldwin & Ross, 2001; Goldstein et al., 2000; Mcdonald, 2007; Nazari & Mahdavi, 2019; Villatoro-Tello et al., 2009):

- **Mayor grado de redundancia:** Dentro de un grupo de documentos relacionados temáticamente, existe un mayor grado de redundancia ya que cada documento puede describir la misma idea principal.
- **Dimensión temporal:** Un grupo de documentos contempla una dimensión temporal sobre un evento en desarrollo. Por lo tanto, la información posterior puede anular información previa y así cambia la perspectiva del suceso. De esta manera, los documentos pueden ser complementarios, superpuestos y contradictorios entre sí (Ma et al., 2020). Esto se debe a que hay información más diversa e incluso contradictoria entre documentos.
- **Mayor nivel de compresión de información:** La relación entre el tamaño del resumen y la cantidad de información contenida en el grupo de documentos de entrada es significativa. Para la GARMD, es más difícil generar un resumen de varios documentos que de uno solo en cuestión. Por esta razón, los modelos que generan resúmenes suelen degradarse. Para estos modelos representa un desafío retener contenido relevante de secuencias de entrada largas y complejas, ya que a su vez deben generar resúmenes coherentes, no redundantes y sin errores fácticos.

2.2.5 Requerimientos de la GARMD

El principal objetivo de la GAR crear sistemas que puedan generar resúmenes similares a como los realizaría el experto humano. Para que un resumen sea generado de múltiples documentos, se consideran los siguientes requerimientos (Goldstein et al., 2000;(Kumar et al., 2021; Yadav et al., 2023) Over et al., 2007; Wang et al., 2010):

- **Cobertura:** Se define como el grado en que el resumen (creado automáticamente) transmite la misma información que otro(s) (documentos fuente).
- **Cohesión:** Es la capacidad de combinar los documentos de la colección para que sea útil al lector. Dentro de este requerimiento, se contemplan los siguientes criterios:
  - **Orden de oraciones:** Se refiere a la capacidad de ordenar oraciones en función del nivel de ponderación de cada oración. Por lo tanto, las oraciones más relevantes son aquellas que tengan la ponderación más alta.
  - **Principio de la noticia:** Se refiere a que primero se presente la información más relevante y diversa para que el lector obtenga el máximo contenido de información, incluso si deja de leer el resumen.
  - **Línea temporal:** Se refiere a considerar el orden en los documentos de la colección según la ocurrencia de eventos en el tiempo.
- **Relevancia de la oración:** Este requerimiento se refiere a que las oraciones seleccionadas deben tener representatividad del tema tratado.
- **Reducción de redundancia:** Se refiere a no incluir oraciones que transmitan la misma información.

2.2.6 Métodos de concatenación de documentos para la GARMD

Para la GARMD, se requiere hacer un análisis sobre la manera de organizar los documentos de entrada. Por esta razón, se utilizan los siguientes métodos de concatenación de documentos.

- **Concatenación Plana:** La concatenación plana es un método de concatenación simple pero poderoso, que consiste en unir los documentos de entrada sin algún orden establecido. La introducción de documentos concatenados planos requiere

de modelos que tengan una gran capacidad de procesamiento de secuencias largas (Ma et al., 2020).

- **Concatenación jerárquica:** A diferencia de la concatenación plana, la concatenación jerárquica preserva la relación cronológica de los documentos de entrada, en lugar de simplemente concatenar documentos sin algún orden. Por lo tanto, facilita que el modelo obtenga una representación rica semánticamente, lo que a su vez mejora la efectividad de los modelos (Ma et al., 2020).

2.3 Ingeniería de características

Las características juegan un papel fundamental para la GARMD, porque ayudan a determinar las ideas clave del texto fuente que se presentarán como resumen. Es decir, las características se extraen de unidades de texto (por ejemplo, palabras, oraciones, párrafos o documentos) de acuerdo con parámetros o criterios que se utilizan para determinar la puntuación de cada unidad de texto. Entre las características más utilizadas se encuentran las siguientes:

- **Frecuencia de términos (TF):** Las palabras con mayor frecuencia en un documento indican cuales son los tópicos más importantes del mismo. La TF se encarga de asignar un puntaje de ocurrencia de cada palabra o término en cada oración del documento en función de su relevancia (Cajueiro et al., 2023; El-Kassas et al., 2021; Hendrastuty & SN, 2021; Hernandez-Castaneda et al., 2020; Mutlu et al., 2019) , como se muestra en la Ecuación 1, donde tf_{ij} es la frecuencia del término j en la oración i.

$$TF(t_j) = tf_{ij}$$

Ecuación 1. Frecuencia de términos.

- **Frecuencia inversa de documentos (IDF):** Esta ponderación se define como la relación entre el número de documentos donde aparece el término y el número de documentos que tiene la colección. La Ecuación 2 describe esta ponderación, donde n_j es el número de documentos u oraciones en los que aparece el término j; N es el número de documentos u oraciones.

$$IDF(t_j) = \log\left(\frac{N}{n_j}\right)$$

Ecuación 2. Frecuencia inversa de documentos.

Esta característica hace que los términos sean más importantes cuando aparecen en pocos documentos, lo que permite caracterizar de mejor manera a cada documento u oración. Por lo tanto, se mide la distribución de cada término en el (los) documentos (Cajueiro et al., 2023; El-Kassas et al., 2021; Hernandez-Castaneda et al., 2020; Mutlu et al., 2019)

- **Frecuencia de términos – Frecuencia Inversa de Documento (TF-IDF):** Consiste en multiplicar TF e IDF de cada término en cuestión ($TF \times IDF$). Es común que ambas ponderaciones se utilicen juntas para determinar la relevancia de cada término, considerando tanto la importancia que tiene el término en la colección de documentos como su importancia en ese documento u oración (Cajueiro et al., 2023; El-Kassas et al., 2021; Hendrastuty & SN, 2021; Hernandez-Castaneda et al., 2020; Mutlu et al., 2019).
- **Frecuencia normalizada de etiquetas POS o NER:** Además de TF, la frecuencia de ciertas etiquetas POS (por ejemplo, sustantivos, pronombres, verbos) o NER (por ejemplo, nombres propios, organizaciones, lugares) puede indicar la relevancia de palabras que conforman una oración. Sin embargo, la frecuencia de algunas palabras que no aportan poca información puede representar ruido en la selección de oraciones. Por lo tanto, se normaliza dicha frecuencia usando la Ecuación 3:

$$TF(t_i) = \frac{t_i}{N}, t_i \in \{POS, NER\}$$

Ecuación 3. Frecuencia normalizada de etiquetas POS o NER.

Donde t_i es la frecuencia de la i-ésima etiqueta en el documento u oración; N representa el número total de etiquetas que contempla el documento d (Cajueiro et al., 2023; Jain et al., 2022; Ray et al., 2023; Wang et al., 2020; Yohannes & Amagasa, 2022). Cada etiqueta t pertenece al conjunto de etiquetas POS o NER, que se enlistan y describen en el Anexo 2.

- **Longitud de oración:** Generalmente, las oraciones cortas no son consideras en el resumen final, porque en la mayoría de los casos no contienen información

significativa (Aote et al., 2023; Hendrastuty & SN, 2021).Por lo tanto, esta característica mide la relación entre el número de palabras de la oración k y la cantidad de palabras de la oración más larga del documento (ver Ecuación 4).

$$Longitud\ de\ las\ oraciones\ (k) = \frac{Número\ de\ palabras\ en\ la\ oración\ \ k}{Cantidad\ de\ palabras\ de\ la\ oración\ más\ larga\ del\ documento}$$

Ecuación 4. Longitud de las oraciones

- **Posición de las oraciones:** Considera que la importancia de una oración disminuye con su distancia respecto al principio del documento. Esta característica se obtiene mediante la Ecuación 5 (Kato et al., 2007; Kiyani & Tas, 2017), donde k es el número que representa la posición de la oración en el documento.

$$Posición\ de\ la\ oración\ (k) = \frac{Número\ total\ de\ oraciones - k}{Número\ total\ de\ oraciones\ en\ el\ documento}$$

Ecuación 5. Posición de la oración

- **Co-ocurrencia entre las palabras de la oración y el título:** Otorga una puntuación alta a las oraciones que contienen palabras en común con respecto al título (Akhtar et al., 2020b; Gao et al., 2020; Kato et al., 2007; Kiyani & Tas, 2017). Esta característica se calcular de la siguiente manera (ver Ecuación 6):

$$Palabras\ oracion \cap Palabras\ título = \frac{Número\ de\ términos\ en\ común\ entre\ el\ título\ y\ la\ oración}{Longitud\ del\ título\ del\ documento}$$

Ecuación 6. Similitud con el titulo

- **Datos numéricos:** La oración que tiene datos numéricos puede contener información importante de los documentos (Cajueiro et al., 2023; Hendrastuty & SN, 2021; Jain et al., 2022; Ray et al., 2023; Wang et al., 2020; Yohannes & Amagasa, 2022). Esta característica es calculada como se muestra en la Ecuación 7.

$$Datos\ Numéricos\ (k) = \frac{Número\ de\ datos\ numericos\ \in oración\ k}{Longitud\ de\ la\ oración\ k}$$

Ecuación 7. Datos numéricos

- **Palabras temáticas:** Son palabras específicas de dominio con la máxima relatividad posible. Por lo tanto, la característica que mide la presencia de palabras temáticas se representa por medio de la relación entre el número de palabras temáticas que

aparecen en una oración, y el número máximo de palabras de una oración (AL-Khassawneh & Hanandeh, 2023; Antony et al., 2023; Aote et al., 2023; Mojrian & Mirroshandel, 2021), como se muestra en la Ecuación 8.

$$Palabras\ Temáticas\ (k) = \frac{Número\ de\ datos\ temáticos\ \in oración\ k}{Número\ máximo\ de\ palabras\ temáticas}$$

Ecuación 8. Palabras temáticas

- **Nombres Propios:** Por lo general, la oración que contiene más nombres propios es importante y lo más probable es que se incluya en el resumen del documento Cajueiro et al., 2023; Jain et al., 2022; Ray et al., 2023; Wang et al., 2020; Yohannes & Amagasa, 2022).

$$Nombres\ Propios\ (k) = \frac{Número\ de\ nombres\ propios\ \in oración\ k}{Longitud\ de\ la\ oración\ k}$$

Ecuación 9. Nombres Propios

- **Palabras clave positivas en la oración:** Dado que las palabras son elementos básicos de una oración, las palabras clave son aquellas que aparecen con mayor frecuencia de uno o varios documentos. Por lo tanto, cuantas más palabras clave tiene una oración, más importante es la oración. Con base en esta premisa, la ponderación de palabras clave positivas se calcula de acuerdo con la Ecuación 10 (El-Kassas et al., 2021; Hendrastuty & SN, 2021; Taieb-MaimMeiravon et al., 2023) donde s representa el contenido de la oración segmentada por n palabras. tf_i es la frecuencia de la palabra i, la cual es multiplicada por su respectiva probabilidad P.

$$Palabras\ positivas(s) = \frac{1}{longitud(s)} \sum_{i=1}^{n} tf_i \times P(s \in S | Palabra\ clave_i)$$

Ecuación 10. Ponderación de palabras clave positivas en una oración.

- **Reducción de Redundancia:** Para eliminar la redundancia en las oraciones del resumen, se plantea la Ecuación 11, que es empleada en diversos trabajos del estado del arte (Abualigah et al., 2020; Aote et al., 2023; Mohamed & Oussalah, 2019)

$$SC(S_i, S_j) = \frac{\sum_{k=1}^{t} w_{ik}\, w_{jk}}{\sqrt{\sum_{k=1}^{t} (w_{jk})^2\ \sum_{k=1}^{t} (w_{ik})^2}}$$

Ecuación 11. Reducción de redundancia

Donde k es el término dentro de un vocabulario t, w_{ik} indica la frecuencia del término k en la oración $S_{i,}$ y w_{jk} la frecuencia del término k en la oración S_j. En general, esta ecuación de similitud mide el grado de semejanza entre dos oraciones. La similitud entre ellas es representada por un valor numérico, regularmente entre 0 y 1, donde los valores más cercanos a 1 indican una similitud más alta entre ambas secuencias.

- **Similitud con la oración principal (central):** Una oración puede considerarse principal o central si ésta muestra una alta semejanza con otras oraciones del(los) documento(s) de entrada (Fattah & Ren, 2008). Por lo tanto, esta ponderación establece que mientras una oración tenga una alta similitud con la oración principal, es más probable que se considere como parte del resumen final. Generalmente se utiliza la similitud coseno (ver Ecuación 11) para calcular la semejanza entre oraciones.

2.4 ¿Cómo se generan resúmenes extractivos a través de un método computacional?

Cuando los humanos escribimos resúmenes somos capaces de comprender el contenido del texto fuente, hacer inferencias y generalizar información. Sin embargo, una computadora carece de estos conocimientos y habilidades. Por lo que es necesario diseñar los procesos cognoscitivos de comprensión del lenguaje. En la GAR, este diseño consiste en preprocesar el texto para crear una representación intermedia del texto original. Posterior a esta actividad se identifican 2 etapas: Modelado de características y selección de oraciones.

2.4.1 Preprocesamiento

El preprocesamiento es la etapa en la que se ingresa el texto fuente sin alteraciones o filtrado de información. El texto se transforma a una forma semiestructurada o estructurada, obteniendo representaciones sencillas que facilitan su análisis. Generalmente, se realiza una limpieza y normalización de (Abualigah et al., 2020; AL-

Khassawneh & Hanandeh, 2023; Ni et al., 2020; Wang et al., 2020). Entre los procesos considerados como preprocesamiento, se encuentran los siguientes:

- **Segmentación de oraciones:** El texto de entrada es segmentado por oraciones, las cuales son unidades fundamentales de información para generar resúmenes extractivos (Alguliyev et al., 2018; Jain et al., 2022; Kumar et al., 2021; Mutlu et al., 2020).
- **Tokenización de palabras:** El proceso de tokenización consiste dividir el texto en tokens o palabras, donde cada una aporta información específica y es útil para caracterizar cada oración. Los delimitadores de los tokens pueden ser símbolos especiales como espacio, punto, coma, dos puntos, punto y coma, guion, corchete de cierre, comillas y signo de exclamación, entre otros (Alami et al., 2020; Jain et al., 2022; Yadav et al., 2023).
- **Reconocimiento de entidades nombradas (NER, por sus siglas en inglés):** El reconocimiento de entidades nombradas, es una tarea de PLN, para clasificar texto en categorías semánticas predefinidas tales como personas, organizaciones, fechas y lugares. Las palabras que son etiquetadas se denominan entidades (Cajueiro et al., 2023; Jain et al., 2022; Ray et al., 2023; Wang et al., 2020; Yohannes & Amagasa, 2022). NER actúa como un importante paso de preprocesamiento para una variedad de aplicaciones posteriores, como recuperación de información, respuesta a preguntas, traducción automática, etc.
- **Normalización:** Dentro de este procedimiento, el texto se normaliza de diferentes maneras, entre las que se encuentran: eliminación de acentos para facilitar la búsqueda y equiparación de cadenas, conversión de mayúsculas a minúsculas o viceversa (Abualigah et al., 2020; Aote et al., 2023; Hendrastuty & SN, 2021)
- **Stemming:** Este procedimiento consiste en reducir cada palabra del texto de entrada a su forma raíz (AL-Khassawneh & Hanandeh, 2023; Antony et al., 2023; Sanchez-Gomez et al., 2022). Generalmente, cada palabra obtenida como raíz aporta información común de los temas en un documento y facilita la comparación entre cadenas.

- **Filtrado de palabras vacías (stopwords):** Este proceso consiste en descartar palabras que carecen de información semántica como artículos, preposiciones, etc. (El-Kassas et al., 2021; Kumar et al., 2021; B. Li et al., 2022; Ni et al., 2020). Por lo tanto, no son necesarias para determinar la relevancia de una oración. El Anexo 1 enlista las Stopwords usadas en este trabajo para el inglés.
- **Etiquetado gramatical:** Es una tarea fundamental en el PLN, consiste en identificar con una etiqueta gramatical (POS, por sus siglas en inglés) a cada palabra de una oración. Este procedimiento es usado para diferenciar cantidades, correos, sustantivos, etc. (Brill, 1992; Sharipov et al., 2023; Singh et al., 2021)
- **Vectorización Word2Vec:** Las palabras son modelos de representación muy usados y útiles en varias tareas del PLN, pero carecen de información semántica sobre los textos o documentos. Para extraer la información contextual de palabras sobre un tema, se utilizan vectores preentrenados de palabras obtenidos de word2vec. Word2Vec es un conjunto de herramientas de código abierto basado en redes neuronales para producir vectores de palabras. Es decir, para cada palabra de entrada, se le asocia un vector de números reales que representa la información contextual de dicha palabra. Actualmente, word2Vec se usa ampliamente en tareas del PLN como identificación de sinónimos, análisis de sentimientos y clasificación de texto (Chengzhang & Dan, 2018; Haider et al., 2020; Rong, 2014). Además, puede utilizar el modelo continuo de bolsa de palabras (CBOW) y el modelo basado en saltos de n-gramas continuo (skip-gram) para producir vectores de palabras

  La idea principal de CBOW es predecir la probabilidad de la palabra central sobre la base del contexto alrededor de la palabra, y el modelo skip-gram es omitir algunos símbolos alrededor de la palabra central para predecir el contexto. Si bien el primero requiere mucho tiempo y el efecto del entrenamiento está limitado por el tamaño del vector, el modelo skip-gram tiene una mejor precisión semántica, pero la complejidad computacional es alta y lleva mucho tiempo el proceso de entrenamiento (Haider et al., 2020).

2.4.2 Modelado de características

El modelado de características es un proceso importante para convertir un formato textual no estructurado en uno estructurado (Hendrastuty & SN, 2021). La calidad de un resumen generado depende de la relevancia que se otorga a un conjunto de características (Abuobieda et al., 2012). Una de las cuestiones importantes en esta etapa es qué características se deben considerar en el proceso de resumen. Para determinar qué características deben ser incluidas en un método deben considerarse los distintos niveles textuales.

2.4.3 Niveles de Ponderación de las características.

Las características mencionadas en la sección 2.3, se clasifican en cuatro niveles que determinan la calidad de una oración (Qaroush et al., 2021). El nivel basado en palabras, el nivel basado en oraciones, el nivel basado en párrafos y las características basadas en grafos o documentos (ver figura 1).

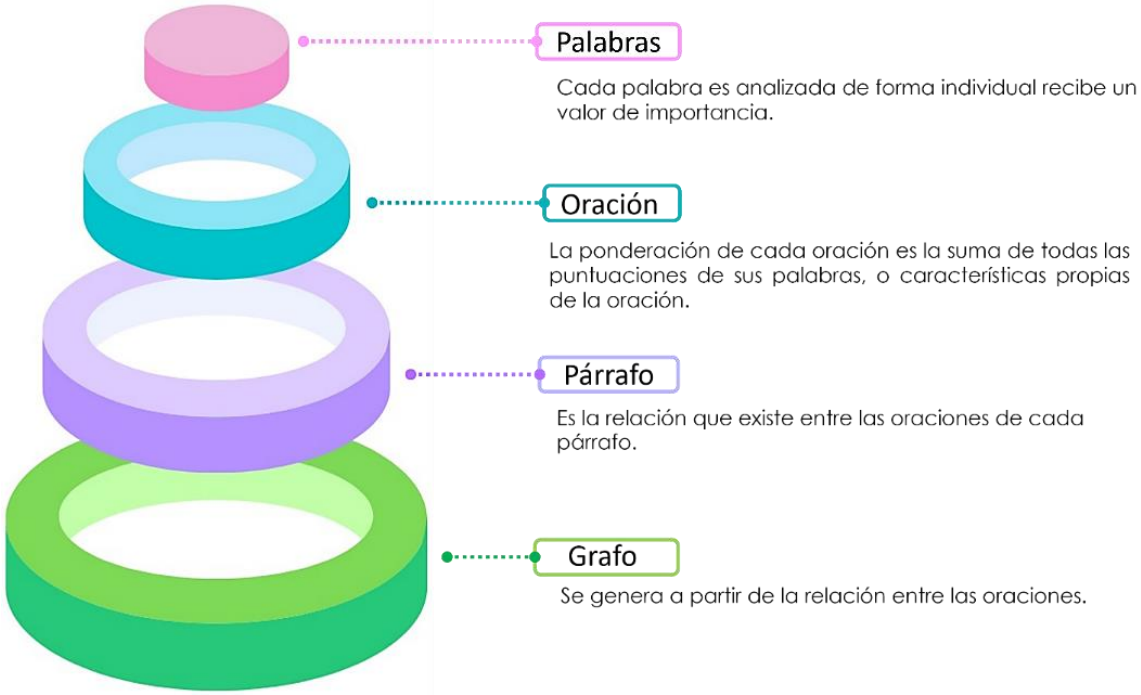


Figura 1. Niveles de ponderación de características

1. En el nivel de palabras, se incluyen características que analizan a las palabras de forma individual asignándoles un valor de importancia.

2. En el nivel oración, se considera la suma de las ponderaciones de las palabras que componen a la oración, así como características relativas a la oración. Como longitud de la oración, su posición dentro del documento, y la similitud entre oraciones.
3. El nivel párrafo describe la relación que existe entre las oraciones que componen al párrafo.
4. Finalmente, el nivel grafo describe la relación que existe entre todas las oraciones que conforman el documento.

2.4.4 Niveles en la estructura del lenguaje

Un lenguaje se puede definir de diferentes formas, sin embargo, desde el punto de vista formal, se define como un conjunto de frases, que generalmente es infinito y se forma con combinaciones de elementos llamado alfabeto, respetando un conjunto de reglas de formación (sintácticas o gramaticales) y de sentido (semánticas).

Lo que es conocimiento para el humano no lo es para las computadoras. Una computadora puede almacenar, copiar, borrar datos y archivos, pero no puede comprender el contenido, hacer inferencias, ni generalizar como hacemos los humanos. Por ello, es necesario modelar los procesos cognoscitivos de la comprensión del lenguaje para el diseño de métodos que realizan tareas lingüísticas como la GAR (Kaljahi et al., 2014). Por lo que es necesario involucrar los diferentes niveles en la estructura del lenguaje porque a través de ellos la computadora interpreta y analiza el contenido que se le proporciona (Abdulateef et al., 2020). La siguiente figura muestra los diferentes niveles de estructura lingüística y posteriormente son descritas.

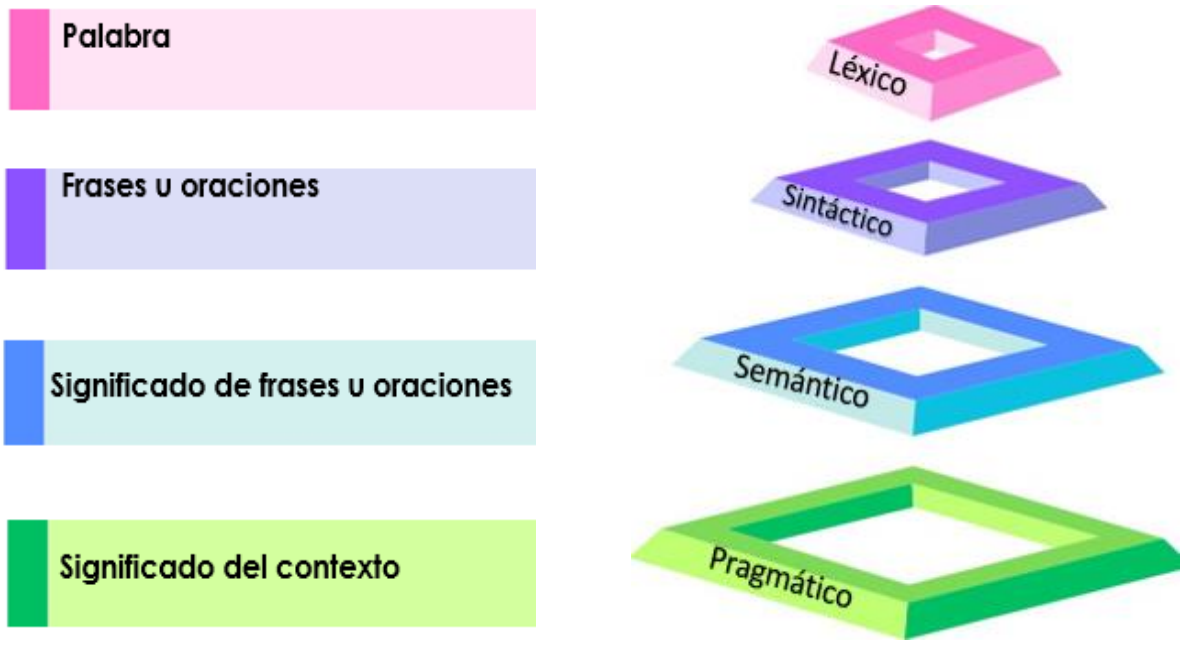


Figura 2. Niveles de estructura lingüística

La estructura lingüística se sustenta por los niveles léxico, sintáctico, semántico y pragmático:

- **Nivel Léxico:** Trata de cómo las palabras se construyen a partir de unidades de significado más pequeñas llamadas morfemas.
- **Nivel Sintáctico:** Se refiere a cómo las palabras pueden unirse para formar oraciones.
- **Nivel Semántico:** Trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independientemente del contexto, es decir la oración aislada.
- **Nivel Pragmático:** Trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un nivel recursivo que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

La estructura del lenguaje ayuda en la interpretación y análisis de las oraciones proporcionadas a un método

2.4.5 Selección de oraciones

Esta etapa consiste en seleccionar el mejor subconjunto de oraciones del texto fuente. el objetivo es extraer oraciones que cubran tantos conceptos importantes como sea posible. Para ello se asignan puntajes a cada oración o palabra. La puntuación de una oración representa qué tan bien la oración explica algunos de los temas más importantes del texto fuente. Los puntajes pueden ser de origen estadístico, lingüístico o una combinación de ambos. Por lo tanto, las oraciones altamente puntuadas son elegidas para formar el resumen, (A. Gupta et al., 2019; V. Gupta & Singh-Lehal, 2010; Ledeneva & García-Hernández, 2017; Mutlu et al., 2019a; Vázquez et al., 2018). No obstante, la selección de oraciones que conforman al resumen final se realiza mediante métodos de optimización o aprendizaje profundo.

2.5 Complejidad de la GARMD

Dentro de la GARMD, la identificación y formalización del problema en la selección de oraciones es fundamental y relevante para los métodos de optimización. Esta tarea se realiza a partir de la teoría de la complejidad computacional, que es la rama de la teoría de computación que estudia los recursos, o coste de computación requerido para resolver un problema dado y definir la solución se puede dar a partir de un método determinista o no determinista (Maldonado, 2013).

El término determinista significa que sin importar lo que haga el algoritmo, sólo hay una cosa que puede hacer a continuación. Los tipos de problemas que se pueden resolver de manera determinista pertenecen a la clase **P**, puesto que su complejidad puede estar acotada por un polinomio. Además, un problema de este tipo puede ser resuelto en tiempo polinomial. Por el contrario, un problema pertenece a la clase **NP** si es resuelto en tiempo polinomial, pero usando métodos no deterministas (Coello, 2004).

Los métodos o modelos no deterministas no existen en el mundo real. El no determinismo es una herramienta imaginaria que hace que los problemas difíciles parezcan triviales. Su mayor valía radica en el hecho de que existe forma de convertir un algoritmo no determinista a uno determinista, aunque a un costo computacional que suele ser muy elevado (Du & Swamy, 2016).

Por otro lado, existe otra clase de problemas denominados **NP Completos** que resultan de gran interés en computación. Estos problemas son muy difíciles de resolver, debido a que el espacio de búsqueda del problema crece conforme a la expresión $\frac{(n-1)!}{2}$, algunos ejemplos de lo que esto significa, son los siguientes:

- Para n = 10, hay unas 181,000 soluciones posibles.
- Para n = 20 hay unas 10,000,000,000,000,000 soluciones posibles.
- Para n = 50, hay unas 100, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000 soluciones posibles.

Como se mencionó anteriormente, la GARMD puede considerarse como una extensión de la GARDI. Sin embargo, la GARMD es más complicada, debido a la cantidad de oraciones contenidas en los documentos de la colección (Alguliev et al., 2013; Mihalcea & Tarau, 2005; Villatoro-Tello et al., 2009). En el estado del arte se ha demostrado que cuanto más largo es un texto, mayor es el número de combinaciones de oraciones (Matías, 2013). La dificultad para resolver este problema radica en la complejidad computacional que conlleva encontrar su solución, ya que no existe un solo procedimiento para encontrar la mejor selección de oraciones. Esto implica que los métodos deterministas de resolución no sean capaces de resolverlo.

2.6 Algoritmos Evolutivos

Los Algoritmos Evolutivos (AE), están inspirados en la capacidad de las especies que tienen para evolucionar y adaptarse a su entorno. En cada iteración, aplican un conjunto de operadores simulando procesos evolutivos. Además, son algoritmos metaheurísticos basados en una población de individuos que emplean mecanismos biológicamente inspirados como la mutación, recombinación, selección natural y supervivencia de los más aptos para ir iterativamente ajustando o refinando un conjunto de soluciones (Andréasson et al., 2005; Emmerich & Deutz, n.d.; *Search and Optimization by Metaheuristics*, 2016).

La mayoría de los AE son heurísticas de búsqueda y optimización derivados de la teoría de evolución clásica. La idea básica es que, si solo se reproducen aquellos individuos de una población que cumplen con ciertos criterios de selección, y los otros individuos

de la población mueren, la población convergerá a aquellos individuos que mejor cumplan con los criterios de selección Los individuos con mejor aptitud tendrán mayor probabilidad de producir descendientes. Esta idea se corresponde con el principio de la evaluación natural: supervivencia de los más adecuados, que permite a la naturaleza adaptarse a entornos cambiantes.

2.6.1 Algoritmos Genéticos (AGs)

Son algoritmos estocásticos que implementan métodos de búsqueda a partir de la herencia genética y el principio Darwiniano de la supervivencia de los más aptos. Los AGs no buscan modelar la evolución biológica sino derivar estrategias de optimización. El concepto se basa en la generación de poblaciones de individuos mediante la reproducción de los padres. Los AGs codifican una posible solución a un problema específico en una estructura de datos conocida como cromosoma o individuo. Para cada individuo, se aplican operadores de recombinación a fin de preservar la información crítica. Los AGs son a menudo vistos como optimizadores de funciones, por lo cual existen una amplia diversidad de problemas a los que se han aplicado. Las aplicaciones más comunes de AGs son la solución de problemas de optimización, donde han mostrado resultados eficientes y fiables (Ledeneva & García-Hernández, 2017).

En Sivanandam & Deepa (2008), se mencionan ventajas que posee esta técnica, las cuales se enlistan a continuación:

- La aleatoriedad juega un papel esencial en los AGs, ya que los operadores de selección, cruza/reproducción y mutación necesitan procedimientos aleatorios.
- Un segundo punto muy importante es que los AGs siempre consideran una población de soluciones. Manteniendo en la memoria más de una sola solución en cada iteración ofrece muchas ventajas. El algoritmo puede recombinar diferentes soluciones para obtener mejores y así, aprovechar la información que proporcionan diversas soluciones. Un algoritmo de base poblacional también es muy susceptible de paralelización. La solidez del algoritmo también debe mencionarse como algo esencial para el éxito del algoritmo.

- Función de aptitud: También llamada *índice de rendimiento* o *criterio de elección*. Este es el elemento utilizado para decidir los valores adecuados de las variables de decisión que resuelven el problema de optimización. La función objetivo permite determinar los mejores valores para las variables de decisión (Streichert, 1995).

2.6.2 Operadores del AG

Para el correcto funcionamiento del AG, se requiere de un conjunto de operadores genéticos, los cuales ayudan a realizar una optimización en la búsqueda de soluciones. A continuación, se describen los operadores comunes del AG.

Una parte fundamental del funcionamiento de un AG es, sin lugar a duda, el proceso de selección de individuos. Este proceso suele realizarse de forma probabilística (es decir, aun los individuos menos aptos tienen una cierta oportunidad de sobrevivir), a diferencia de las estrategias evolutivas, en las que la selección es extintiva (los menos aptos tienen cero probabilidades de sobrevivir) (Coello, 2004). Los operadores de selección comúnmente usados en el AG son los siguientes:

- **Ruleta:** Este operador se encuentra dentro de los operadores de selección proporcional, en los cuales se eligen individuos de acuerdo a su contribución de aptitud con respecto al total de la población (Coello, 2004; Holland, 1992). Este operador fue propuesto en (De Jong, 1975) y ha sido el método más comúnmente usado desde los orígenes de los AGs. El algoritmo es simple, pero eficiente (su complejidad es $O(n^2)$. El algoritmo del operador Ruleta es el siguiente (Coello, 2004; Du & Swamy, 2016).

  Calcular la suma de valores esperados T

  1. Repetir N veces (N es el tamaño de la población):
     a. Generar un numero aleatorio r entre 0.0 y T
     b. Ciclar a través de los individuos de la población sumando los valores esperados hasta que la suma sea mayor o igual a r.
     c. El individuo que haga que esta suma exceda el límite es el seleccionado.

- **Torneo:** Los métodos de selección proporcional, requieren de dos pasos a través de toda la población en cada generación (Coello, 2004; Du & Swamy, 2016; Holland, 1992):
  1. Calcular la aptitud media (y, si se usa escalamiento sigma, la desviación estándar).
  2. Calcular el valor esperado de cada individuo.

  El uso de jerarquías requiere que se ordene toda la población (una operación cuyo costo puede volverse significativo en poblaciones grandes). Esta técnica fue propuesta por Wetzel (1983). La idea básica del método es seleccionar con base en comparaciones directas de los individuos. El algoritmo es el siguiente (Coello, 2004):
  1. Barajar los individuos de la población.
  2. Escoger un número p de individuos (típicamente 2).
  3. Compararlos con base en su aptitud.
     a) El ganador del "torneo" es el individuo más apto.
     b) Debe barajarse la población un total de p veces para seleccionar N padres (donde N es el tamaño de la población).

Posterior a la selección, el AG pasa por un proceso de cruza de individuos. En los sistemas biológicos, la cruza es un proceso complejo que ocurre entre parejas de individuos. Estos individuos se alinean, luego se fraccionan en ciertas partes y posteriormente intercambian fragmentos entre sí. En computación evolutiva, se simula la cruza intercambiando segmentos de cadenas lineales de longitud fija. Aunque las técnicas de cruza básicas suelen aplicarse a la representación binaria, estas son generalizables a alfabetos de cardinalidad mayor (Coello, 2004; Du & Swamy, 2016).

Una vez realizada la selección de individuos, se procede con la mutación de estos. La mutación se considera como un operador secundario en el AG canónico. Es decir, su uso es menos frecuente que el de la cruza. En la práctica, se suelen recomendar porcentajes de mutación de entre 0.001% y 0.01%para la representación binaria. Algunos investigadores, sin embargo, han sugerido que el usar porcentajes altos de mutación al inicio de la búsqueda, y luego decrementarlos exponencialmente,

favoreciendo el desempeño de un AG. Otros autores sugieren que $p_m = \frac{1}{L}$ (donde L es la longitud de la cadena cromosómica) es un límite inferior para el porcentaje óptimo de mutación (Coello, 2004).

2.7 Evaluación de resúmenes

El ajuste de parámetros de cualquier método de la GARMD depende de qué tan parecido son los resúmenes que genera, con respecto a los resúmenes de referencia. De esta manera, se busca mejorar la calidad de los resúmenes automáticos y producir puntajes confiables y estables. Todos los métodos de evaluación automatizados existentes funcionan comparando el resumen automático con uno o más resúmenes de referencia escritos por humanos.

Comúnmente, la evaluación de resúmenes involucra utilizar diversas medidas de calidad que requieren juicios humanos, como coherencia, concisión, gramaticalidad, legibilidad y contenido. Sin embargo, realizar una evaluación manual simple de resúmenes a gran escala empleando criterios lingüísticos y cobertura de contenido requiere de más de 3.000 horas de esfuerzo humanos. Por lo que esto es muy costoso y difícil de realizar (Mani, 2001) .

Lin y Hovy (2007) propusieron ROUGE (Recall-Oriented Understudy for Gisting Evaluation), un método de evaluación automática que mide la similitud entre resúmenes generados de forma automática y resúmenes generados por expertos humanos. Dentro de ROUGE, el método de evaluación más común es ROUGE-N, el cual mide la co-ocurrencia de n-gramas entre un resumen candidato (generado con una herramienta de la GAR) y uno o más resúmenes de referencia generado por humanos. Formalmente ROUGE-N se calcula de la siguiente manera:

$$ROUGE - N = \frac{\sum s \in \{Reference\ Summaries\} \sum gram_n \in s\ Count_{match}\ (gram_n)}{\sum s \in \{Reference\ Summaries\} \sum gram_n \in s\ Count(gram_n)}$$

Ecuación 12. Fórmula para calcular ROUGE-N.

Donde n representa la longitud del n-grama, $(gram_n)$ y $Count_{match}\ (gram_n)$ es el número máximo de $n - gramas$ que se producen conjuntamente en un resumen candidato y un conjunto de resúmenes candidato y un conjunto de resúmenes de referencia. Está claro

que ROUGE-N es una medida relacionada con la especificidad o Recall porque el denominador de la ecuación es la suma total de la cantidad de n-gramas que ocurren en el lado del resumen de la referencia.

2.8 Resumen del capítulo

Dentro de este capítulo se describieron varios conceptos fundamentales sobre el PLN y la GARMD, con el propósito de contextualizar el tema de estudio de esta tesis. De manera general, el PLN es una ciencia que se apoya de la lingüística computacional y estudia la comprensión del lenguaje, mismo que puede ser verbal, escrito, o incluso en imágenes. Donde el lenguaje natural es transformado a una representación para manipularlo.

La GAR, que tiene por objetivo que un software acepte un texto fuente, extraiga la información más importante y la presente al usuario. Entre las ventajas que brindan los resúmenes de texto se encuentran: reducen el de tiempo de lectura, ayudan a búsquedas más sencillas y con resultados más precisos, son más imparciales. Además, facilitan el estudio en la academia.

Existen diversos tipos de resúmenes, esto obedece a dos motivos: el principal es que existen diferentes fuentes de documentos y la segunda, es que se le da diferentes usos al resumen. De acuerdo con su salida, los resúmenes pueden ser abstractivos, extractivos o híbridos. Por otro lado, los resúmenes pueden ser indicativos o informativos de acuerdo con su función. Mientras que, de acuerdo con su entrada, los resúmenes pueden ser generados de un solo documento (GARDI) o de múltiples documentos (GARMD). En esta última tarea, se requiere generar un resumen a partir de un grupo de dos o más documentos, donde la mayoría de las aplicaciones se encuentra bajo el dominio de noticias.

En este capítulo también se abordaron las diferencias entre la GARSD y GARMD. Las diferencias más importantes entre ambas es que la GARMD tiene que enfrentar mayor grado de redundancia, aborda la dimensión temporal, ya que un grupo de documentos con información posterior puede anular o modificar información previa. Además, la

GARMD se enfrenta a una mayor relación de compresión, la cual se refiere a la situación en que generar un resumen se vuelve significativamente más difícil cuando aumenta el volumen de los documentos.

Se expusieron las principales características que han sido propuestas en diversos trabajos para la GAR, entre las que se encuentran las siguientes:

- Frecuencia de Términos (TF)
- Frecuencia Inversa de Documentos (IDF)
- TF-IDF
- Frecuencia de etiquetas POS y NER
- Longitud de la oración
- Posición de la oración
- Co-ocurrencia entre las palabras de la oración y el título
- Datos numéricos
- Palabras temáticas
- Nombres propios
- Palabras clave positivas en la oración
- Reducción de redundancia
- Similitud con la oración principal o central

Por otro lado, se describen las etapas para generar resúmenes a través de métodos computacionales: Preprocesamiento, modelado de características y selección de oraciones.

Asimismo, se expuso la complejidad que involucra la tarea de la GAR y algunos conceptos relacionados con optimización que ayudan a conceptualizar los términos a lo largo de esta tesis. Uno de los métodos de optimización muy conocidos son los Algoritmos Genéticos (AGs). Los AGs realizan una búsqueda de soluciones, a partir de la herencia genética y el principio darwiniano de la supervivencia de los más aptos. Durante el proceso, el AG emplea operadores de selección, cruza y mutación, los cuales ayudan al proceso de búsqueda y selección de mejores soluciones.

Por último, este capítulo finaliza con la evaluación de resúmenes. Esta evaluación se realiza mediante una comparación entre el resumen generado de forma automática y resúmenes de referencia escritos por humanos, la cual es realizada a través de ROUGE, que mide la similitud entre ambos resúmenes. En particular, se utiliza ROUGE-N como una medida que utiliza n-gramas para calcular la similitud entre ambos documentos.

CAPÍTULO 3
Estado del arte

Desde mediados de los 90s, comenzó el desarrollo de programas que fomentan la GAR, tales como la conferencia de comprensión de documentos o DUC (*Document Understanding Conferences*) (Over & Dang, 2007) y la conferencia de análisis de texto o TAC (*Text Analysis Conferences*) (Galanis & Malakasiotis, n.d.) (Das & Martins, 2007). Desde entonces, han surgido varios esfuerzos que generan resúmenes a partir de múltiples documentos, utilizando diferentes modelados de características que ponderan la importancia de oraciones. A continuación, se presentan trabajos relacionados que han logrado avances significativos en el análisis, del modelado de características para la GARMD.

3.1 Modelado de características

Una de las partes más importantes en la GAR es el modelado de características, la cual tiene como objetivo determinar la relevancia de las oraciones que formarán parte de

un resumen (Ferreira, De Souza Cabral, et al., 2014; Sanchez-Gomez et al., 2021). Dentro de este modelado existen características de estadísticas y lingüísticas, cuyo uso en la GARMD depende de cómo se exploten y combinen, para diferenciar la relevancia de cada oración de los documentos fuente (Goularte et al., 2019). La cantidad de características que se han propuesto para la selección de oraciones es amplia, entre las que se encuentran:

- **Posición de las oraciones dentro del texto:** La información importante a menudo es cubierta por los escritores al principio de un documento. Por lo tanto, las oraciones iniciales generalmente contienen el contenido más relevante.
- **Cobertura:** Es la capacidad de buscar y extraer los puntos o tópicos principales de un conjunto de documentos. Mientras más puntos principales tenga una oración, esta se considera es más relevante.
- **Frecuencia de palabras:** La idea de considerar esta característica es que las palabras importantes aparecen con mayor frecuencia en el documento u oración.
- **Similitud con el título:** La semejanza léxica o semántica entre el título y una oración candidata a resumen indica que dicha oración es relevante e informativa.
- **Longitud de las oraciones:** La longitud de oración es una característica que ayuda a diferenciar la importancia de cada oración en el resumen. Usualmente las oraciones cortas no son consideradas en el resumen porque no contienen información relevante (Gandotra & Arora, 2021).
- **Reducción de redundancia:** Las oraciones que provienen de varios documentos teóricamente tienen problemas de redundancia. El manejo de la redundancia es un factor muy importante, debido a que el resumen generado debe evitar contener información repetida (Akhtar et al., 2020a; Alguliyev et al., 2015).
- **Coherencia:** Esta característica se encarga de medir la concordancia estructural de cada oración seleccionada en el resumen. Por lo que esta se mide una vez que se realice la selección de oraciones.
- **Inclusión de nombre propios:** Por lo general, la oración que contiene más nombres propios es importante y es altamente probable que se incluya en el resumen (Fattah & Ren, 2008).

- **Inclusión de palabras temáticas:** Como se mencionó anteriormente, las palabras temáticas son términos de dominio específico que resumen las ideas principales de uno o varios documentos de entrada. Tomando como referencia esta característica, una oración se considera altamente relevante si contiene varias palabras temáticas.
- **Aparición de verbos y adverbios:** Se consideran importantes las oraciones que cuyas palabras indican acciones, las cuales dan significado al contenido
- **Inclusión de entidades nombradas:** Las oraciones que contienen palabras que representan entidades como nombre de alguna organización o lugar son importantes. Por lo tanto, la frecuencia de entidades nombradas indica el nivel de relevancia de dichas palabras, por lo que deberían ser incluidas en el resumen final.
- **Inclusión de números:** Dado que las cifras siempre son cruciales para presentar hechos, esta característica da importancia a las oraciones que tienen ciertas cifras.
- **Frecuencia inversa de oraciones:** Se define como la frecuencia relativa de aparición de un término en una oración. El concepto básico de utilizar esta puntuación es evaluar cada palabra en relación con su distribución en todo el documento.
- **Frases clave:** Es un conjunto de palabras que indican que la oración lleva un mensaje importante en el documento (por ejemplo, "en este reporte", "en resumen", "propósito", "significativamente", "en conclusión") (Roul, 2021; Singh et al., 2021).
- **Palabras clave positivas:** Son aquellas que generalmente son más frecuentes en el documento (Roul, 2021; Singh et al., 2021). Su frecuencia indica la relevancia de palabras y si una oración incluye varias de estas palabras, entonces indica que esa oración es importante.
- **Palabras clave negativas**: A diferencia de las palabras clave positivas, las palabras clave negativas son poco probables que ocurran en el resumen, por lo que la frecuencia de estos términos es también relevante en una oración (Roul, 2021; Singh et al., 2021).
- **Similitud entre oraciones:** La similitud entre oraciones es una propiedad que mide el grado de semejanza y relación dos o más oraciones. Los valores de similitud son representados por valores numéricos regularmente entre 0 y 1, donde los más

cercanos a 1 indican una similitud más alta entre ambas secuencias. Mientras una oración tenga una alta similitud con las demás, se considera importante y apta para el resumen final.

- **Número de oraciones en el documento:** Proporciona un número de oraciones en el documento de manera general, ya que es muy probable que la importancia de una oración varie para los documentos fuente largos y cortos (Mutlu et al., 2019b)
- **Posición de oración relativa al párrafo:** Esta característica viene directamente de la observación de que, al comienzo de cada párrafo, se inicia una nueva discusión y al final de cada párrafo, tenemos un cierre concluyente. Por lo tanto, esta característica indica que las primeras y últimas oraciones de cada párrafo son importantes para el resumen.
- **Similitud de palabras entre párrafos:** Esta característica es similar a similitud entre oraciones. Sin embargo, esta característica calcula la similitud entre párrafos completos, en lugar de oraciones individuales (Fattah & Fattah, 2014).
- **Coocurrencia de información no esencial:** En (Fattah & Fattah, 2014) se considera que algunas palabras son indicadores que información no esencial, y suelen aparece al principio de una oración.
- **Palabras influyentes:** Se puede establecer una lista de palabras para un dominio en particular (Goularte et al., 2019). A partir de esta lista, las oraciones se consideran relevantes si contienen alguna de las palabras pertenecientes a la lista.
- **Estilo de palabras:** Las palabras con énfasis (negrita, cursiva y subrayado) o en mayúsculas pueden considerarse importantes (Goularte et al., 2019).
- **Inclusión de Pronombres:** Los pronombres no se incluyen en las características importantes, a menos que vayan acompañados de un sustantivo correspondiente (Goularte et al., 2019).
- **Análisis del discurso:** La información sobre el nivel del discurso en un texto también se considera una buena característica. Analizando el discurso es posible identificar la estructura del texto (Goularte et al., 2019), y así seleccionar las oraciones relevantes del mismo.

La Figura 3 muestra una gráfica con la frecuencia de implementación (en número de trabajos) de cada característica que se ha propuesto dentro del estado del arte de la GARMD.

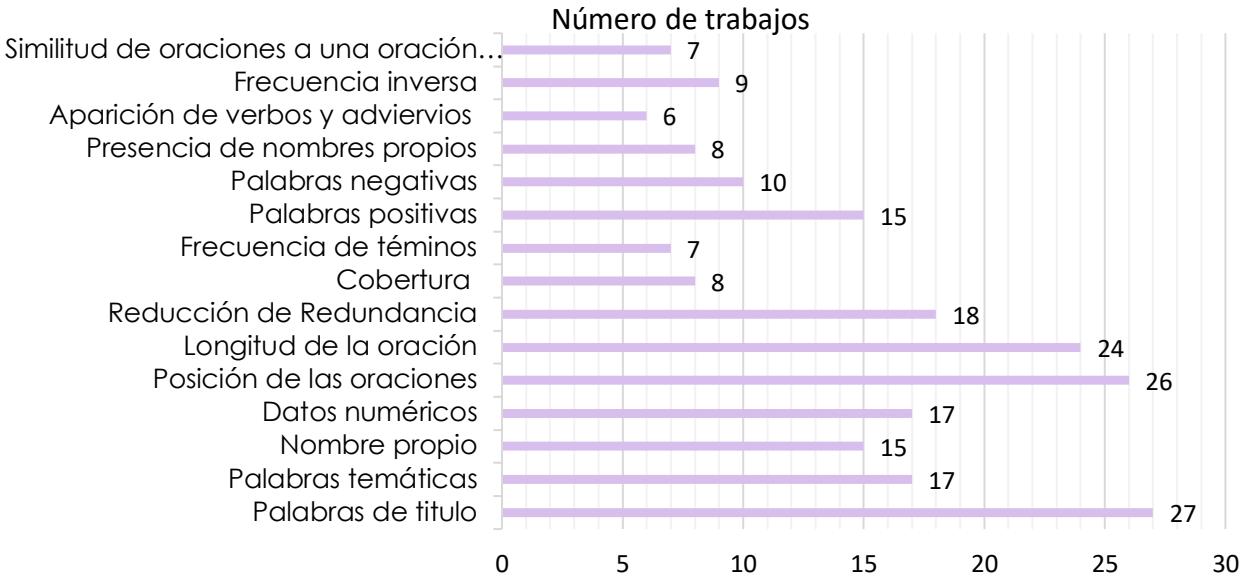


Figura 3. Número de trabajos que abordan cada característica para la GARMD.

De la Figura 3, se puede notar que hay características más utilizadas que otras. Por ejemplo, la posición de oración es más utilizada que la coherencia o cobertura. Esto se debe a que ciertas características tienden a ser más factibles de implementar o el costo computacional puede ser razonable.

Al escribir un resumen los humanos tenemos en cuenta tanto el significado como las relaciones semánticas de las oraciones, así como algunas propiedades estructurales del texto (palabras, oraciones, párrafos y documento). Por este motivo, en los métodos propuestos en el EA, se ha buscado incluir características que consideren los distintos niveles de estructura del lenguaje. Así los niveles léxico y sintáctico, analizan la distribución de características a nivel palabra y oración respectivamente (Fattah & Ren, 2009; Ferreira et al., 2013; Goularte, Nassar, Fileto, & Saggion, 2019; Meena & Gopalani, 2014; Mutlu, Sezer y Akcayol, 2019; 2020; Oliveira et al., 2016; Suanmali, Salim y Binwahlan, 2009; Wan, 2010; Wang, Li, Wang y Zheng, 2017). En el nivel semántico, en cambio, se ha tenido en cuenta el significado de las palabras y oraciones, así como las relaciones semánticas de las mismas (Chen, Liu, Chen, & Wang, 2017; Cheng & Lapata, 2016; Denil,

Demiraj, & De Freitas, 2015; Mohamed & Oussalah, 2019; Narayan, Cohen, & Lapata, 2018; Ren et al., 2018; Yin & Pei, 2015; Zhang, Lapata, Wei, & Zhou, 2018).

El uso de diferentes características y las combinaciones de ellas son factores que influyen en la calidad de un buen resumen (Sanchez-Gomez et al., 2021). Las principales dificultades en este sentido son:

a) Determinar cómo seleccionar las características que deben implementarse en un método (selección de características).
b) Determinar el nivel de relevancia de cada característica.

Dentro del EA, se ha abordado la selección de características de distintas formas. A continuación, se presentan algunas aproximaciones y trabajos previos que se han abordado estos temas.

3.1.1 Selección de características

Debido a que todas las características estudiadas en EA son demasiadas, es necesario definir un criterio de selección. Donde se puede hacer una selección considerando la frecuencia de las características candidatas como criterio de selección (Jo, 2019). Sin embargo, algunas investigaciones han propuesto diferentes enfoques y métodos de selección de características.

En la investigación de (Sanchez-Gomez et al., 2021) se analizó, implementó, comparó y seleccionó características para seleccionar oraciones. En este trabajo, se consideraron tres esquemas se ponderación de términos como TF-IDF (ver Ecuación 1 y Ecuación 2), RTF-SISF (variante de TF-IDF) y TF con longitud de documento, la cual esta última es una ponderación establecida por el framework Okapi BM25. Estas ponderaciones se utilizaron bajo una selección de cinco medidas de similitud para generar resúmenes, las cuales son: similitud Coseno, Jaccard, de Traslape, RRN y NGD, con la finalidad de satisfacer los siguientes requerimientos: mejorar cobertura y reducir redundancia. En este trabajo emplea una selección de características, ya que considera una selección entre:

- Esquemas de ponderación de términos
- Medidas de similitud

Para la evaluación de cada selección de características, se utilizó el corpus DUC02, generando resúmenes de 200 palabras. Sin embargo, existe un sesgo en la selección de colecciones de documentos, pues en su análisis y generación de resúmenes solo fueron consideradas 10 colecciones de 59. Además, no se tomaron en cuenta otras características que pudieran ser relevantes, como semejanza de oraciones con el título, centralidad de la oración, aparición de palabras clave, entre otras. No obstante, es un trabajo que considera la selección de características para modelar la caracterización y selección de oraciones.

Por otro lado, en (Mutlu et al., 2019b) se propuso un método para la GARMD que analiza las siguientes características tanto de forma individual, como sus diferentes combinaciones:

- Frecuencia de términos (TF)
- Similitud con el título
- Frecuencia de la oración inversa (IDF)
- Posición de la oración
- Longitud de la oración
- Similitud entre oraciones
- Inclusión de nombres propios

En esta investigación, se examinaron varias combinaciones de estas características para identificar la manera más eficiente de extraer oraciones relevantes en DUC02. Después de la ponderación de oraciones, se llevó a cabo la selección las mismas, la cual se llevó a través de un perceptrón multicapa poco profundo y dos sistemas de inferencia difusos para extraer oraciones, considerando que la longitud de los resúmenes es de 200 y 400 palabras. Los resultados de esta investigación sugieren que el método propuesto supera a los propuestos en el estado del arte. No obstante, en los experimentos realizados no se hizo un análisis adecuado y conclusivo en determinar las mejores combinaciones de características para diferentes longitudes de resúmenes, ya que tanto los documentos fuente, como los resúmenes de referencia fueron unidos.

Otro trabajo que ha abordado la selección de características fue en (Fattah & Fattah, 2014), donde se investigó el efecto de utilizar ocho características de oraciones, las cuales se enlistan a continuación:

1. Similitud de palabras entre oraciones
2. Similitud de palabras entre párrafos
3. Estilo de palabras
4. Palabras clave
5. TF-IDF
6. Coocurrencia de palabras entre la oración y el título
7. Posición de la oración
8. Coocurrencia de información no esencial

Después de determinar el efecto de estas características de manera individual, luego se emplearon en combinación para construir resúmenes por medio de un modelo de máxima entropía, clasificador Naive Bayes y máquina de soporte vectorial o SVM. Cada modelo o clasificador es entrenado con una muestra de datos del corpus DUC01. Posterior al entrenamiento, el desempeño o aprendizaje de cada modelo o clasificador fue probado en resúmenes de múltiples documentos de 100 palabras, utilizando el corpus DUC02. Los resultados que se obtuvieron indican que al seleccionar las características 1, 2, 5 y 6 generan resultados de evaluación razonables. No obstante, la selección propuesta depende de modelos de aprendizaje supervisado, siendo una dificultad ante la generación resúmenes en otros dominios o lenguajes.

3.1.2 Relevancia de características

Además de la selección de las características, en el modelado también es necesario otorgar relevancia a cada una de ellas. En el EA se han propuesto diferentes formas de obtener esta relevancia: A través de la puntuación de los documentos de entrada o asignando un coeficiente de relevancia para cada característica.

3.1.2.1 Relevancia basada en la puntuación de las oraciones de los documentos de entrada

En este enfoque, a partir de los documentos de entrada se construye la relevancia de las características, es decir, en función del texto que se proporciona se generarán

ponderaciones para cada característica considerada.

En (Luhn, 1958) se introdujo por primera vez un método simple, pero eficaz, para generar resúmenes basados en la frecuencia de términos, pues se demostró que la frecuencia podría servir como criterio para detectar oraciones más relevantes. En su método se identificaron las raíces de las palabras, finalmente, las oraciones se ordenaron por su puntuación. Mientras que en (Baxendale, 1958) se consideró la posición de las oraciones, así como la frecuencia de las palabras para resumir, más tarde en (Edmundson, 1969), se introdujeron características como: Palabras emblemáticas, similitud con el título.

3.1.2.2 Relevancia basada en coeficientes calculados mediante optimización

En (Jain et al., 2022), se utilizó un método de optimización para la selección de ocho características que ponderan la importancia de cada oración, entre las que están: posición de la oración en el párrafo, frecuencia de términos, longitud de la oración, frecuencia inversa de documentos, oraciones que incluyen números, similitud con las oraciones. En esta investigación, el tamaño de los cromosomas ($C = 8$) representa el número total de características. Cada cromosoma es una combinación de 8 valores de coeficientes de relevancia calculados en un rango de 0 a 1. Posteriormente, cada coeficiente de relevancia participa en la generación de resúmenes. La función de aptitud es una medida unitaria que determina un cromosoma que conduce a la mejor solución entre un conjunto de cromosomas. La función de aptitud fue la siguiente:

$$F(s) = \sum_{j=1}^{C} f_{j\,(s)}$$

Ecuación 13. Función de aptitud (Jain et al., 2022)

Donde: f_j es el coeficiente de relevancia para cada j-ésima característica, (s) es el puntaje de la oración y C es el número total de características.

Mientras que en (Verma & Om, 2019) se generaron combinaciones de coeficientes de relevancia de para extraer oraciones. Con 20 características entre las que se encuentran: similitud agregada, *Bushy path*, palabras clave, relación léxica, entidades nombradas, oraciones nominales y verbales, inclusión de datos numéricos, relaciones

abiertas, inclusión de nombres propios, centralidad de la oración, longitud de la oración, posición de la oración, inclusión de palabras del título, significado de la oración, frecuencia de palabras, inclusión de palabras mayúsculas. Los coeficientes de relevancia iniciales de las características fueron generados de forma aleatoria, en el rango de 0 a 1. Posteriormente, estos coeficientes se actualizaron a través del AG. Las oraciones se evalúan por medio de los coeficientes de relevancia por la puntuación de la oración a través de la siguiente función de aptitud:

$$X_q = \{X_{q1} + X_{q2} + \cdots X_{q20}\}$$

Ecuación 14. Función de Aptitud (Verma & Om, 2019)

Donde X es el coeficiente de relevancia de cada característica y q_i es el valor de puntuación de la oración.

En (Fattah, 2016) fue empleado un AG para obtener una combinación adecuada de coeficientes de relevancia de 10 características, entre las que se encuentran las siguientes: posición de las oraciones, semejanza con el título, inclusión de entidades que incluyan nombres, longitud de la oración entre otras. Primero se investigó el efecto de cada característica para generar resúmenes. Posteriormente, consideraron todas las características mencionadas anteriormente para entrenar a un AG y un algoritmo de regresión matemática para obtener una combinación adecuada de características y coeficientes de relevancia.

Para evaluar una oración (s) se utilizó una función de puntuación ponderada para integrar las 10 características como se muestra en la siguiente ecuación:

$$Puntuación\ Oración\ (s) = w_1 C_1\ (s) + w_2 C_2\ (s) + w_3 C_3\ (s) + w_4 C_4\ (s)$$
$$+ w_5 C_5\ (s) + w_6 C_6 (s) + w_7 C_7\ (s) + w_8 C_8\ (s) + w_9 C_9\ (s) + w_{10} C_{10}\ (s)$$

Ecuación 15. Puntuación de la oración (Fattah, 2016)

Donde: w_i son los coeficientes de relevancia calculados por el AG, y C_i la puntuación de cada característica en la oración (s).

3.1.2.3 Relevancia basada en coeficientes calculados mediante aprendizaje automático

En (Jo, 2019) los coeficientes de relevancia fueron calculados proporcionalmente, a través de cálculo de similitudes entre vectores (similitud coseno y distancia euclidiana). Se consideraron conjuntos de datos con ejemplos de entrenamiento. En cuanto a la selección de oraciones, en este trabajo se aborda como un problema de clasificación binaria, mediante el algoritmo KNN, donde cada oración se clasifica como parte del resumen o no.

Por otro lado, en (Mahalleh & Gharehchopogh, 2022) se seleccionaron oraciones a través de 5 características semánticas y estadísticas (frecuencia de palabras, similitud con el título, posición de las oraciones, centralidad y número de palabras clave en la oración). A cada de una estas características se les asignó un coeficiente de relevancia que es superior a otro. Este coeficiente fue obtenido mediante aprendizaje automático, donde a través de conjuntos de datos externos que incluyen textos y resúmenes de noticias de la BBC, el método propuesto aprendió el coeficiente de relevancia de cada característica. Una vez determinados los coeficientes de relevancia, se calculó la puntuación de cada oración por medio de la siguiente ecuación:

$$Puntuación\ (oración) = \sum_{i=1}^{5} w_i * Puntuación_{f_i}(oración)$$

Ecuación 16. Puntuación de la oración (Mahalleh & Gharehchopogh, 2022)

Donde w_i es el coeficiente de relevancia de la característica i, f_i es la puntuación de la característica en la oración en cuestión. En otras palabras, cada coeficiente indica el nivel de contribución que tiene cada característica. Después de calcular todas las puntuaciones de las oraciones, se realizó un ranqueo para determinar qué oraciones formarían parte del resumen final.

En (Li et al., 2020), se presenta un modelo de clasificación basado en grafos no supervisado, en el que por la similitud entre palabras y oraciones se puede estimar la distancia entre vectores de incrustaciones de palabras. Los coeficientes de relevancia de las características se obtuvieron de la frecuencia de palabras que reflejan mejor contenido de un documento. Una oración es destacada si es recomendada por otras

oraciones y también debe ser votada por palabras destacadas, debido a que expresa un significado similar y una palabra es importante si aparece en muchas oraciones destacadas. Con el objetivo de ranquear las oraciones y palabras, se definió una matriz de similitud palabras por oración que fue calculada mediante la siguiente ecuación:

$$P = \alpha + \beta + \gamma$$

Ecuación 17. Suma ponderada (Li et al., 2020)

Donde la puntuación de los vectores es la suma ponderada de α, β, γ tienen valores entre 0 y 1.

En este mismo sentido, en (Binwahlan et al., 2009) el conjunto de datos se dividió en 2 secciones (entrenamiento y prueba) para calcular los coeficientes de relevancia de las características. Las características estudiadas fueron cinco, las cuales son: Centralidad de la oración, similitud con el título, palabras clave, similitud con la primera oración y frecuencia de palabras. Una vez que fueron obtenidos los coeficientes, fueron multiplicados por la importancia de cada oración a través de la siguiente ecuación:

$$Puntuación\ (oración_i) = \sum_{j=1}^{5} Cf_j * puntuación_{oración}$$

Ecuación 18. Fórmula para determinar la importancia de la oración

Donde: Cf_j es el coeficiente de relevancia de cada característica. En consecuencia, las oraciones seleccionadas para el resumen final se ordenaron de forma descendente.

3.1.1.4 Relevancia basada en coeficientes manuales

En (Mendoza, 2015) se utilizó un conjunto de características independientes del dominio y lenguaje, para determinar la calidad un resumen. Entre las que se encuentran: Posición de las oraciones (c1), similitud con el título (c2), longitud de las oraciones (c3), cohesión entre las oraciones (c4) y cobertura (c5). Estas características fueron maximizadas con la función objetivo que fue optimizada por un algoritmo memético:

$$Max\left(f(X_k)\right) = \alpha(c1) + \beta(c2) + \gamma(c3) + \delta(c4) + \varepsilon(c5)$$

Ecuación 19. Función objetivo (Mendoza, 2015)

Donde: $\alpha, \beta, \gamma, \delta, \varepsilon$ son coeficientes que permiten dar un peso ponderado a cada característica. Los coeficientes de relevancia son valores de 0 a 1.

De manera semejante, en (Qaroush et al, 2021) se propuso un método, con la combinación de características semánticas y estadísticas que incluyen oraciones clave, longitud de la oración clave, inclusión de nombres propios, posición de la oración, similitud con el título, centralidad de la oración, inclusión de números. En la etapa de selección de oraciones fueron evaluadas las oraciones con una suma lineal definida como:

$$Puntuación\ de\ la\ oración = \sum_{i=1} w_i * s_i$$

Ecuación 20. Suma lineal de características (Qaroush et al, 2021)

Donde w_i representa el coeficiente de relevancia de las características definido y s_i la puntuación de la característica. La sumatoria de esta puntuación refleja la contribución de la característica en la puntuación total que es igual a 1. Después de calcular la puntuación total de las oraciones, estas se ordenaron en forma descendente según su puntuación total.

En el trabajo de Vázquez et al. (2018), se propuso realizar una optimización en la selección de oraciones mediante el Algoritmo Genético (AG) para la GARDI. Para diferenciar la importancia de cada selección de oraciones, se combinaron cuatro características, los cuales se enlistan a continuación: Cobertura (α), Posición de la oración (β), longitud de la oración (γ) y similitud con el título (δ).

La forma en que se combinaron estas características depende del grado de relevancia de cada una de ellas. Es decir, a cada una se le proporciona un coeficiente con valor decimal entre 0 y 1, donde 0 representa el valor mínimo de relevancia y 1 representa el máximo valor posible. De manera formal, esta combinación se representa a través de la siguiente ecuación:

$$Puntaje(S) = w_1\alpha + w_2\beta + w_3\gamma + w_4\delta$$

Ecuación 21. Combinación de características propuesta en Vázquez (2018).

Donde S representa el contenido del resumen generado; w_1, w_2, w_3 y w_4 son los coeficientes de relevancia que fueron asignados a las características α, β, γ y δ, respectivamente. Los resultados obtenidos muestran mejoras en la selección de oraciones. Sin embargo, la forma en la que obtuvieron estos coeficientes fue de manera manual, asumiendo que estos valores son los adecuados para mejorar la selección de oraciones. Por lo tanto, esto representa optar por criterios subjetivos, que a su vez limita obtener una mejora en la selección de oraciones.

3.1.3 Selección de oraciones

El problema crucial al crear resúmenes cercanos al humano es seleccionar las oraciones más relevantes del texto, de manera que conserven los temas principales y se evite la inclusión de oraciones redundantes (Hosseinabadi et al., 2022). En esta etapa se han incluido diversas técnicas entre las que se encuentran:

3.1.3.1 Árboles de decisión

En (Chuang & Yang, 2000) introdujeron otro algoritmo de extracción de oraciones, que se basa en el árbol de decisión que representa un documento de texto. Los métodos basados en árboles de decisión funcionan bien en un dominio específico (Andhale & Bewoor, 2017; Sabuna & Setyohadi, 2018). El árbol de decisión es uno de los métodos de clasificación que utiliza una representación de estructura de árbol, donde cada uno de los nodos del árbol representa los atributos. Cada rama es la división de resultados probados y cada nodo de hoja representa ciertos grupos de clases. En la GAR el proceso comienza con la clasificación de los datos aleatorios para que sean reglas de decisión. Generalmente, el árbol de decisiones utiliza una estrategia de búsqueda de arriba hacia abajo. El árbol se construye dividiendo los datos de forma recursiva para que cada parte de los datos provenga de la misma clase. El algoritmo es uno de los métodos para crear árboles de decisiones basados en el entrenamiento (Aries et al., 2019; Nasar et al., 2019; Sabuna & Setyohadi, 2018)

3.1.3.2 Cadenas léxicas

(Barzilay & Elhadad, 1997) introdujo un método para resumir textos basado en cadenas léxicas. Usó distribución de palabras y conexiones léxicas para generar una estructura léxica de presentación del texto. Las cadenas léxicas consideran relaciones entre palabras y la cadena léxica se crea tomando una nueva frase y encontrando una cadena relacionada según la cohesión léxica. Estas frases clave son los elementos clave del resumen (Arya et al., 2023; Paul & Salim, 2023). En general, esta técnica intenta identificar los conceptos más importantes en el documento de entrada con la ayuda de cadenas léxicas y luego extraer las oraciones que contienen esos conceptos (Divya & Sripriya, 2022; Waseemullah et al., 2022).

3.1.3.3 Clustering

Por lo general, el documento se redacta de forma coordinada donde se clasifican varios temas uno tras otro. Normalmente, estos documentos se dividen en segmentos de forma explícita o implícita. Debido a que en la GAR se deben abordar todos los temas que aparecen en los documentos. Para producir un buen resumen significativo, es necesario agrupar el texto. Si los documentos fuente abordan diversos temas es útil el uso de esta técnica (Alguliyev et al., 2019; Belwal et al., 2023; Cajueiro et al., 2023; Divya & Sripriya, 2022; Gupta et al., 2023). La agrupación es un paso común y también crucial en el resumen de texto mediante el cual se puede extraer valiosa información relacional oculta (Belwal et al., 2023). En (P. Verma & Om, 2019), las oraciones se agrupan según la distancia semántica entre ellas. A continuación, en cada grupo, se calcula la similitud acumulativa de oraciones basada en el método de combinación de múltiples características.

3.1.3.4 LSA (Análisis Semántico Latente)

Este mecanismo ha surgido en el tratamiento de lenguajes naturales para analizar las relaciones entre los documentos y la terminología contenida en ellos. LSA se representa con una matriz que contiene recuentos de palabras por párrafos (las filas son las palabras únicas y las columnas representan cada párrafo). Es una técnica basada en la descomposición de valores singulares (SVD), que se utiliza para minimizar el número

de filas, manteniendo la estructura de similitud entre columnas. Luego las palabras se comparan tomando el coseno del ángulo entre los 2 vectores (formado por 2 filas cualesquiera). Los valores más cercanos a 1 representan palabras muy similares, mientras que los valores más cercanos a 0 representan palabras muy diferentes. En general, para el resumen automático de texto existen 3 pasos: Creación de la matriz de entrada, descomposición de valores singulares y selección de oraciones (Belwal et al., 2023; Rajalakshmi et al., 2023; Ramani et al., 2023; Reddy & Guha, 2023).

3.1.3.5 Redes Neuronales

Estudios de investigación recientes han incorporado el aprendizaje profundo, además de incrustaciones de palabras y mecanismos de aprendizaje de para abordar el problema de la GAR. Las redes neuronales han demostrado que tienen buen rendimiento. Entre las redes neuronales que se han aplicado se encuentran: red neuronal recurrente (Abdi et al., 2021; Sharma et al., 2022; S. Verma & Nidhi, 2018), red neuronal convolucional (Ghadimi & Beigy, 2023; Joshi et al., 2023), redes neuronales basadas en transformadores (Glazkova & Morozov, 2023), redes neuronales basadas en codificador automático variacional (Xiong et al., 2023), red neuronal basada en grafos (Jalil et al., 2023).

3.1.3.6 Optimización

Los métodos de resumen también pueden formularse matemáticamente como un enfoque de optimización que puede ser de un solo objetivo o de múltiples objetivos. En los enfoques de objetivo único, se optimiza una función objetiva única que incluye todos los criterios ponderados. Por otro lado, los enfoques multiobjetivo pueden optimizar simultáneamente muchas funciones objetivo sin ponderación (Abuobieda et al., 2012; Binwahlan et al., 2009; Jain et al., 2022; Mendoza, 2015; Mojrian & Mirroshandel, 2021; Mosa et al., 2019; Sanchez-Gomez et al., 2022). Entre los principales algoritmos que se han empleado en el EA se encuentran:

- Colonia de Abejas Multiobjetivo
- Algoritmo Codicioso
- Algoritmo de optimización evolutivo

- Algoritmo memético
- Algoritmo de Evolución Diferencial
- Algoritmo genético

En particular, el AG ha obtenido buenos resultados en las aproximaciones de la GARMD. Simula la evolución de los organismos naturales en una computadora para resolver problemas de optimización en el campo real que genera aleatoriamente diferentes tipos de soluciones de problemas y selecciona soluciones más favorables de acuerdo con el principio de supervivencia donde el más apto iterará y optimizará aún más las soluciones a través de la herencia y la variación que es similar a la evolución biológica en la naturaleza (Abualigah et al., 2020; Han & Xiao, 2022).

3.2 Resumen del capítulo

En este capítulo se describieron los aspectos que considera el modelado de características. Dichos aspectos fueron los siguientes:

Selección de características

Determina un subconjunto de características que contribuyan a generar resúmenes que incluyan oraciones relevantes, que cubran los temas que se abordan en los documentos fuente y no transmitan la misma información. En el EA, la selección de características se ha realizado en trabajos previos estableciendo una selección por frecuencias o por métodos más sofisticados y costosos como los expuestos en la sección 3.1.1.

Relevancia de las características

En el EA la mayoría de las investigaciones han buscado incorporar un coeficiente de relevancia. Porque tratar todas las características del texto con el mismo nivel de importancia puede considerarse el factor principal que provoca resúmenes de baja calidad (Mosa et al., 2019). Los coeficientes de relevancia han sido calculados mediante optimización, aprendizaje automático, *clustering*, y coeficientes manuales que tienden a ser subjetivos.

Selección de oraciones

Es la etapa crucial para generar el resumen. Es aquí en donde las características seleccionadas, con sus coeficientes de relevancia asignados establecen qué oraciones son las representativas del documento. En el estado del arte esta etapa se ha abordado por diversas técnicas tales como: árboles de decisión, cadenas léxicas, clustering, análisis semántico latente, redes neuronales y optimización.

No obstante, existen algunas limitaciones. Por ejemplo, en *clustering* a pesar de que es un método simple e intuitivo, los elementos están limitados a ser asignados a un solo grupo (Alguliyev et al., 2019; Belwal et al., 2023; Cajueiro et al., 2023; Divya & Sripriya, 2022; Gupta et al., 2023). Por otro lado, en los métodos basados en grafos, es posible modelar documentos, sin embargo, su construcción y almacenamiento es complejo. Además de que no reflejan el significado de las palabras u oraciones, motivo por el que diversas palabras u oraciones que se refieren al mismo tema se representan como dos nodos diferentes (El-Kassas et al., 2021; Z. Li et al., 2020).

En cuanto los métodos basados en aprendizaje profundo tienen un buen rendimiento, pero se enfrentan al desafío de encontrar conjuntos de datos adecuados que normalmente, constan de una gran cantidad de datos de entrenamiento (Abdi et al., 2021; El-Kassas et al., 2021; Ghadimi & Beigy, 2023; Sharma et al., 2022; S. Verma & Nidhi, 2018; Xiong et al., 2023). Con respecto a los métodos basados en análisis semántico latente, el resumen generado depende de la calidad de la representación semántica del texto de entrada (Belwal et al., 2023; Rajalakshmi et al., 2023; Ramani et al., 2023; Reddy & Guha, 2023). En árboles de decisión no es posible identificar las relaciones entre oraciones sin descubrir las frases compartidas entre estas oraciones (Aries et al., 2019; El-Kassas et al., 2021; Nasar et al., 2019; Sabuna & Setyohadi, 2018).
Por lo anterior expuesto, es necesario modelar las características obteniendo un coeficiente de ponderación a través de un método que mejore la selección de oraciones.

CAPÍTULO 4
Método Propuesto

A partir del problema planteado en el Capítulo 1, donde se desconoce qué tan útil puede ser el modelado de características híbridas obtenidas de los resúmenes de referencia que pueda mejorar la selección de oraciones para la GARMD extractivos, se estableció la siguiente hipótesis: si los resúmenes de referencia son el modelo objetivo de la tarea de la GARMD extractivos, entonces la obtención de sus características híbridas permitirá mejorar la selección de oraciones en dicha tarea. Para comprobar esta hipótesis, se propone una metodología que consta de las siguientes etapas:

- Modelado de características.
- Preprocesamiento, etiquetado y vectorización.
- Selección de oraciones.

Posteriormente, cada una de las etapas es descrita.

4.1 Modelado y cálculo de coeficientes de relevancia de las características

La finalidad de esta etapa fue obtener los coeficientes de relevancia de cada característica a partir de los resúmenes de referencia escritos por humanos, que incluyen los conjuntos de datos. Estos coeficientes de relevancia afectan la forma en que se evaluaron las oraciones del texto fuente, para ser seleccionadas e incluidas en el resumen final. La figura 5, se muestra el proceso de esta etapa y posteriormente su descripción.

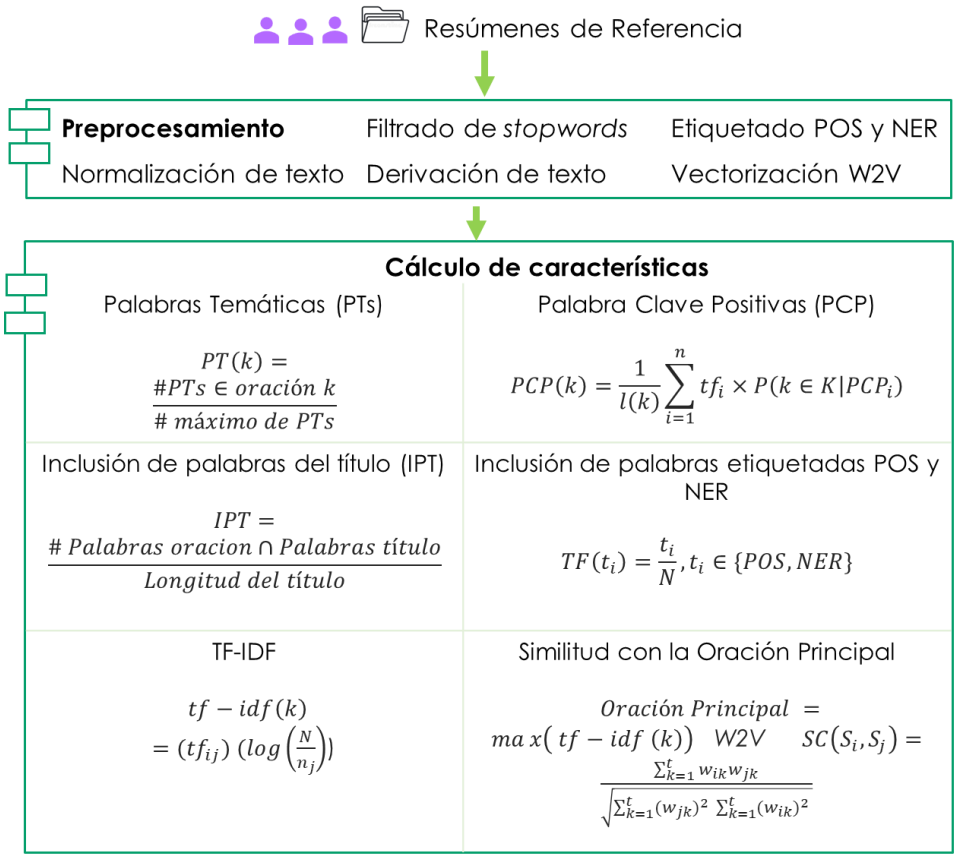


Figura 4. Extracción de características.

Entrada: La entrada en esta etapa son los resúmenes de referencia escritos por humanos (documentos fuente).

Preprocesamiento: Los documentos fuente fueron preprocesados normalizando y derivando el texto, haciendo filtrado de *stopwords,* posteriormente el texto fue etiquetado asignando categorías gramaticales (Etiquetado POS), así como etiquetas

de reconocimiento de entidades nombradas (NER). Además, el texto fue vectorizado con el modelo de incrustación de palabras *Word2vec* para codificar el significado de las palabras y construir conceptos lingüísticos de las oraciones.

Cálculo de características: De los documentos fuente preprocesados se extrajeron un conjunto de características híbridas que consideran los niveles de estructura textual y del lenguaje, así como los requerimientos de la GARMD y se describen en la siguiente tabla:

Tabla 1.Características calculadas

Palabras temáticas: Está característica es relacionada con palabras de dominio específico que aparecen con frecuencia en un documento. En el método propuesto debido a que los documentos suelen ser muy largos se ha considerado al 7% de las palabras más frecuentes como palabras temáticas.			
Contribución**:** Relevancia de la oración y cobertura.	Nivel Textual: Palabra.	Nivel de estructura del lenguaje: Léxico	Categoría: Estadística.
Palabras Clave positivas: Dado que las palabras son los elementos básicos de una oración, cuantas más palabras clave con contenido tenga una oración, más importante será. Por lo tanto, las palabras clave positivas se definen como el 7% de las palabras que se incluyen con frecuencia en el resumen.			
Contribución: Relevancia de la oración y cobertura	Nivel Textual: Palabra	Nivel de estructura del lenguaje: Léxico	Categoría: Estadística
Inclusión de palabras del título: La oración recibe una puntuación alta si contiene palabras que aparecen en el título del documento.			
Contribución: Relevancia de la oración	Nivel Textual: Palabra	Nivel de estructura del lenguaje: Léxico	Categoría: Estadística
Inclusión de palabras etiquetadas (POS, NER): La frecuencia de etiquetas POS o NER puede indicar la relevancia de palabras que conforman una oración. En el método propuesto fueron calculadas 56 diferentes categorías. Sin embargo, debido a que algunas tuvieron una			

ponderación muy baja solo fueron consideradas 14 de ellas: CC, CD, DT, IN, JJ, NN, NNS, VB, VBD, VBN, PERSON, ORG, GPE y DATE, (Ver anexo 2).			
Contribución: Relevancia de la oración y cobertura.	Nivel Textual: Palabra.	Nivel de estructura del lenguaje: Sintáctico	Categoría: Lingüística
TF-IDF: Reconoce palabras esenciales. El término frecuencia (TF) mide la cantidad de veces que aparece una palabra en el documento, mientras que la frecuencia inversa del documento (IDF) procesa la cantidad de oraciones en los que aparece la palabra. En el momento en que la palabra es más frecuente en la oración, pero menos frecuente en todo el documento, el valor TF-IDF es mayor.			
Contribución: Cobertura, reducción de redundancia	Nivel Textual: palabra, oración, documento	Nivel de estructura del lenguaje: Léxico	categoría: Estadística
Similitud con la oración principal: Se define como la similitud entre una oración y otras oraciones en el documento, el empleo de centralidad aumenta la diversidad. En el método propuesto, la oración principal fue la que obtuvo la puntuación más alta en el cálculo de TF-IDF. Una vez identificada la oración principal, se calculó la similitud con otras oraciones con similitud coseno a través de la vectorización Word2vec.			
Contribución: Reducción de redundancia y cobertura.	Nivel Textual: Oración y documento	Nivel de estructura del lenguaje: Semántico y Léxico.	Categoría: Lingüística y estadística

A partir del cálculo de las características de la tabla anterior se realizaron los siguientes pasos (que se ejemplifican en la figura 5).

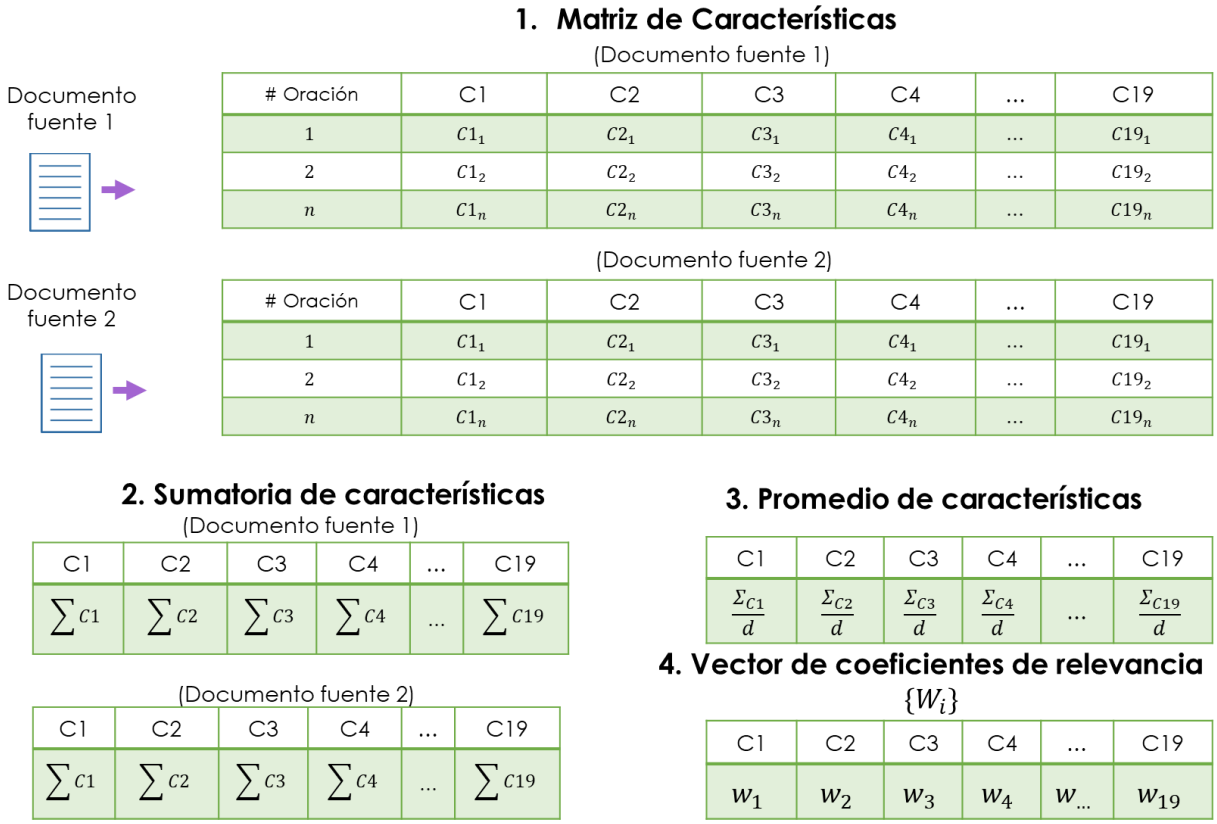


Figura 5. Ponderación de los pesos de las características.

1. Para cada documento fuente, se generó una matriz de características. Las columnas representan el cálculo de cada característica (en la figura 5 representada por C_i), y las filas a las oraciones del documento fuente.
2. Al terminar el cálculo de las características de todas las oraciones del documento fuente se obtuvo una sumatoria de cada característica $\sum C_i$.
3. Se obtuvo el promedio de características de los documentos fuente $\frac{\sum C_i}{d}$ (donde d es el número de documentos fuente), para tener un valor representativo de cada característica.
4. Considerando los promedios del paso anterior, se calcularon los coeficientes de relevancia de cada característica mediante la probabilidad bayesiana (ecuación 22) de cada valor. Una ventaja importante de esta concepción de probabilidad es que admite la asignación de probabilidades a sucesos únicos y permite calcular la probabilidad de un evento, a partir de valores conocidos de

otras probabilidades relacionadas al evento (Tubau Sala & Alonso Cánovas, 2002).

$$Coeficiente\ de\ relevancia = \frac{\left(\frac{\sum C_i}{d} * 1\right)}{\sum \frac{\sum C_i}{d}} = 1$$

Ecuación 22. Cálculo de coeficientes de relevancia

Donde: $\frac{\sum C_i}{d}$ es el promedio de cada característica.

En función del cálculo anterior, se generó el vector de coeficientes de relevancia. Cada coeficiente obtuvo un valor entre 0 y 1.

Salida: Un vector de coeficientes, denominado $\{w_i\}$.

4.2 Concatenación de documentos y preprocesamiento

La figura 6 ejemplifica la segunda etapa del método propuesto y posteriormente, se describen cada una de las partes de esta etapa.

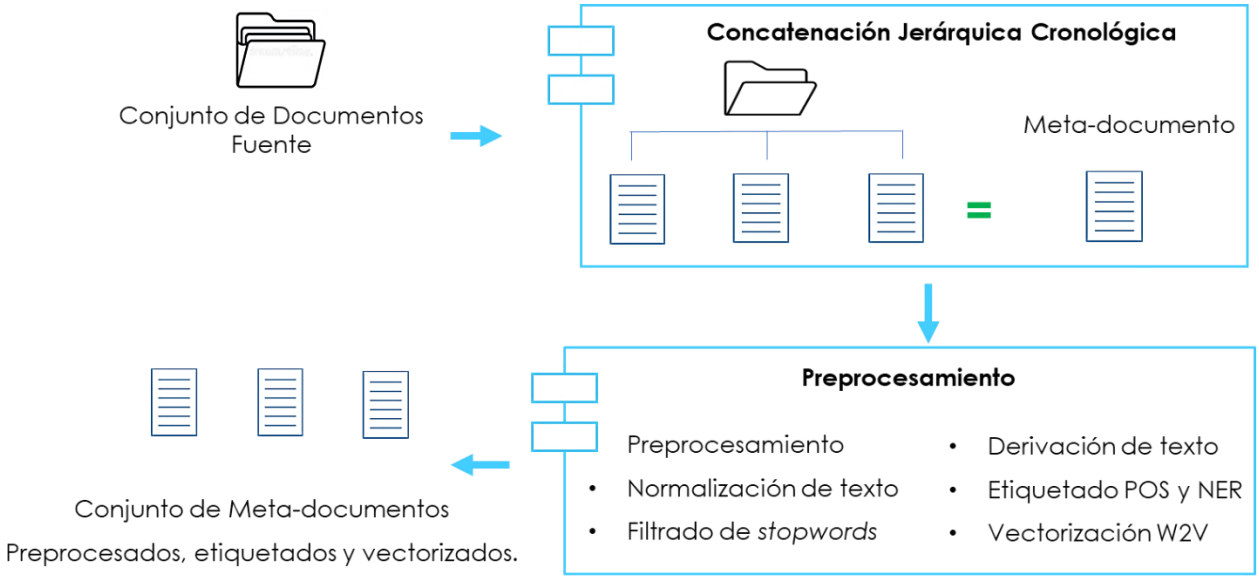


Figura 6 Concatenación y preprocesamiento

Entrada: Un conjunto de documentos para ser resumidos.

Concatenación de documentos: En la GARDM la concatenación implica combinar los documentos de la colección de manera útil para presentar la información más relevante y diversa, así como considerar el orden en los documentos según la ocurrencia de eventos en el tiempo. Por lo que los documentos de la colección fueron concatenados de forma jerárquica para crear un meta-documento, considerando su cronología, desde las noticias más antiguas, hasta la más reciente.

Preprocesamiento: Una vez que los documentos fueron concatenados se aplicó preprocesamiento. Se normalizó el texto considerando lematización, se filtraron de *stopwords.* Además de que se realizó etiquetado POS y NER. Finalmente, las oraciones fueron vectorizadas con *Word2vec* para codificar el significado de las palabras y construir conceptos lingüísticos.

Salida: La salida de esta etapa es un conjunto de meta-documentos preprocesados, etiquetados y vectorizados.

4.3 Optimización de selección de oraciones

En esta etapa se consideró aplicar AG para abordar la selección de oraciones como un problema de optimización combinatoria, por medio de los distintos operadores de selección, cruza y mutación. Para realizar una búsqueda que evalúa una función de aptitud para explorar y presentar una solución óptima para un problema. La figura 7 muestra esta etapa.

Entrada: La entrada para esta etapa fue la salida de la etapa descrita en la sección 4.2, un conjunto de documentos preprocesados, etiquetados y vectorizados.

AG: A partir de la entrada, se generó a la población inicial aleatoria, con codificación binaria, en donde un gen representa a una oración y un individuo representa un resumen candidato.

Consecutivamente, a partir de la población se generan resúmenes candidatos para posteriormente calcular sus características.

Una vez que fueron calculadas las características de cada resumen candidato se siguió el proceso que se muestra en la figura 8 y se describe a continuación:

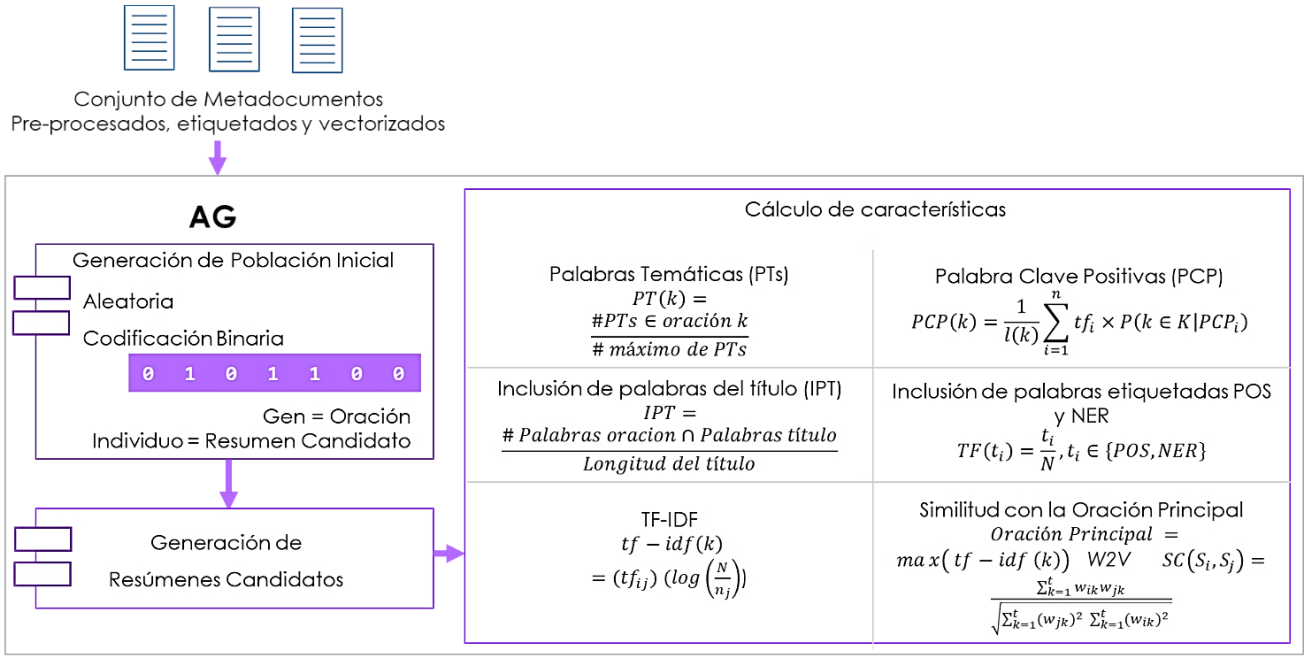


Figura 7. Extracción de características y selección de oraciones.

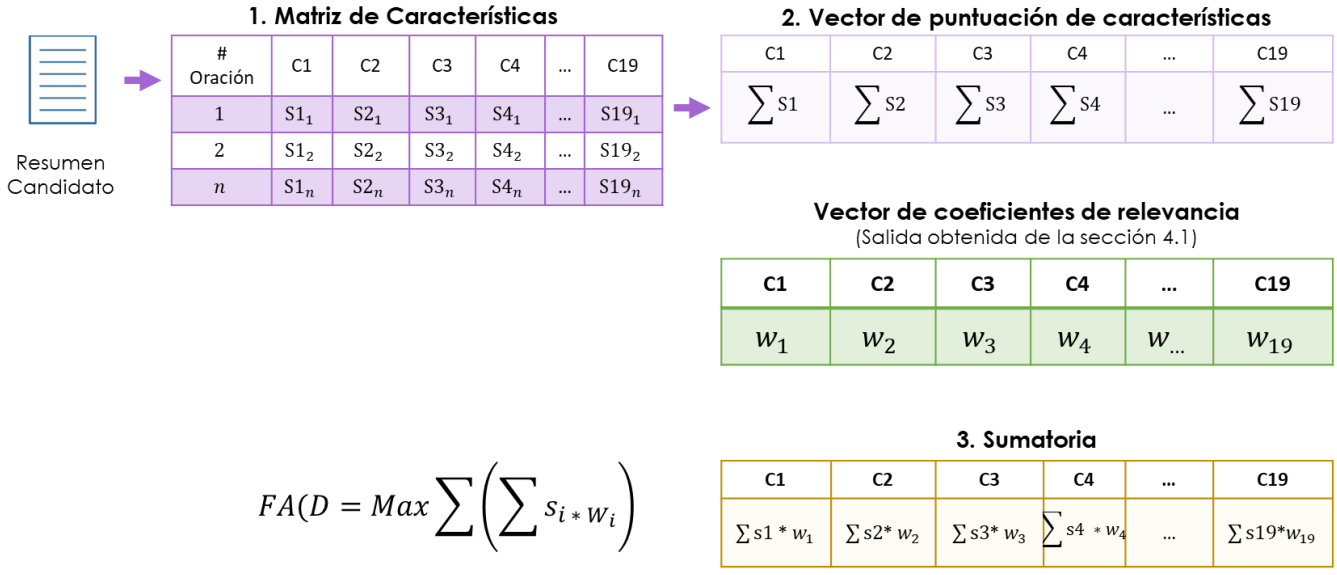


Figura 8. Evaluación de resúmenes candidatos

1- Con base en las características cálculadas para cada resumen candidato, se generó una matriz de características. Las columnas representan los valores S_i de puntuación de cada característica y las filas a las oraciones que constituyen al resumen candidato.
2- Partiendo de las puntuaciones S_i fue generado el vector de puntuación de características a través del cálculo de la sumaria de cada característica $\sum S_i$.
3- Posteriormente, se evaluó la aptitud del resumen candidato por medio de la siguiente ecuación.

$$FA(D = Max \sum \left(\sum s_{i*W_i} \right)$$

Ecuación 23. Función de Aptitud

Donde: la ponderación del documento candidato D_i, es la maximización de la suma lineal de la sumatoria de las puntuaciones obtenidas $\sum S_i$ por los coeficientes de relevancia $\{W_i\}$(vector de coeficientes de relevancia, obtenido en la etapa 4.1).

Posteriormente, se generaron nuevos resúmenes candidatos a través de los operadores de selección, cruza y mutación. El criterio de parada que se estableció fue el número de generaciones.

Salida: Resumen de un conjunto de documentos.

4.4 Resumen del capítulo

En este capítulo se describió la metodología propuesta. Además, se dio cumplimiento a los objetivos específicos:

- Analizar los conjuntos de datos (sección 4.1).
- Seleccionar un enfoque de concatenación de documentos de entrada (Sección 4.2).
- Analizar las características que se han propuesto en el estado del arte para la GARMD (sección 4.1).
- Modelar las características de los documentos de referencia escritos por humanos (sección 4.1 y 4.3).

- Con base en el modelo de características obtenido, optimizar la selección de oraciones a través del AG (sección 4.3).

CAPÍTULO 5
Experimentos y resultados

En este capítulo se describe la experimentación realizada del método propuesto. Con base en los resultados obtenidos del método propuesto, se busca comprobar la siguiente hipótesis: si se obtiene un modelado de características híbridas a partir de los resúmenes de referencia escritos por humanos, entonces se mejorará selección de oraciones para la GARMD extractivos.

Primero, se describen los conjuntos de datos utilizados para probar el desempeño del método propuesto. Posteriormente, se proporcionan descripciones a los métodos del Estado del Arte (EA) y heurísticas, los cuales se utilizaron como puntos de comparación y referencia del método propuesto. Finalmente, se muestran los parámetros utilizados del método propuesto y sus respectivos resultados en cada colección de documentos y longitud del resumen.

5.1 Conjuntos de datos

Para evaluar el desempeño del método propuesto, se consideraron dos tareas: GARMD Genéricos y de Actualización. Ambas tareas fueron consideradas con el objetivo de tener una perspectiva más amplia del método propuesto, así como su flexibilidad frente a diferentes condiciones en el dominio de noticias. En la sección 5.1.1 se muestra una breve descripción del conjunto de datos DUC01. Por otro lado, en la sección 5.1.2 se proporciona la descripción de TAC08.

5.1.1 DUC01

Para evaluar empíricamente los resultados del método propuesto, para la GARMD genéricos, se consideró el corpus DUC01, que es un punto de referencia abierto para la evaluación de resúmenes escrito en inglés y compuesto por 309 documentos divididos en 30 colecciones. El propósito de esta tarea es generar un resumen informativo de cierto periodo de tiempo. Las tasas de compresión a la que fueron generados estos resúmenes fueron 50, 100, 200 y 40 palabras. Además, este corpus incluye 2 resúmenes de referencia (escritos de forma abstractiva) para su evaluación.

5.1.2 TAC08

Para evaluar la tarea de actualización se consideró el conjunto de datos TAC08, que comprende 48 temas cada uno con 20 documentos divididos en dos conjuntos A y B. El conjunto A precede cronológicamente a los documentos del conjunto B. El objetivo de esta tarea es generar dos resúmenes, el primero para dar contexto de la noticia hasta cierto periodo de tiempo (resumen A), y el segundo con información de actualización sobre la noticia (resumen B), asumiendo que el lector ya conoce la información previa y solo requiere conocer la nueva información. La extensión de los resúmenes es de 100 palabras, mismos que deben ser comparados con 4 resúmenes de referencia para su evaluación.

5.2 Evaluación

En la sección 5.2.1 se describen las medidas con las que fueron evaluados los resúmenes generados por método propuesto para las diferentes tareas. Por otro lado, en la sección 5.2.2 se describe los métodos del estado del arte y las heurísticas con las que es comparado el desempeño del método propuesto.

5.2.1 Medidas

Los resultados obtenidos fueron evaluados con el sistema ROUGE (Recall-Oriented Understudy for Gisting Evaluation), para establecer la calidad de un resumen creado por el método propuesto contrastándolo con resúmenes ideales creados por humanos. Estas medidas cuentan el número de unidades superpuestas como n-gramas, secuencias de palabras y pares de palabras mediante Rouge-1 y Rouge-2 para la GARMD genéricos. Mientras que para GARMD de actualización se consideró Rouge-2 y Rouge-SU4 (Dang & Owczarzak, 2008).

5.3 Resultados

Los resultados derivados de este trabajo están organizados de la siguiente manera: Primero se muestran los resultados respecto a la selección de características hibridas y generación de coeficientes de relevancia de los resúmenes de referencia escritos por humanos. A partir estos coeficientes de relevancia, se evaluó la selección de oraciones a través del AG descrito en el capítulo anterior.

5.3.1 Selección de características

La GAR extractivos implica realizar una selección de oraciones a través de características de diferentes niveles textuales y lingüísticos. De esta manera, se busca generar resúmenes que expongan las ideas principales de un conjunto de documentos fuente.

A partir del cálculo de características textuales, se debe cubrir el contenido del texto y mantener la diversidad de oraciones. A pesar de ello, el uso de características

estadísticas por si solas puede no capturar información relevante, por no considerar el significado del texto. Por otro lado, al confiar únicamente en características lingüísticas, no se capturan características estadísticas importantes. En vista de esto, se calcularon 68 características de tipo estadísticas y lingüísticas de los resúmenes de referencia escritos por humanos. Esto fue realizado con el propósito de determinar cuáles de ellas implícitamente considera el humano al redactar un resumen.

Los resultados obtenidos de dicho calculo se muestran en la figura 9, donde puede observarse lo siguiente de ambos conjuntos de datos:

- Las características que predominan sobre las demás en ambos conjuntos de datos son las siguientes: palabras temáticas, sustantivo común no plural, similitud con la oración principal y TF-IDF. Incluso la suma de estas características en ambos conjuntos de datos representa el 58% en DUC01 y 53% en TAC08.
- Las palabras temáticas son las que obtuvieron la más alta ponderación en ambos conjuntos de datos (32% en DUC01 y 26% en TAC08). De estos indicadores, se puede suponer que el humano tiende a seleccionar en mayor medida las palabras temáticas para generar resúmenes.
- A diferencia de lo anterior, existen 48 características en las que cada una representa menos del 1% del total. Sin embargo, en conjunto representan el 11% de relevancia en DUC01 y 13% de relevancia en TAC08. Además de estas, se muestran otras con porcentajes cercanos del 1%, como por ejemplo algunas entidades nombradas (DATE y GPE) o algunas etiquetas POS (CC, VB y VBN).
- En el caso del cálculo en DUC01, se observa que los resúmenes escritos por humanos contienen un porcentaje de redundancia del 1%. Esto puede suponer que dichos resúmenes cuentan con un grado de redundancia, aunque sea mínimo.

Si se emplearan todas las características analizadas, implicaría realizar un gran esfuerzo computacional. Por lo tanto, es necesario definir un criterio de selección. En este caso se determinó que el criterio para seleccionar las características a considerar sería a partir de la frecuencia de las características candidatas como se estableció en (Jo, 2019). En

consecuencia, debido a la baja ponderación que obtuvieron varias características, se omitieron 48 características con menor relevancia para los siguientes experimentos. Por lo tanto, se consideraron únicamente 19 con mayor puntuación para el método propuesto.

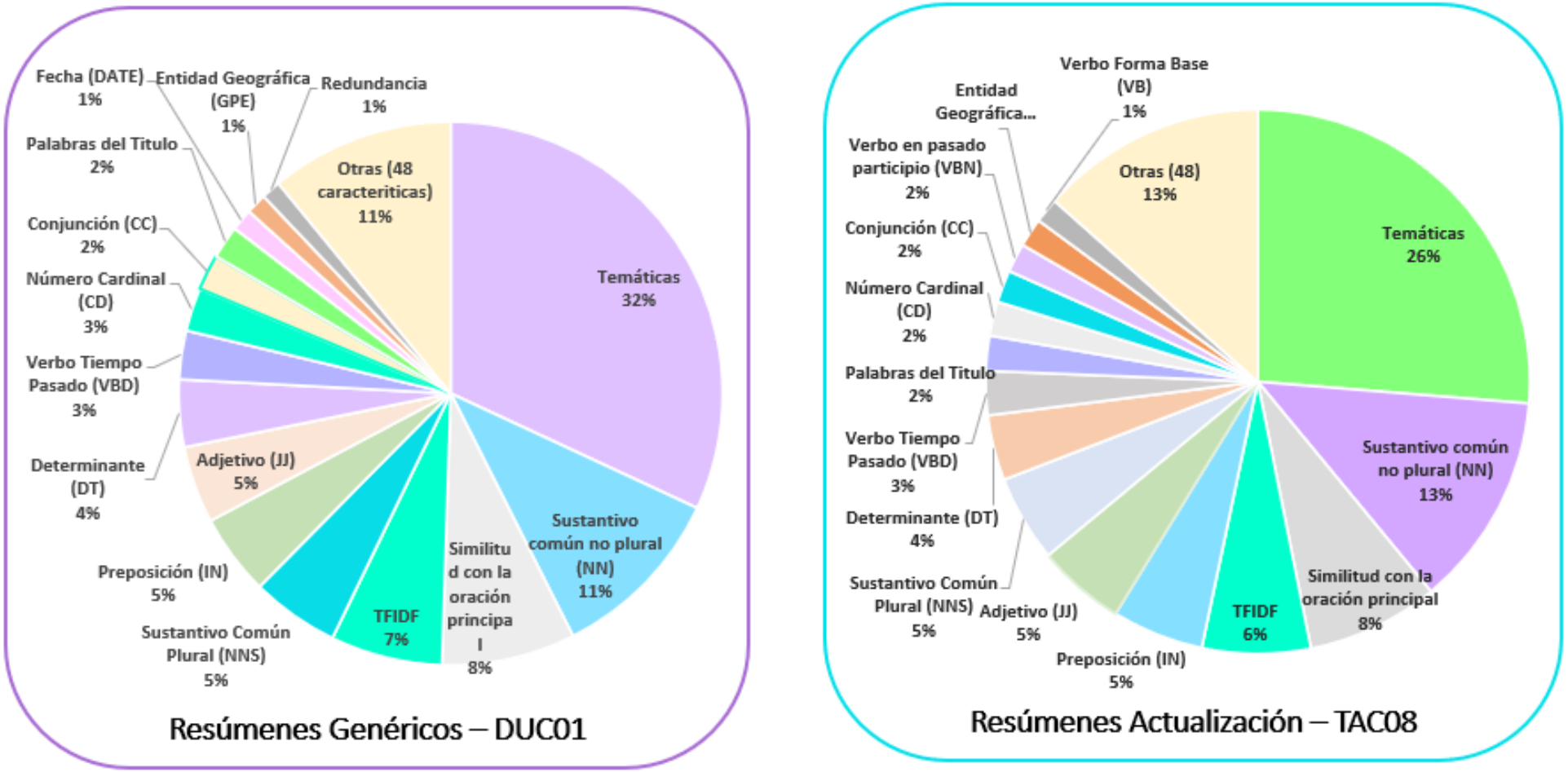


Figura 9. Ponderación de características

5.3.2 Coeficientes de Relevancia de características

Para determinar si los coeficientes de relevancia asignados a cada característica afecta a la calidad de los resúmenes generados. Se realizó una prueba en el AG con 5 generaciones para observar su efecto en la selección de oraciones.

Tabla 2. Prueba con y sin coeficientes de relevancia

<table>
<tr><td colspan="6">Resúmenes Genéricos 100 Palabras</td></tr>
<tr><td colspan="6">Parámetros</td></tr>
<tr><td># de individuos</td><td>Operador de Selección</td><td>Elite</td><td>Cruza</td><td>Mutación</td><td># Generaciones</td></tr>
<tr><td>2 * Cantidad de oraciones</td><td>Ruleta</td><td>3</td><td>Uniforme (98%)</td><td>Inversión (0.09%)</td><td>5</td></tr>
<tr><td colspan="3">Selección de oraciones sin coeficientes de relevancia
$\mathrm{FA}(D) = Max\left(\sum_{i=1} s_i\right)$</td><td colspan="3">Selección de oraciones con coeficientes de relevancia:
$FA(D = Max \sum \left(\sum s_{i*W_i}\right)$</td></tr>
<tr><td colspan="3">ROUGE-1: 25.04</td><td colspan="3">ROUGE-1: 27.9</td></tr>
</table>

Como puede observarse en la tabla 2, al incorporar los coeficientes de relevancia se tiene una mejora de +2.9 en relación con la selección de oraciones que no considera la incorporación de coeficientes de relevancia.

5.3.3 Selección de oraciones

Para evaluar la efectividad de la selección de características y sus coeficientes de relevancia obtenidos, se evaluó el método propuesto en resúmenes genéricos y de actualización. En la sección 5.3.3.1 se exponen los resultados obtenidos del modelado para la GARMD de genéricos (DUC01), mientras en la sección 5.3.3.2 se muestran los resultados para la GARMD de actualización (TAC08).

Además, en función de la tabla 1, donde se asignaron niveles de contribución a los requerimientos de la GARMD, niveles textuales y de estructura de lenguaje, así como categoría a cada una de las características que fueron consideradas en el modelo propuesto. Fueron calculados los porcentajes de participación de las características en el resumen generado, por medio de los coeficientes para cada categoría mediante la siguiente ecuación.

$$\%\ Contribución\ = \frac{(Coeficiente\ Relevancia * 100)}{\sum Coeficientes\ Relevancia\ (categoria)}$$

Ecuación 24. Ecuación para obtener el porcentaje de participación de las características.

5.3.3.1 Resúmenes genéricos (DUC01)

En este conjunto de datos fueron generados resúmenes de 4 longitudes (50, 100, 200 y 400 palabras), para probar el desempeño del método propuesto en diferentes longitudes de resumen.

Métodos del Estado del Arte y heurísticas

A continuación, se describen brevemente las heurísticas que se utilizaron para comparar el desempeño del método propuesto.

- **Topline:** Se considera una heurística que tiene el objetivo de generar los mejores resúmenes extractivos. Los resultados de dichos resúmenes se utilizan como referencia para establecer el límite superior el cual es posible lograr mediante métodos del estado del arte.
- **Baseline-first:** Es una heurística que toma las primeras oraciones de la colección de documentos en secuencia cronológica para generar resúmenes extractivos, hasta alcanzar el tamaño de resumen objetivo. Cualquier método del EA que genere mejores resúmenes que el Baseline-first se puede considerar inteligente.
- **Baseline-random:** A diferencia del Baseline-first, esta heurística consiste en seleccionar oraciones al azar del documento fuente para incorporarlas como un resumen extractivo, hasta que se logre la longitud requerida.

- **Baseline-first document:** Es una heurística que se basa en incluir las primeras palabras del primer documento, hasta que se cumpla el tamaño de resumen objetivo. En otras palabras, toma el contenido del primer documento escrito.
- **Lead Baseline:** Toma las primeras 50, 100, 200 y 400 palabras del último documento (más reciente) para incluirlas como resumen. Para esta heurística, los documentos se deben ordenar cronológicamente.

Además de las heurísticas, el siguiente listado describe los métodos del estado del arte con los que el método propuesto es comparado.

- **CBA:** Es un método basado en agrupamiento de oraciones para generar resúmenes extractivos. Este agrupamiento es realizado bajo la suposición de que cada oración se considera un tópico. En particular, CBA utiliza dos tipos de agrupamiento: jerárquico y de particionamiento (K-means) para seleccionar las oraciones para el resumen final (Boros et al., 2001).
- **NeATS:** Este método considera las siguientes características para ponderar cada oración de los documentos fuente: frecuencia términos, posición de la oración, palabras estigma y una versión simplificada de la Máxima Relevancia Marginal (MMR). Además de estas características, NeATS utiliza agrupamiento de términos, un "sistema de compañeros" de oraciones emparejadas y anotación temporal explícita para la selección de oraciones (Lin & Hovy, 2002).
- **AG (2 características):** Además del AG propuesto, en (Mendoza et al., 2022) se propuso optimizar la selección de oraciones usando un AG. Sin embargo, en este trabajo solo se emplearon dos características para la selección de oraciones, las cuales son: cobertura y posición de oraciones.
- **MBR:** Además del AG, en este trabajo también se realizó una aproximación con aprendizaje profundo,a través de la Máquina de Boltzmann Restringida (MBR). A diferencia de redes neuronales tradicionales como el Perceptrón Multicapa (MLP), la MBR es de bajo costo computacional y es útil para descubrir nuevas relaciones entre características.
- **Baldwin:** Este método se basa en el concepto de emplear "frases y palabras interesantes", las cuales se consideran candidatas para la extracción de

oraciones (Baldwin & Ross, 2001). De acuerdo con los autores de este método, una frase interesante es aquella que contiene palabras cuya entropía relativa es baja respecto a la colección de documentos.

Resultados en resúmenes genéricos (DUC01)

100 palabras: La tabla 3 muestra la comparación del método propuesto con los métodos del EA y heurísticas (parte superior izquierda) en resúmenes de 50 palabras, así como los parámetros del AG (parte superior derecha) y ponderación de características que fueron empleadas en esta tarea (parte inferior). De acuerdo con los resultados mostrados en la parte superior derecha, el método propuesto supera a las heurísticas y métodos del estado del arte, obteniendo 27.405 en ROUGE-1. Sin embargo, es superado por el AG (2 características), obteniendo 28.023 en ROUGE-1.

En cuanto al procesamiento de resúmenes de esta longitud, se consideró ágil, pues con 15 generaciones se obtuvo el mejor resultado. El número de individuos por generación fue de 2 * el número de oraciones. Como operadores genéticos se utilizó ruleta, cruza uniforme del 98% y mutación por inversión de 0.009%.

Por último, en la parte inferior de la tabla se observa que las características estadísticas están más presentes en los resúmenes, ya que representan el 91% de relevancia. Esto se debe a que la longitud de resúmenes es muy corta, por lo que la inclusión de etiquetas POS y NER son poco incluidas. En cuanto a los niveles de lenguaje, la mayoría de las características se encuentran en el nivel léxico.

200 palabras: La tabla 4 muestra los resultados obtenidos del método propuesto en la tarea de resúmenes genéricos de 100 palabras. Además, muestra una comparación del método propuesto con heurísticas y métodos del estado del arte, así como los parámetros del AG y la ponderación de características. En la parte superior izquierda se observa que el método propuesto supera a todos los métodos del estado del arte y heurísticas, obteniendo 34.053 en ROUGE-1. Asimismo, el método propuesto es el que muestra más cercanía al Topline (47.256). Para obtener estos resultados, se utilizaron los parámetros mostrados en la parte superior derecha. De estos parámetros, se observa

que fueron similares a los presentados en la tabla 2. Sin embargo, para esta longitud se incrementó el número de generaciones a 85. Por lo tanto, se puede suponer que, para la obtención de resúmenes de esta longitud el AG debe realizar más exploración, ya que el espacio de búsqueda para la selección de oraciones es mayor.

En cuanto a la relevancia de características, se muestran diferencias respecto a la tabla anterior, donde la participación de las características lingüísticas (54.6%) es mayor a las estadísticas (45.4%). En otras palabras, la mayoría de las características provienen de la identificación de un subconjunto de etiquetas POS y NER. Por otro lado, la mayoría de las características se encuentran en el nivel léxico, seguido por el nivel sintáctico. Al final de esta comparación, se encuentran en menor medida las características de nivel semántico. Tomando en cuenta los requerimientos de la GARMD (ver Capítulo 2), los resúmenes generados cumplen en mayor medida en Cobertura y Relevancia, mientras la reducción de redundancia se toma en menor medida.

200 palabras: Al igual que en los resultados previos, la tabla 5 muestra una comparación entre el método propuesto, los métodos del EA y heurísticas (parte superior izquierda) en resúmenes de 200 palabras, así como los parámetros del AG (parte superior derecha) y relevancia de características que fueron empleadas en esta tarea (parte inferior). En la comparación de métodos y heurísticas se observa que el método propuesto muestra el mejor desempeño, obteniendo 41.432 en ROUGE-1, seguido de AG (2 características), obteniendo 40.372 en ROUGE-1. Además, la MBR obtuvo un buen desempeño en esta comparación (39.584), pero resultó menor al método propuesto y al AG (2 características). Posterior a esto, se encuentra el Baseline-first (39.28) y NeATS (37.883).

En cuanto a los parámetros del AG, el número de individuos por generación permaneció con la misma cantidad que en los experimentos anteriores (2 * el número de oraciones). El operador utilizado para estos experimentos fue Torneo, con tamaños de k igual a 2, el número de individuos élite por generación fue de 3. Además, se utilizó la cruza uniforme con una probabilidad del 98%, mutación por inversión con una probabilidad de 0.012% y 135 generaciones. De estos parámetros se observa que, se incrementó el número de generaciones y la probabilidad de mutación para favorecer la exploración del AG.

Finalmente, en la parte inferior de la tabla 5 se muestra que hay una mayor participación de las características lingüísticas (57.3%) que las estadísticas (42.7%). Por lo tanto, se muestra mayor relevancia en las etiquetas POS y NER que las características estadísticas como TF-IDF o palabras temáticas. Con relación a los niveles de estructura del lenguaje, la mayor parte de las características se encuentran en el nivel léxico, seguido del nivel sintáctico y semántico. Además de esto, también es importante mencionar que los resúmenes generados consideran en mayor medida la Cobertura y Relevancia, mientras que la reducción de redundancia se toma en menor medida.

400 palabras: En la tabla 6, se muestran los resultados obtenidos del método propuesto en la tarea de resúmenes genéricos de 400 palabras. Adicionalmente, se muestran los parámetros del AG propuesto y la ponderación de características. En cuanto a la comparación de métodos del estado del arte y heurísticas se observa que el método propuesto supera los métodos del estado del arte y heurísticas, obteniendo 49.533 en ROUGE-1. Comparado con el AG (2 características) y la MBR, el método propuesto muestra una diferencia más significativa a los experimentos previos.

Por otro lado, los parámetros del AG fueron similares a los experimentos previos: número de individuos: 2 * número de oraciones, selección por Torneo con $k = 2$, selección elitista de 3 individuos por generación, cruza uniforme de 98% y mutación por inversión de 0.012%. No obstante, se optó por incrementar el número de generaciones a 180, debido a que mientras la longitud del resumen incrementa, el espacio de búsqueda del AG es mayor.

Finalmente, de la ponderación de características se observa lo siguiente:

- Las características estadísticas cuentan con un mayor porcentaje de relevancia (54.6%) que las características lingüísticas (44.4%). En otras palabras, se observa que las características estadísticas como TF-IDF, palabras positivas y palabras temáticas son más útiles para generar resúmenes de 400 palabras. Sin embargo, también hay un alto porcentaje de características derivadas de las etiquetas POS y NER.

- Con relación a los niveles de estructura del lenguaje, la mayoría de las características están en el nivel léxico, seguido del nivel sintáctico y semántico. Esta misma tendencia se observa en los resultados de las tablas 2, 3 y 4.
- Considerando los requerimientos de la GARMD, las características utilizadas se enfocan en mayor medida a la Cobertura y Relevancia de términos/oraciones, mientras se emplea en menor medida la reducción de redundancia.
- En cuanto a los niveles de granularidad, el método propuesto se enfoca en mayor grado a ponderar en función de las palabras, que a nivel de oración o documento. Sin embargo, la mayor parte de estas características cumplen con este nivel de ponderación.

Tabla 3. Resultados de resúmenes genéricos (50 Palabras)

Resúmenes Genéricos 50 Palabras

Método	**Rouge-1**
Topline	40.395
AG (2 características)	28.023
Método Propuesto	27.405
MBR	27.369
Baseline-first-document	25.435
Baseline-first	25.194
Baldwin	22.906
CBA	22.679
Lead Baseline	22.620
NeATS	22.594
Baseline-random	20.027

Parámetros del AG	
Número de individuos	$2 * Cantidad\ de\ oraciones$
Operador de selección	Ruleta
Número de individuos Elite	3
Cruza	Uniforme (98%)
Mutación	Inversión (0.009%)
Número de Generaciones	15

Modelado de características:

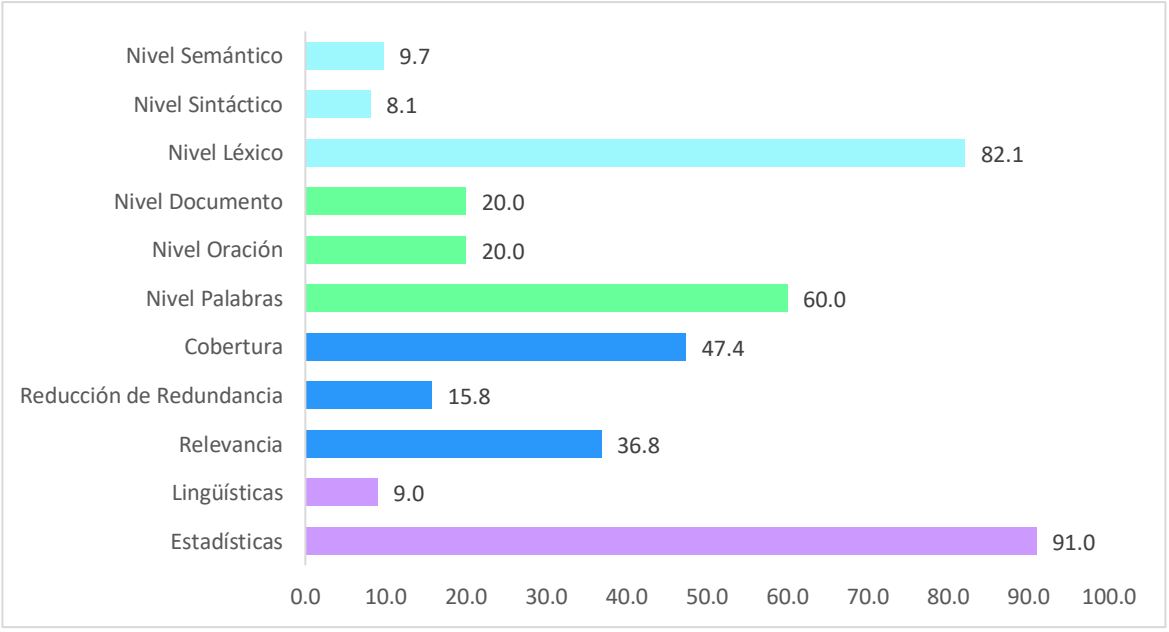


Tabla 4. Resultados de resúmenes genéricos (100 Palabras)

Resúmenes Genéricos 100 Palabras

Método	**Rouge-1**
Topline	47.256
Método Propuesto	34.053
AG (2 Características).	33.985
MBR	32.923
Baseline-first	31.716
Baseline-first-document	30.462
Baldwin	28.647
NeATS	28.195
Lead Baseline	28.195
Baseline-random	26.994
CBA	26.741

Parámetros del AG	
Número de individuos	$2 * Cantidad\ de\ oraciones$
Operador de selección	Ruleta
Número de individuos Elite	3
Cruza	Uniforme (98%)
Mutación	Inversión (0.009%)
Número de Generaciones	85

Modelado de características

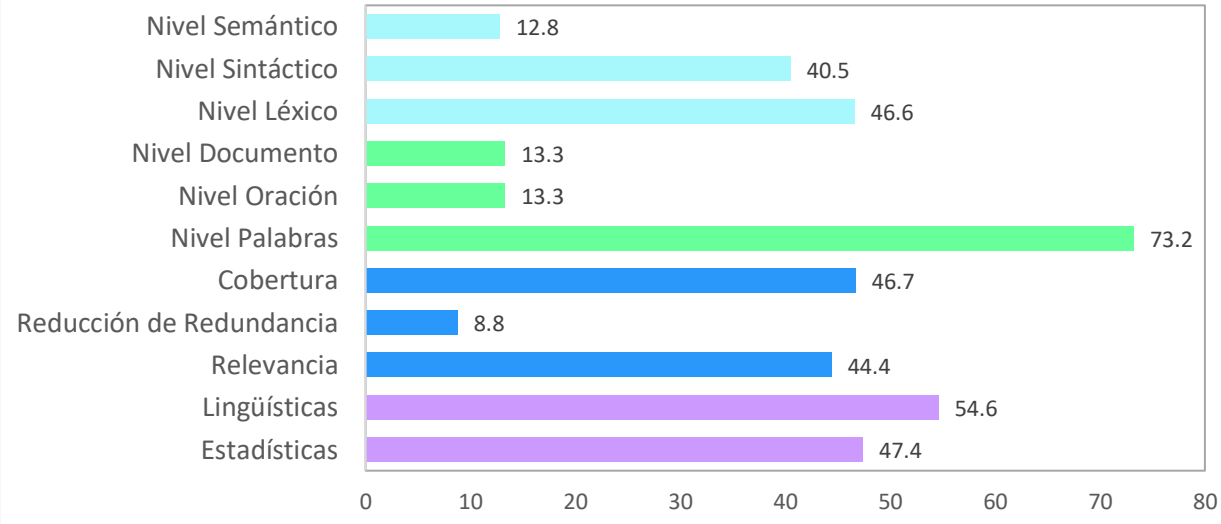


Tabla 5 Resultados de resúmenes genéricos (200 Palabras)

Resúmenes Genéricos 200 Palabras

Método	**Rouge-1**
Topline	53.63
Método Propuesto	41.432
AG (2 Características).	40.372
MBR	39.584
Baseline-first	39.28
NeATS	37.883
Baldwin	35.89
BFD	35.472
CBA	34.108
Baseline-random	34.057
Lead Baseline	34.009

Parámetros del AG	
Número de individuos	$2 * Cantidad\ de\ oraciones$
Operador de selección	Torneo (k=2)
Número de individuos Elite	3
Cruza	Uniforme (98%)
Mutación	Inversión (0.012%)
Número de Generaciones	135

Modelado de características

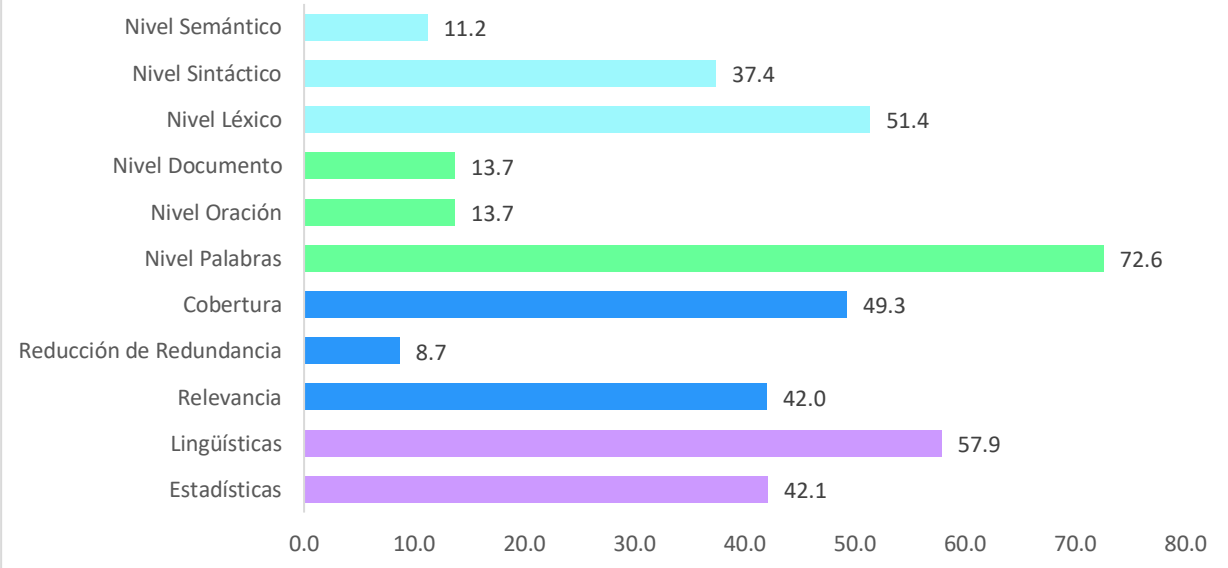


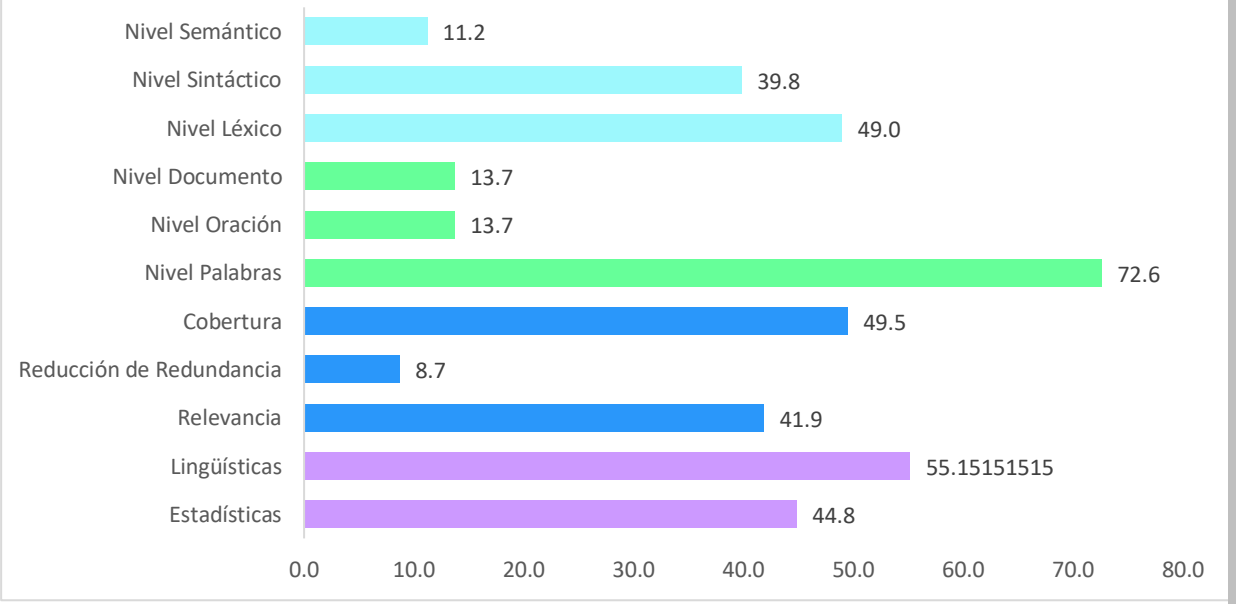
Tabla 6 Resultados de resúmenes genéricos (400 Palabras)

Resúmenes Genéricos 400 Palabras

Método	**Rouge-1**
Topline	60.691
Método Propuesto	49.533
AG (2 Características)	47.619
MBR	47.275
Baseline-first	47.198
NeATS	45.551
Baseline-random	42.131
CBA	41.259
Baseline-first-document	41.161
Lead Baseline	39.961

Parámetros del AG	
Número de individuos	$2 * Cantidad\ de\ oraciones$
Operador de selección	Torneo (k=2)
Número de individuos Elite	3
Cruza	Uniforme (98%)
Mutación	Inversión (0.012%)
Número de Generaciones	180

Modelado de características



5.3.3.2 Resultados en resúmenes de actualización (TAC08)

El taller de resúmenes de actualización (denominado TAC08) contó con 33 participantes de todo el mundo, proponiendo diferentes enfoques y métodos para generar resúmenes de actualización. En conjunto presentaron un total de 71 resultados diferentes. Para comprar los resultados del método propuesto se consideró el mejor método (ICSI), de mediano (Abawakid) y de más bajo desempeño (LIPN).

- **ICSI:** Se basa en un marco general que proyecta el resumen como un problema de optimización global con una solución de programación lineal entera (Gillick et al., 2008)
- **Abawakid:** Este trabajo utilizó una función de puntuación para identificar las oraciones más relevantes. Entre las características textuales utilizadas están la posición de las oraciones, la ubicación de las oraciones, la similitud oración-oración (Dang & Owczarzak, 2008).
- **LIPN:** Se empleó el algoritmo K-means para calcular la similitud de oraciones para detectar la novedad entre los resúmenes A y B.

En la tabla 7 se muestran los resultados obtenidos del método propuesto en la GARMD iniciales (A) y de actualización (B) a 100 palabras. Adicionalmente, se muestran los parámetros del AG propuesto y la ponderación de características. En cuanto a la comparación de métodos del estado del arte y heurísticas, se observa que el método propuesto supera la mayoría de los métodos del estado del arte y heurísticas en los resúmenes tipo A, obteniendo 10.889 en ROUGE-2. Sin embargo, es superado por ICSI (10.900) en una diferencia muy corta de 0.011. Por otro lado, el método propuesto obtiene el mejor desempeño en los resúmenes de tipo B (10.066), superando a todas las heurísticas y métodos.

Respecto a los parámetros del AG, el número de individuos por generación incrementó, debido al espacio de búsqueda que implica realizar la selección de oraciones. Entonces el número de individuos fue dado de la siguiente manera: 3 * el número de oraciones. Los operadores de selección empleados fueron Ruleta y Elitismo, éste último de 3 individuos por generación. Además de la selección, se utilizó la cruza uniforme con una

probabilidad del 98%, mutación por inversión con una probabilidad de 0.016% y 120 generaciones. Estos parámetros fueron seleccionados para favorecer la exploración en el espacio de búsqueda.

Finalmente, para la ponderación de características, el método propuesto presenta diferencias significativas respecto a los experimentos previos. Por lo tanto, se observa lo siguiente:

- Las características estadísticas cuentan con un mayor porcentaje de relevancia, en comparación con las características estadísticas. En otras palabras, se puede decir que TF-IDF, palabras positivas, palabras temáticas y centralidad de la oración son importantes para generar resúmenes de actualización.
- Con relación a los niveles de estructura del lenguaje, el método propuesto considera en mayor grado las características léxicas, seguido de las características sintácticas y semánticas.
- En cuanto a los requerimientos de la GARMD mencionados en el Capítulo 2, el método propuesto tiene una mayor relevancia hacia la Cobertura y Relevancia, por lo que se considera en menor grado la reducción de redundancia.
- En cuanto a los niveles de granularidad (oración, palabra, documento), el método propuesto se centra a ponderar en palabras de mayor grado, mientras a nivel de oración y documento permanece en igualdad porcentaje

Tabla 7. Resultados de resúmenes de actualización

Resúmenes de Actualización

Resumen A

Método	**Rouge-2**
Topline	15.607
ICSI	10.900
Método Propuesto	10.889
AG (2 Características)	8.592
Abawakid	7.900
Baseline-first-document	6.512
Baseline-first	5.851
Lead Baseline	5.800
LIPN	4.400
Baseline-random	4.253

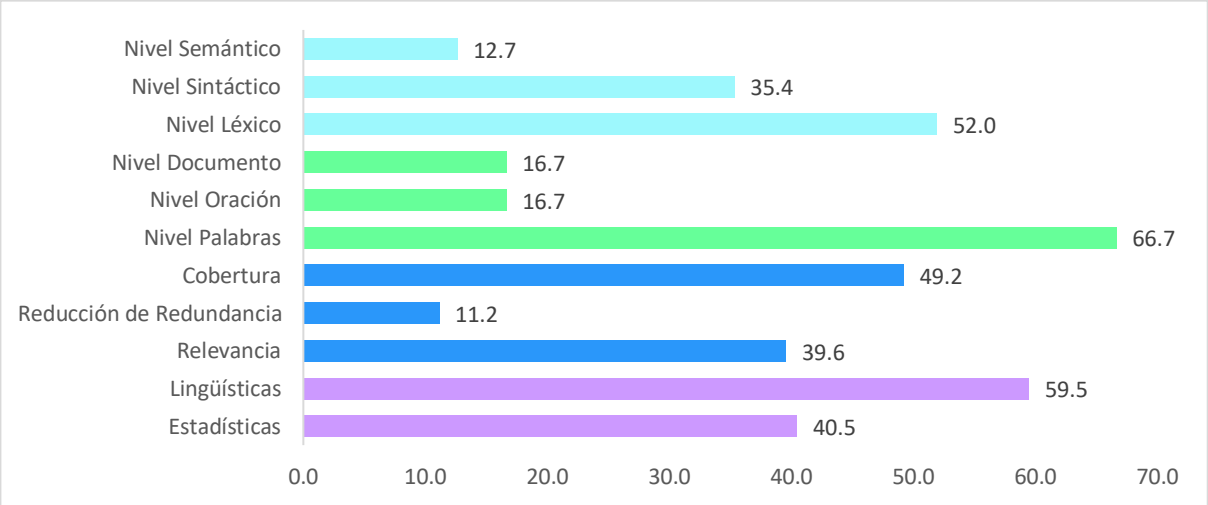
Resumen B

Método	**Rouge-2**
Topline	16.119
Método Propuesto	10.066
AG (2 Características)	9.700
ICSI	9.400
Abawakid	8.100
Baseline-first-document	5.502
Baseline-first	6.359
Lead Baseline	6.000
LIPN	3.400
Baseline-random	4.091

Parámetros del AG

Número de individuos	Operador selección	Número de individuos Elite	Cruza	Mutación	Número de Generaciones
$3 * Cantidad\ de\ oraciones$	Ruleta	3	Uniforme (98%)	Inversión (0.016%)	120

Modelado de características



5.4 Resumen del capítulo

En este capítulo, se mostraron los resultados obtenidos del método propuesto para la GARMD genéricos y de actualización. Para estos experimentos, se emplearon y describieron conjuntos de datos como DUC01 para la GARMD genéricos y TAC08 para la GARMD de actualización. Además, se definieron medidas de evaluación de ROUGE (ROUGE-1 y ROUGE-2) para establecer una comparación entre métodos del estado del arte y heurísticas. Asimismo, se describieron brevemente las heurísticas y métodos del estado del arte empleados para la comparación.

Respecto al modelado de características en los resúmenes de referencia escritos por humanos, se observó que algunas características estadísticas como palabras temáticas, TF-IDF y similitud con la oración principal mantienen un alto coeficiente de relevancia en estos resúmenes. El criterio para seleccionar las características a considerar fue a partir de la frecuencia de las características candidatas, por lo que se consideraron únicamente 19 de ellas para reducir el costo computacional que implica utilizar todas en la selección de oraciones.

Posterior a la generación de coeficientes de relevancia de las características, se utilizaron estos coeficientes como referencia para realizar una selección de oraciones con el AG mencionado en el capítulo anterior. En cuanto a los resúmenes genéricos, el método propuesto supera a todas las heurísticas y métodos del estado del arte. En particular, en los resúmenes de 400 palabras se observan diferencias significativas entre el método propuesto y las heurísticas/métodos del estado del arte. Para obtener estos resultados, se realizaron experimentos con el AG propuesto a través de diferentes parámetros. Los parámetros colocados en las tablas 2, 3, 4 y 5 fueron seleccionados, debido a los resultados obtenidos en la evaluación de ROUGE.

Por otro lado, en la GARMD de actualización se muestran los resultados del método propuesto al generar resúmenes iniciales (A) y de actualización (B). De acuerdo con estos resultados, el método propuesto supera a las heurísticas y métodos del estado del arte (a excepción de ICSI en los resúmenes A). Además, se emplearon algunos cambios

en los parámetros del AG para obtener una mejor selección de oraciones, incrementando el número de individuos por generación y la probabilidad de mutación.

Respecto al análisis de ponderación de características, aquellas de carácter estadístico son las que mantienen una mayor relevancia que las lingüísticas. En cuanto a los requerimientos de la GARMD, la Cobertura y Relevancia son requerimientos que tomaron mayor relevancia para la selección de oraciones. Sin embargo, la reducción de redundancia es un requerimiento tomado en menor medida.

En general, de acuerdo los resultados mostrados en este capítulo, se comprobó la hipótesis planteada, pues al obtener un modelado de características híbridas a partir de los resúmenes de referencia escritos por humanos, se logró mejorar la selección de oraciones para la GARMD extractivos.

Asimismo, con estos experimentos se cumplieron los siguientes objetivos específicos:

1. Modelar las características de los documentos de referencia escritos por humanos.
2. Optimizar la selección de oraciones a través del AG.
3. Evaluar y analizar el desempeño de los resúmenes generados a través de ROUGE.
4. Comparar los resultados obtenidos del modelo propuesto con los trabajos del estado del arte.

CAPÍTULO 6
Conclusiones y trabajo futuro

En esta investigación se realizó un modelado de características híbridas a partir de los resúmenes de referencia escritos por humanos para la GARMD extractivos. De acuerdo con los resultados mostrados en el Capítulo 5, el modelo propuesto mejora la selección de oraciones de dicha tarea. Asimismo, se cumplen con los siguientes objetivos:

1. Se generaron resúmenes genéricos y de actualización a partir de 2 conjuntos de datos (DUC01 y TAC08) con diferentes características y requerimientos.
2. Se analizaron diferentes formas de concatenación de documentos. Sin embargo, se optó por la concatenación jerárquica cronológica, debido a que ésta ha sido de fácil implementación y ha proporcionado coherencia cronológica en la selección de oraciones.
3. Con base en el análisis de características propuestas en el Estado del Arte (EA), en esta investigación se utilizaron 63 de ellas para la GARMD extractivos. No

obstante, se emplearon 19 características, ya que mantienen altos valores en sus coeficientes de relevancia.

4. Una vez obtenido el modelo de características de los resúmenes de referencia escritos por humanos, se optimizó la selección de oraciones a través del Algoritmo Genético (AG).

En comparación con otros métodos del EA y heurísticas, el modelado de características propuesto logra el mejor desempeño al generar resúmenes genéricos de longitudes largas (200 y 400 palabras). Por otra parte, el modelado logra ser competitivo al generar resúmenes genéricos de longitudes cortas (50 y 100 palabras). Por lo tanto, no se rechaza la hipótesis planteada en el Capítulo 1.

En general, la inclusión de características híbridas en la GARMD ayuda a extraer información relevante de los resúmenes de referencia escritos por humanos. De esta manera, se buscó determinar que oraciones incluir en el resumen, considerando diferentes niveles textuales (palabra, oración, párrafo y documento) y requerimientos (relevancia, cobertura y reducción de redundancia). Como se observó en los resultados mostrados en el Capítulo 5, las características lingüísticas fueron de mayor relevancia que las estadísticas. En cuanto a los niveles textuales, las características a nivel de palabra fueron más relevantes que a nivel de oración, párrafo o documento. Finalmente, la mayor parte de las características empleadas se encuentran en el análisis léxico.

Dentro del EA, se ha estudiado ampliamente un conjunto de características textuales. Sin embargo, no todas son ponderadas con el mismo nivel de relevancia. En la literatura, esta relevancia se ha calculado mediante dos enfoques:

1. Desde del texto fuente: se calcula las características y derivado de este cálculo se seleccionan las oraciones que serán incluidas en el resumen.
2. Asignando coeficientes de relevancia: Además de obtener las características del texto fuente, a cada una de ellas se le asocia un coeficiente de relevancia para determinar la importancia de cada oración del texto fuente. Estos coeficientes han sido calculados a través de optimización (Fattah, 2016; Jain et al., 2022; Verma & Om, 2019), aprendizaje automático (Binwahlan et al., 2009; Jo, 2019; Li

et al., 2020; Mahalleh & Gharehchopogh, 2022), o por asignación manual (Mendoza, 2015; Qaroush et al, 2021; Vázquez et al, 2018).

Derivado de esta investigación, fue posible conocer la utilidad de modelar las características textuales a partir de los resúmenes de referencia escritos por humanos. Para conocer dicha utilidad, se realizó un cálculo de coeficientes de relevancia de las características híbridas consideradas en el método propuesto. Posteriormente, se realizó una selección de oraciones a través del AG.

De acuerdo con la experimentación realizada en documentos pertenecientes al dominio de noticias y en idioma inglés, los resultados obtenidos en resúmenes genéricos (DUC01), el método propuesto presenta los mejores resultados en la generación de resúmenes de 200 y 400 palabras. Mientras que para las longitudes cortas (50 y 100 palabras), existe dificultad para superar a los métodos del EA.

En cuanto a los resúmenes iniciales y de actualización (TAC08), el método propuesto fue capaz de reconocer información actualizada, ya que en los resultados en resúmenes de tipo B (de actualización) superó a los métodos del estado del arte y heurísticas. Mientras que en los resúmenes de tipo A (iniciales) presenta limitaciones o dificultades para reconocer información introductoria.

6.1 Aportaciones

Calcular coeficientes de relevancia de las características híbridas en los resúmenes de referencia ha sido una alternativa útil como paso previo a la selección de oraciones, ya que estos documentos se consideran resúmenes objetivo para la GARMD extractivos. En otras palabras, mientras los resúmenes automáticos se asemejen más a los resúmenes de referencia, se cumple el principal propósito de la GARMD: crear sistemas que puedan generar resúmenes similares a cómo los realizaría el humano. Además, se centró en calcular de manera objetiva la importancia de las oraciones, generando modelos que permitan mejorar la selección de oraciones. No obstante, a pesar de que los resúmenes de referencia se suelen usar para evaluar sistemas de la GARMD extractivos, no se utilizaban para determinar la importancia de cada característica empleada.

Por otro lado, también se centró en analizar el grado que aportan las características híbridas con base en múltiples perspectivas. Es decir, si los resúmenes fueron construidos con diferentes características textuales (palabra, oración, párrafo o documento) y lingüísticas (a nivel léxico, sintáctico o semántico). O bien de acuerdo con los requerimientos de la GARMD (cobertura, reducción de redundancia y relevancia) y con base en tipo de extracción de información (estadística o lingüística).

De acuerdo con los resultados mostrados en el Capítulo 5, el modelado propuesto como paso previo a la selección de oraciones favorece la generación de resúmenes de calidad, pues el modelo propuesto supera al EA bajo la evaluación de resúmenes en ROUGE-1 y ROUGE-2 para resúmenes genéricos y de actualización, respectivamente.

6.2 Trabajo futuro

Como trabajo futuro, se sugiere que el modelado de características propuesto sea probado en textos de dominios no relacionados a noticias (académico, legal, científico o literario). De esta manera, se busca comprobar si el modelado propuesto es capaz de detectar relevancia en oraciones, detectar cobertura de temas y reducir redundancia. Además, sería de utilidad probar el modelado de características en la GARMD de otros idiomas (por ejemplo, español, portugués, ruso, etc.).

Por otro lado, sería interesante analizar el desempeño del modelado propuesto en la GAR de documentos individuales. Y en su caso, calcular o ajustar los coeficientes de relevancia para mejorar la selección de oraciones en dicha tarea. Finalmente, es importante incorporar otras características que no fueron mencionadas a lo largo de esta investigación.

6.3 Implicaciones éticas

Generar un buen resumen implica elegir fragmentos del texto original, para representar la idea del contenido de acuerdo con una longitud determinada. En los últimos años, la disponibilidad de herramientas de la GAR ha aumentado, tanto para uso personal como comercial. Sin embargo, un resumen contiene información sesgada, existe el riesgo de pasar ese sesgo a los lectores como un hecho. En consecuencia, el uso de herramientas de la GAR puede implicar sesgos que conllevan a un análisis de consideraciones éticas

(Hannah Brown, 2023; Liu et al., 2023). En este trabajo, se analizaron las consideraciones éticas desde las siguientes perspectivas:

- **Sesgo de datos:** Los resúmenes generados por el método propuesto presentan información relevante de diferentes documentos fuente. Por lo que estos resúmenes no buscan tergiversar o manipular los hechos en beneficio de otros interesados.
- **Sesgo estructural:** Al generar resúmenes de noticias, generalmente, se toman en cuenta las primeras oraciones de los documentos fuente. En este trabajo, el método propuesto genera resúmenes sin tomar en cuenta la posición de la oración en el documento fuente. De esta manera, se pretende dar la misma posibilidad de selección a cada una de las oraciones del documento fuente.
- **Responsabilidad:** El método propuesto, es una herramienta para ayudar al usuario en el proceso de adquirir información relevante sobre una serie de hechos, considerando su desarrollo en un tiempo determinado. Sin embargo, no pretende sustituir las actividades del humano, ya que los resúmenes generados solo proporcionan información para que el humano pueda formar su propio criterio.
- **Uso de datos y privacidad:** Los textos utilizados para evaluar el método propuesto no contienen información sensible y/o personal de otras personas u organizaciones. No obstante, el uso de estos textos requirió la autorización de NIST (Instituto Nacional de Estándares y Tecnología, por sus siglas en inglés). Acceso: https://www-nlpir.nist.gov/projects/duc/data.html
- **Medio ambiente:** A diferencia de otros modelos de lenguaje que requieren de un masivo aprendizaje de datos (*BERT, RoBERTa, Bloom)*, los modelos generados del método propuesto tienen un menor impacto ambiental, requiriendo una menor cantidad de recursos computacionales y eléctricos. Además, el modelo de lenguaje *word2vec,* el cual subyace al método propuesto no requirió entrenamiento adicional.

Referencias

Abdi, A., Hasan, S., Shamsuddin, S. M., Idris, N., & Piran, J. (2021). A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion. *Knowledge-Based Systems*, *213*, 106658. https://doi.org/10.1016/J.KNOSYS.2020.106658

Abdulateef, S., Khan, N. A., Chen, B., & Shang, X. (2020). Multidocument Arabic text summarization based on clustering and word2vec to reduce redundancy. *Information (Switzerland)*, *11*(2), 59. https://doi.org/10.3390/info11020059

Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text Summarization: A Brief Review. *Studies in Computational Intelligence*, *874*, 1–15. https://doi.org/10.1007/978-3-030-34614-0_1

Abuobieda, A., Salim, N., Albaham, A. T., Osman, A. H., & Kumar, Y. J. (2012). Text summarization features selection method using pseudo genetic-based model. *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, 193–197. https://doi.org/10.1109/INFRKM.2012.6204980

AL-Khassawneh, Y. A., & Hanandeh, E. S. (2023). Extractive Arabic Text Summarization-Graph-Based Approach. *Electronics 2023, Vol. 12, Page 437*, *12*(2), 437. https://doi.org/10.3390/ELECTRONICS12020437

Alami, N., El Mallahi, M., Amakdouf, H., Qjidaa, H., & Tools, M. (2020). *Hybrid method for text summarization based on statistical and semantic treatment*. https://doi.org/10.1007/s11042-021-10613-9

Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. *Expert Systems*, *36*(1), e12340. https://doi.org/10.1111/EXSY.12340

Alguliyev, R. M., Ramiz, |, Aliguliyev, M., Isazade, N. R., Abdi, A., & Idris, N. (2018). *COSUM: Text summarization based on clustering and optimization*. https://doi.org/10.1111/exsy.12340

Andhale, N., & Bewoor, L. A. (2017). An overview of text summarization techniques. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*. https://doi.org/10.1109/ICCUBEA.2016.7860024

Antony, D., Abhishek, S., Singh, S., Kodagali, S., Darapaneni, N., Rao, M., Paduri, A. R., & Bg, S. (2023). A Survey of Advanced Methods for Efficient Text Summarization. *2023 IEEE 13th Annual Computing and Communication Workshop and Conference, CCWC 2023*, 962–968. https://doi.org/10.1109/CCWC57344.2023.10099322

Aote, S. S., Pimpalshende, A., Potnurwar, A., & Lohi, S. (2023). Binary Particle Swarm Optimization with an improved genetic algorithm to solve multi-document text

summarization problem of Hindi documents. *Engineering Applications of Artificial Intelligence*, *117*, 105575. https://doi.org/10.1016/J.ENGAPPAI.2022.105575

Aries, A., Zegour, D. eddine, & Hidouci, W. K. (2019). *Automatic text summarization: What has been done and what has to be done*. https://arxiv.org/abs/1904.00688v1

Arya, C., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Multi-Document News Web Page Summarization Using Content Extraction and Lexical Chain Based Key Phrase Extraction. *Mathematics 2023, Vol. 11, Page 1762*, *11*(8), 1762. https://doi.org/10.3390/MATH11081762

Baldwin, B., & Ross, A. (2001). Baldwin language technology's DUC summarization system. *Proceedings of the 1st Document Understanding Conference, New Orleans, LA*.

Barzilay, R., & Elhadad, M. (1997). *Using lexical chains for text summarization*. https://doi.org/10.7916/D85B09VZ

Baxendale, P. B. (1958). Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, *2*(4), 354–361. https://doi.org/10.1147/rd.24.0354

Belwal, R. C., Rai, S., & Gupta, A. (2023). Extractive text summarization using clustering-based topic modeling. *Soft Computing*, *27*(7), 3965–3982. https://doi.org/10.1007/S00500-022-07534-6/METRICS

Binwahlan, M. S., Salim, N., & Suanmali, L. (2009). Swarm Based Text Summarization. *2009 International Association of Computer Science and Information Technology - Spring Conference, IACSIT-SC 2009*, 145–150. https://doi.org/10.1109/IACSIT-SC.2009.61

Boros, E., Kantor, P. B., & Neu, D. J. (2001). *A Clustering Based Approach to Creating Multi-Document Summaries*. https://www-nlpir.nist.gov/projects/duc/pubs/2001papers/rutgers_final.pdf

Brill, E. (1992). *A simple rule-based part of speech tagger*. 152. https://doi.org/10.3115/974499.974526

Cajueiro, D. O., Nery, A. G., Tavares, I., De Melo, M. K., Dos Reis, S. A., Weigang, L., & Celestino, V. R. R. (2023). *A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding*. https://zenodo.org/record/7500273.

Chuang, W. T., & Yang, J. (2000). Text summarization by sentence segment extraction using machine learning algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1805*, 454–457. https://doi.org/10.1007/3-540-45571-X_52/COVER

Dang, H., & Owczarzak, K. (2008). Overview of the TAC 2008 update summarization task. *Tac*, 1–16. http://www.nist.gov/tac/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf

Divya, S., & Sripriya, N. (2022). Unsupervised hierarchical text summarization. *AIP Conference Proceedings*, *2670*(1). https://doi.org/10.1063/5.0116918/2832163

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, *16*(2), 264–285. https://doi.org/10.1145/321510.321519

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. In *Expert Systems with Applications* (Vol. 165, p. 113679). Pergamon. https://doi.org/10.1016/j.eswa.2020.113679

Ghadimi, A., & Beigy, H. (2023). SGCSumm: An extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. *Expert Systems with Applications*, *215*, 119308. https://doi.org/10.1016/J.ESWA.2022.119308

Gillick, D., Favre, B., & Hakkani-t�r, D. (2008). The ICSI Summarization System at TAC 2008. *In Proceedings of the Text Analysis Conf. Workshop*, 801–815. https://pageperso.lis-lab.fr/benoit.favre/papers/favre_tac2008.pdf

Glazkova, A. V., & Morozov, D. A. (2023). Applying Transformer-Based Text Summarization for Keyphrase Generation. *Lobachevskii Journal of Mathematics*, *44*(1), 123–136. https://doi.org/10.1134/S1995080223010134/METRICS

Gupta, S., Sharaff, A., & Nagwani, N. K. (2023). Frequent item-set mining and clustering based ranked biomedical text summarization. *Journal of Supercomputing*, *79*(1), 139–159. https://doi.org/10.1007/S11227-022-04578-1/METRICS

Hannah Brown, R. S. (2023). HOW (UN)FAIR IS TEXT SUMMARIZATION? *Eleventh International Conference on Learning Representations*. https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/summarization/overview

Hendrastuty, N., & SN, A. (2021). Text Summarization in Multi Document Using Genetic Algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(4), 327–338. https://doi.org/10.22146/ijccs.66026

Hernandez-Castaneda, A., Garcia-Hernandez, R. A., Ledeneva, Y., & Millan-Hernandez, C. E. (2020). Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords. *IEEE Access*, *8*, 49896–49907. https://doi.org/10.1109/ACCESS.2020.2980226

Hosseinabadi, S., Kelarestaghi, M., & Eshghi, F. (2022). ISSE: a new iterative sentence scoring and extraction scheme for automatic text summarization. *International Journal of Computers and Applications*, *44*(6), 535–540. https://doi.org/10.1080/1206212X.2020.1829844

Jain, A., Arora, A., Morato, J., Yadav, D., & Kumar, K. V. (2022). Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm. *Applied Sciences (Switzerland)*, *12*(13), 6584. https://doi.org/10.3390/app12136584

Jalil, Z., Nasir, M., Alazab, M., Nasir, J., Amjad, T., & Alqammaz, A. (2023). Grapharizer: A

Graph-Based Technique for Extractive Multi-Document Summarization. *Electronics 2023, Vol. 12, Page 1895*, *12*(8), 1895. https://doi.org/10.3390/ELECTRONICS12081895

Jo, T. (2019). *Text Summarization* (pp. 271–294). Springer, Cham. https://doi.org/10.1007/978-3-319-91815-0_13

Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2023). DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. *Expert Systems with Applications*, *211*, 118442. https://doi.org/10.1016/J.ESWA.2022.118442

Kaljahi, R., Foster, J., & Roturier, J. (2014). Semantic Role Labelling with minimal resources: Experiments with French. *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, 87–92. https://doi.org/10.3115/v1/S14-1012

Kumar, Y., Kaur, K., & Kaur, S. (2021). Study of automatic text summarization approaches in different languages. *Artificial Intelligence Review*, *54*(8), 5897–5929. https://doi.org/10.1007/S10462-021-09964-4/TABLES/2

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*. https://doi.org/10.1016/J.AIOPEN.2022.03.001

Li, Z., Zheng, X., & He, J. (2020). *Unsupervised Summarization by Jointly Extracting Sentences and Keywords*. https://arxiv.org/abs/2009.07481v2

Lin, C.-Y., & Hovy, E. (2002). From single to multi-document summarization. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 457. https://doi.org/10.3115/1073083.1073160

Liu, Y. L., Cao, M., Blodgett, S. L., Cheung, J. C. K., Olteanu, A., & Trischler, A. (2023). *Responsible AI Considerations in Text Summarization Research: A Review of Current Practices*. http://arxiv.org/abs/2311.11103

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

Mahalleh, E. R., & Gharehchopogh, F. S. (2022). An automatic text summarization based on valuable sentences selection. *International Journal of Information Technology (Singapore)*, *14*(6), 2963–2969. https://doi.org/10.1007/s41870-022-01049-x

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

Mohamed, M., & Oussalah, M. (2019). SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing and Management*, *56*(4), 1356–1372. https://doi.org/10.1016/j.ipm.2019.04.003

Mojrian, M., & Mirroshandel, S. A. (2021). A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA. *Expert Systems with Applications*, *171*, 114555. https://doi.org/10.1016/j.eswa.2020.114555

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. *Knowledge-Based Systems*, *163*, 518–532. https://doi.org/10.1016/j.knosys.2018.09.008

Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*, *183*, 104848. https://doi.org/10.1016/j.knosys.2019.07.019

Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2020). Candidate sentence selection for extractive text summarization. *Information Processing & Management*, *57*(6), 102359. https://doi.org/10.1016/J.IPM.2020.102359

Nasar, Z., Syed, W., Jaffry, M., & Kamran, M. (2019). *Textual keyword extraction and summarization: State-of-the-art*. https://doi.org/10.1016/j.ipm.2019.102088

Ni, Y., Barzman, D., Bachtel, A., Griffey, M., Osborn, A., & Sorter, M. (2020). Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence. *International Journal of Medical Informatics*, *139*. https://doi.org/10.1016/J.IJMEDINF.2020.104137

Paul, A. M., & Salim, A. (2023). Text Summarization Using Lexical Chaining and Concept Generalization. *Lecture Notes on Data Engineering and Communications Technologies*, *131*, 793–809. https://doi.org/10.1007/978-981-19-1844-5_63/COVER

Qaroush, A., Abu Farha, I., Ghanem, W., Washaha, M., & Maali, E. (2021). An efficient single document Arabic text summarization using a combination of statistical and semantic features. *Journal of King Saud University - Computer and Information Sciences*, *33*(6), 677–692. https://doi.org/10.1016/J.JKSUCI.2019.03.010

Rajalakshmi, R., Vidhya, S., Harina, D., Karna, R., & Sowmya, A. (2023). Text Summarization for News Articles using Latent Semantic Analysis Technique. *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, 1421–1425. https://doi.org/10.1109/ICESC57686.2023.10193508

Ramani, K., Bhavana, K., Akshaya, A., Harshita, K. S., Thoran Kumar, C. R., & Srikanth, M. (2023). An Explorative Study on Extractive Text Summarization through k-means, LSA, and TextRank. *WiSPNET 2023 - International Conference on Wireless Communications, Signal Processing and Networking*. https://doi.org/10.1109/WISPNET57748.2023.10134303

Ray, A. T., Fischer, O. P., White, R., Fischer, O. J., Mavris, D. N., White, R. T., & Cole, B. F. (2023). *aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT Data-driven Aviation Safety Enhancement using Machine Learning View project Advanced Vehicles View project aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT*. https://doi.org/10.2514/6.2023-2583

Reddy, K. M., & Guha, R. (2023). Automatic Text Summarization for Conversational Chatbot. *2023 IEEE 8th International Conference for Convergence in Technology,*

I2CT 2023. https://doi.org/10.1109/I2CT57861.2023.10126161

Sabuna, P. M., & Setyohadi, D. B. (2018). Summarizing Indonesian text automatically by using sentence scoring and decision tree. *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, *2018-January*, 1–6. https://doi.org/10.1109/ICITISEE.2017.8285473

Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2022). A multi-objective memetic algorithm for query-oriented text summarization: Medicine texts as a case study. *Expert Systems with Applications*, *198*. https://doi.org/10.1016/J.ESWA.2022.116769

Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: The rule-based POS tagger for Uzbek language. *ArXiv Preprint ArXiv:2301.12711*. https://arxiv.org/abs/2301.12711v2

Sharma, G., Gupta, S., & Sharma, D. (2022). Extractive Text Summarization Using Feature-Based Unsupervised RBM Method. *Lecture Notes in Networks and Systems*, *370*, 105–115. https://doi.org/10.1007/978-981-16-8664-1_10/COVER

Singh, R. K., Khetarpaul, S., Gorantla, R., & Allada, S. G. (2021). SHEG: summarization and headline generation of news articles using deep learning. *Neural Computing and Applications*, *33*(8), 3251–3265. https://doi.org/10.1007/s00521-020-05188-9

Taieb-MaimMeiravon, Romanovski-Chernik, A., Last, M., Litvak, M., & Elhadad, M. (2023). Mining Eye-Tracking Data for Text Summarization. *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2023.2227827

Tubau Sala, E., & Alonso Cánovas, D. (2002). Inferencias bayesianas: una revisión teórica. *Anuario de Psicología, ISSN 0066-5126, Vol. 33, Nº. 1, 2002, Págs. 25-48*, *33*(1), 25–48. https://dialnet.unirioja.es/servlet/articulo?codigo=259815&info=resumen&idioma=SPA

Verma, P., & Om, H. (2019). MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Systems with Applications*, *120*, 43–56. https://doi.org/10.1016/j.eswa.2018.11.022

Verma, S., & Nidhi, V. (2018). Extractive Summarization using Deep Learning. *Research in Computing Science*, *147*(10), 107–117. https://doi.org/10.13053/rcs-147-10-9

Wang, Z., Wu, Y., Lei, P., Jain, R., Sharma, A., Sankar Mishra, G., Nand, P., & Chakraborty, S. (2020). You may also like Named entity recognition in steel field based on BiLSTM-CRF model Zhai Chenhao and Wang Chengyao-Named Entity Recognition Method of Brazilian Legal Text based on pre-training model Named Entity Recognition in English Text. *Journal of Physics: Conference Series*, *1712*, 12013. https://doi.org/10.1088/1742-6596/1712/1/012013

Waseemullah, Fatima, Z., Zardari, S., Fahim, M., Siddiqui, M. A., Ibrahim, A. A. A., Nisar, K., & Naz, L. F. (2022). A Novel Approach for Semantic Extractive Text Summarization. *Applied Sciences 2022, Vol. 12, Page 4479*, *12*(9), 4479.

https://doi.org/10.3390/APP12094479

Xiong, Y., Yan, M., Hu, X., Ren, C., & Tian, H. (2023). An unsupervised opinion summarization model fused joint attention and dictionary learning. *Journal of Supercomputing*, *79*(16), 17759–17783. https://doi.org/10.1007/S11227-023-05316-X/METRICS

Yadav, A. K., Ranvijay, Yadav, R. S., & Maurya, A. K. (2023). State-of-the-art approach to extractive text summarization: a comprehensive review. *Multimedia Tools and Applications 2023*, 1–63. https://doi.org/10.1007/S11042-023-14613-9

Yohannes, H. M., & Amagasa, T. (2022). Named-entity recognition for a low-resource language using pre-trained language model. *Proceedings of the ACM Symposium on Applied Computing*, 837–844. https://doi.org/10.1145/3477314.3507066

Anexos

Anexo 1. Lista de Stopwords

El siguiente listado muestra las Stopwords usadas en este trabajo para la GAR en inglés

i	its	being	by	under	nor
me	itself	have	for	again	not
my	they	has	with	further	only
myself	them	had	about	then	own
we	their	having	against	once	same
our	theirs	do	between	here	so
ours	themselves	does	into	there	than
ourselves	what	did	through	when	too
you	which	doing	during	where	very
your	who	a	before	why	s
yours	whom	an	after	how	t
yourself	this	the	above	all	can
yourselves	that	and	below	any	will
he	these	but	to	both	just
him	those	if	from	each	don
his	am	or	up	few	should
himself	is	because	down	more	now
she	are	as	in	most	
her	was	until	out	other	
hers	were	while	on	some	
herself	be	of	off	such	
it	been	at	over	no	

Anexo 2. Lista de etiquetas POS y NER

En las siguientes tablas se muestran el listado de etiquetas POS y NER.

Tabla 8. Lista de etiquetas POS.

Etiqueta	Nombre	Etiqueta	Nombre	Etiqueta	Nombre
CC	Conjunción de coordinación	NNS	Sustantivo plural	TO	Palabra "to"
CD	Número cardinal	NNP	Sustantivo propio singular	UH	Interjección
DT	Determinante	NNPS	Sustantivo propio plural	VB	Verbo en forma base
EX	Existencial "there"	PDT	Predeterminante	VBD	Verbo en pasado
FW	Palabra extranjera	POS	Terminación posesiva	VBG	Verbo gerundio o participio presente
IN	Preposición o conjunción subordinante	PRP	Pronombre personal	VBN	Verbo participio pasado
JJ	Adjetivo	PRP$	Pronombre posesivo	VBP	Verbo presente en 3ra persona singular
JJR	Adjetivo, comparativo	RB	Adverbio	VBZ	Verbo en 3ra persona singular presente
JJS	Adjetivo, superlativo	RBR	Adverbio comparativo	WDT	Determinante Wh
LS	Marcador de elemento de lista	RBS	Adverbio superlativo	WP	Pronombre Wh
MD	Modal	RP	Partícula adverbial	WP$	Pronombre posesivo wh
NN	Sustantivo singular o masa	SYM	Símbolo	WRB	Wh-adverbio

Tabla 9. Lista de etiquetas NER.

Etiqueta	Nombre	Etiqueta	Nombre
PERSON	Personas (incluyendo nombres de ficción)	LAW	Documentos relevantes convertidos en leyes
NORP	Nacionalidades, grupos religiosos o políticos	LANGUAGE	Cualquier idioma con nombre
FACILITY	Edificios, aeropuertos, carreteras, puentes, etc.	DATE	Fechas o periodos absolutos o relativos
ORGANIZATION	Empresas, agencias, instituciones, etc.	TIME	Tiempos más pequeños que un día
GPE	Países, ciudades, estados	PERCENT	Porcentaje (incluido "%")
LOCATION	Ubicaciones no GPE, cadenas montañosas, masas de agua	MONEY	Valores monetarios, incluida la unidad
PRODUCT	Vehículos, armas, alimentos, etc. (No servicios)	QUANTITY	Medidas como el peso o la distancia
EVENT	Nombres de huracanes, batallas, guerras, eventos deportivos, etc.	ORDINAL	Nombres de números ordinales (p. ej. "primero", "segundo", "tercero")
WORK OF ART	Títulos de libros, canciones, etc.	CARDINAL	Números que no pertenecen a otro tipo