Capítulo 1

Enfoque teórico para la sumarización de artículos científicos mediante modelos largos de lenguaje

Giovanna Inosuli Campos Flores[1]
Yavé Emmanuel Vargas Márquez[2]

https://doi.org/10.61728/AE20253714

[1] Instituto Politécnico Nacional.
[2] Instituto Politécnico Nacional.

Definición del problema

En el ámbito académico y científico, la capacidad de acceder y procesar rápidamente información relevante de artículos científicos es crucial para investigadores, académicos y estudiantes. La constante generación de conocimiento en diversas disciplinas, sumada al creciente volumen de publicaciones científicas, plantea un desafío significativo para aquellos que buscan mantenerse actualizados en sus campos. El trabajo de Larsen y Von In (Larsen y Von Ins, 2010) demuestra que el número de publicaciones científicas ha experimentado un crecimiento constante, con tasas anuales cercanas al 4.7 %. Este aumento, aunque desigual entre disciplinas, refleja la necesidad urgente de herramientas que permitan a los investigadores manejar eficazmente el flujo de información. El fenómeno, conocido como "sobrecarga de información", dificulta la identificación de contenido relevante y preciso, especialmente en áreas donde el avance científico ocurre a un ritmo acelerado.

La lectura y comprensión detallada de los artículos científicos puede ser una tarea tediosa y demandante en términos de tiempo, debido a su léxico especializado, estructuras gramaticales complejas y variedad de temas abordados. Por su parte, estudios como los de Luis Parra y Fernando Álvarez (Parra-Medina, 2021) destacan la importancia de filtrar contenido relevante en un mar de información que puede resultar abrumador, especialmente para quienes carecen de experiencia en la evaluación de fuentes académicas.

En este contexto, la automatización del proceso de sumarización se presenta como una solución prometedora para condensar textos extensos en versiones más comprensibles y accesibles, manteniendo los hallazgos clave y eliminando información irrelevante. La sumarización automática no solo ayuda a los investigadores a ahorrar tiempo, sino que también optimiza el acceso al conocimiento al destacar los puntos más relevantes de los textos originales.

Como señala el Ing. Uriel Ramírez (2020) en su tesis, la generación automática de resúmenes puede mitigar los efectos negativos de la sobrecarga de información, permitiendo a los usuarios transformar datos en conocimiento de manera eficiente. Este proceso, además de ser útil para la

navegación y selección de artículos, desempeña un rol crucial como paso intermedio en otras tareas de PLN (Procesamiento de Lenguaje Natural), como la resolución de referencias, la eliminación de redundancias y la simplificación de oraciones.

Sin embargo, la sumarización de artículos científicos enfrenta desafíos significativos. Los modelos actuales, aunque efectivos para textos generales, tienen limitaciones al procesar documentos académicos debido a su naturaleza especializada. La jerga técnica, las referencias cruzadas y las estructuras argumentativas complejas pueden dificultar la identificación de información relevante y la generación de resúmenes coherentes. Además, los sistemas de sumarización deben lidiar con la variabilidad en el formato y estilo de los artículos, lo que complica el desarrollo de enfoques universales.

Además, autores como Eugenio López y colaboradores (E. J. L. Gómez, 2011) señalan que la terminología técnica y la ambigüedad contextual en los artículos científicos representan retos adicionales para los modelos de lenguaje. Esto subraya la necesidad de avanzar en el desarrollo de sistemas de sumarización que no solo sean precisos y eficientes, sino que también puedan adaptarse a las particularidades del lenguaje científico.

La automatización de la sumarización de artículos científicos es un área de gran relevancia y potencial impacto en el ámbito académico. Aunque existen retos técnicos y conceptuales que deben abordarse, los avances en Modelos Largos de Lenguaje preentrenados ofrecen una base sólida para enfrentar este desafío y proporcionar soluciones efectivas que optimicen el acceso a la información científica.

Modelos largos de lenguaje preentrenados

La sumarización es una tarea propia del campo del PLN o NLP (por en sus siglas en inglés, Natural Language Processing), que es un campo de la inteligencia artificial que investiga la comunicación entre personas y computadoras utilizando lenguaje natural que es aquel que utilizan los humanos para comunicarse, ya sea en forma escrita u oral (Talamé, Cardoso y Amor, 2019).

Generalmente, en el PLN se realizan cambios en los textos para analizarlos de una manera más sencilla, conocida como procesamiento de datos, que consiste en aplicar diversas técnicas clave, las cuales ayudan en la limpieza de los datos, tales como:

- Tokenización: Divide el texto en unidades mínimas llamadas "tokens", como palabras o caracteres. Por ejemplo, en el texto "Los alumnos estudian mucho", los tokens son: {"Los", "alumnos", "estudian", "mucho"} (Vajjala, Majumder, Gupta y Surana, 2020).
- Segmentación: Divide el texto en partes como párrafos u oraciones, lo que facilita su análisis y procesamiento. Esta técnica es esencial para evitar ambigüedades causadas por puntuación o abreviaturas (Ma, 2018).
- Normalización: Estandariza el texto convirtiendo caracteres a minúsculas, eliminando duplicados y expandiendo abreviaturas para optimizar el rendimiento de los modelos (HaCohen-Kerner, Miller y Yigal, 2020).
- Lematización y stemming: Transforman palabras a sus formas base, reduciendo la dimensionalidad y mejorando tareas como recuperación de información (Nasimba Tipan, 2023).

Los avances significativos que obtuvo PNL fueron gracias a que, durante los años recientes, se ha desarrollado algoritmos como el transformer, que es una arquitectura de red neuronal concebida en 2017 por investigadores de Google en el artículo "Attention Is All You Need", la cual es una arquitectura que se fundamenta en un mecanismo de atención destinado a la detección de dependencias entre las entidades (Briones Viqueira, 2023).

El transformer fue planteado como un método de atención para enlazar un codificador-decodificador de una RNR (Red Neuronal Recurrente) o una RNC (Red Neuronal Convolucional), pero su ventaja es evitar las recurrencias y convoluciones propias de los RNR o RNC, de esta manera, buscando un mejor desempeño y rendimiento mediante el mecanismo de atención (Beltrán Beltrán y Rodríguez Mojica, 2021).

A partir del transformer se crearon diversos modelos preentrenados que aprovechan sus componentes centrales: el encoder y el decoder. Al-

gunos de los modelos preentrenados también se pueden llamar MLL o en inglés LLM (Large Language Models), ya que estos modelos requieren cantidades enormes de datos (en gigabytes o terabytes) y parámetros (de decenas a cientos de miles de millones) y sobresalen en el PLN (Brown y otros, 2020).

Para crear estos MLL, son esenciales varias fases, empezando por el preprocesamiento de los datos, ya que mejora la calidad y consistencia de los datos, facilitando su uso en tareas complejas. Estas tareas son las mencionadas para la limpieza de datos, y se utiliza un conjunto de técnicas según el desarrollo lo requiera (HaCohen-Kerner, Miller y Yigal, 2020).

Luego se realiza una fase de entrenamiento, también conocida como preentrenamiento donde el modelo aprende utilizando el enorme corpus de datos de texto con el objetivo que el modelo desarrolle una comprensión general del lenguaje, lo que incluye gramática, semántica y patrones contextuales, con ello se desarrolla una base sólida al modelo antes de ser adaptado para tareas específicas (Gururangan y otros, 2020).

Finalmente, si se requiere, se aplica una fase de ajuste fino (o fine-tuning en inglés), que es un proceso en el cual un modelo previamente entrenado se adapta a tareas específicas utilizando datos relacionados con dichas tareas. Existen varias técnicas para realizar este ajuste fino. Una de ellas es el *transfer learning*, donde los modelos preentrenados, que ya tienen un rendimiento aceptable para tareas generales, se afinan con datos específicos de una tarea para mejorar su rendimiento en esa área (Gururangan y otros, 2020).

Otra estrategia es la instruction-tuning, donde el modelo se ajusta con datos formateados como pares de instrucciones y entradas/salidas. Este enfoque mejora la capacidad del modelo para generalizar de manera efectiva a nuevos problemas y mejora su desempeño en tareas posteriores (Peng, Li, He, Galley y Gao, 2023).

El alignment-tuning busca alinear el modelo con los valores humanos para evitar la generación de textos falsos, sesgados o dañinos. Se utiliza retroalimentación humana para ajustar los parámetros y garantizar respuestas útiles, honestas y seguras. Esto se logra mediante RLHF (reinforcement learning with human feedback), donde el modelo se entrena inicialmente con demostraciones y luego se ajusta usando aprendizajes

por recompensas (reward modeling) y reforzamiento de aprendizaje (reinforcement learning) (Naveed y otros, 2024).

Los MLL representan un avance significativo en el campo del NLP, gracias a su capacidad para comprender y generar texto de manera altamente coherente. Estos modelos, entrenados en vastas cantidades de datos textuales provenientes de diversas fuentes como libros, artículos científicos y páginas web, son capaces de capturar patrones lingüísticos, contextos semánticos y estructuras gramaticales complejas (Wang, Li, Wu, Hovy y Sun, 2023).

Los modelos que destacan por abordar problemáticas complejas, ofrecer soluciones eficientes y establecer estándares que inspiran futuras investigaciones. A continuación, se presenta una descripción de los modelos más populares o conocidos:

- GPT: Es un modelo generativo que utiliza una arquitectura de transformer para procesar el texto de manera secuencial, desarrollado por OpenAI. Es ampliamente utilizado para tareas como la generación de texto, respuestas a preguntas, traducción automática (Chintagunta, Katariya, Amatriain y Kannan, 2021).
- BERT: Es un modelo de lenguaje bidireccional que lee el texto en ambas direcciones (de izquierda a derecha y de derecha a izquierda) para entender el contexto en profundidad, desarrollado por Google (Devlin y Chang, 2018).

A continuación, se muestra en la Tabla 1 las características principales de los modelos de GPT y BERT:

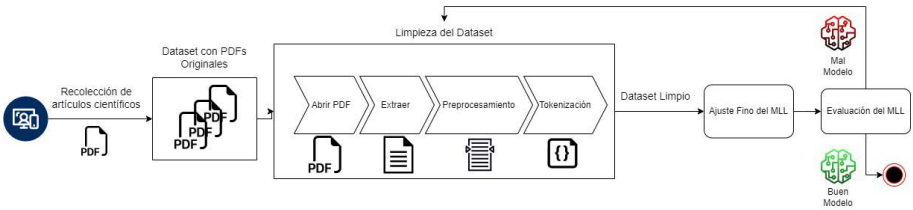
Tabla 1. Característica de Modelos GPT y BERT

Característica	GPT	BERT
Enfoque	Generativo (crear texto nuevo)	Extractivo (identifica texto existente)
Direccionalidad	Unidireccional	Bidireccional
Adaptabilidad	Alta	Medios (para generación, requiere más ajustes)
Uso común	Resumen abstractivo	Resumen extractivo

Metodología teórica para el ajuste fino

En esta parte se expondrá como lograr el ajuste fino de un modelo largo de lenguaje, en específico GPT-3, para la tarea de sumarización de artículos científicos, para reducir la complejidad estos artículos serán en español. A continuación, en la Ilustración 1 se describe el flujo teórico propuesto:

Ilustración 1. Diagrama teórico



Preparación del dataset

La calidad del dataset es fundamental para entrenar un modelo, ya que los datos de entrada inadecuados pueden llevar a resultados erróneos. Por ello, es crucial garantizar que los datos sean adecuados para cumplir con los objetivos deseados del modelo. La limpieza de datos es importante porque permite eliminar valores inconsistentes, ausentes o irrelevantes que podrían afectar el rendimiento del modelo. Un dataset limpio asegura que el modelo reciba información precisa, relevante y libre de ruido, lo que mejora tanto la precisión como la generalización del sistema.

El concepto GIGO (Garbage In, Garbage Out) ilustra que, si se ingresan datos incorrectos de baja calidad en un sistema, los resultados obtenidos serán igualmente incorrectos o inútiles. Por tanto, la calidad de los datos iniciales es esencial para garantizar la validez de los resultados:

1. Recolección de datos:

- Obtener artículos científicos en español que incluyan texto completo, texto extraíble y no en imágenes, resumen, palabras clave y contenido. Para la delimitación en la obtención de artículos, se puede tomar un

área en específico (como, por ejemplo, el área de inteligencia artificial, que llega a ser multidisciplinaria).
- Seleccionar fuentes confiables, como revistas indexadas, buscadores académicos, repositorios de acceso abierto (e.g., SciELO, Redalyc) y bases de datos académicas.

2. Limpieza de datos:

- Eliminar partes innecesarias como lo son metadatos, encabezados, laterales, pies de página, elementos no textuales (como imágenes, figuras, diagramas, tablas), notas al pie de página y referencias. Así mismo se eliminarían sus menciones y contenido relacionado (como los títulos de las imágenes).
- Normalizar el formato de los artículos para asegurar la uniformidad de los textos (por ejemplo, convertir todo a minúsculas).
- Preprocesar los textos mediante tokenización, eliminación de caracteres especiales y manejo de términos específicos.

3. Generación de conjuntos de datos:

- Del dataset limpio se dividen los datos en conjuntos de entrenamiento y prueba.
- Asegurando que cada artículo esté acompañado de un resumen de referencia del artículo, este mismo puede ser creado por un humano, pero en este caso se usará el resumen incluido en cada artículo para la cuestión del tiempo.

Selección de Hiperparámetros para el entrenamiento

Los hiperparámetros determinan cómo se ajustará el modelo durante el entrenamiento. Los principales a considerar son:

1. Tasa de aprendizaje (learning rate):

- Esta se establece con un valor inicial que permita un ajuste gradual sin provocar oscilaciones excesivas y se podrá modificar en posteriores

entrenamientos según los resultados obtenidos.

2. Tamaño de lote (batch size):

- Determinar el tamaño óptimo con base en los recursos computacionales disponibles donde se ajusta según se vea necesario.

3. Número de épocas:

- Definir un número suficiente para lograr la convergencia sin sobreajustar o subajustar el modelo.

Métricas de evaluación

Evaluar la calidad de los resúmenes generados es crucial para medir el éxito del ajuste fino. Las métricas más comunes incluyen:

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Que es un conjunto de métricas ampliamente utilizadas para evaluar la calidad de resúmenes generados automáticamente, comparándolos con resúmenes de referencia realizados por humanos. Estas métricas calculan la coincidencia entre n-gramas, palabras o frases entre el resumen generado y el de referencia, proporcionando una medida de similitud que se interpreta como una aproximación a la calidad del resumen (Ng y Abrecht, 2015). A continuación, se mencionan algunas de sus métricas:

- ROUGE-N (e.g., ROUGE-1, ROUGE-2) mide coincidencias de unigramas y bigramas, es decir, compara palabras individuales (unigramas) y pares consecutivos de palabras (bigramas) entre el texto generado y el de referencia.
- ROUGE-L evalúa la subsecuencia común más larga entre el texto generado y el de referencia, es decir, la mayor secuencia de palabras que aparece en ambos textos en el mismo orden, aunque no necesariamente de forma continua.

Retos del ajuste fino

El ajuste fino de modelos largos de lenguaje como el modelo GPT-3 puede ser extremadamente útil, pero también presenta varios desafíos que deben abordarse para obtener resultados óptimos y minimizar problemas técnicos. Estos desafíos incluyen:

1. Recursos Computacionales:

- El ajuste fino requiere un hardware avanzado, como GPUs o TPUs, para poder manejar el tamaño y la complejidad computacional del modelo.
- El tiempo de entrenamiento puede ser extenso, además de requerir un uso considerable de recursos de las computadoras y un consumo significativo de energía.

2. Calidad de los datos:

- En el ámbito científico, los datos suelen estar dispersos en múltiples fuentes, con diferentes formatos y estilos de redacción. Actualmente, no existe un dataset estandarizado que facilite el entrenamiento para la sumarización de artículos científicos. Esto dificulta la recopilación de un corpus homogéneo y representativo, afectando la capacidad del modelo para aprender patrones generalizables. Crear un dataset desde cero requiere tiempo, esfuerzo y recursos, ya que implica recopilar, limpiar, estructurar y anotar una gran cantidad de textos científicos.
- La mayoría de los resúmenes disponibles para evaluar de sumarización están en inglés; sin embargo, en español, las opciones son limitadas, y los resúmenes disponibles suelen ser los propios abstracts de los artículos. Esto puede restringir al modelo a aprender a generar resúmenes que simplemente repliquen los abstracts. Un abstract es un resumen breve, conciso y claro de un documento académico o científico que generalmente se presenta en idioma inglés. O anotar manualmente resúmenes en español requiere expertos en diferentes áreas científicas, lo cual es costoso y difícil de escalar.

- Los datos recopilados pueden contener sesgos inherentes (por ejemplo, hacia ciertos campos de estudio, autores o regiones) o errores (gramaticales, tipográficos, incompletos). Los sesgos en los datos pueden influir negativamente en las respuestas del modelo, y los errores pueden afectar la calidad y precisión de las predicciones, Garantizar que los datos sean representativos y libres de sesgos requiere una revisión exhaustiva, y la identificación de errores puede ser compleja si el corpus es grande.

Esta metodología teórica proporciona un marco inicial para guiar el ajuste fino del modelo, destacando los pasos clave, las decisiones importantes y los desafíos inherentes al proceso.

Conclusiones

En el ámbito académico y científico, la creciente cantidad de literatura hace imprescindible contar con herramientas que ayuden a optimizar el tiempo y los recursos. Los MLL como GPT o BERT, al especializarlos para la tarea de sumarización, pueden ofrecer resúmenes claros y concisos que destacan los puntos clave de los documentos. Esto generaría un impacto significativo, especialmente para investigadores, académicos y estudiantes que necesitan procesar grandes volúmenes de información.

Un modelo de sumarización automatizado simplifica el análisis de textos científicos, permitiendo identificar rápidamente información relevante. Por ejemplo, en investigaciones multidisciplinarias, estos modelos ayudan a los usuarios a comprender conceptos clave incluso fuera de su especialidad.

Además, la accesibilidad es un beneficio importante para los investigadores con limitaciones de tiempo, debido a que los resúmenes generados automáticamente ofrecen una solución valiosa y una visión general del contenido, para que se pueda saber rápidamente si el trabajo les resulta relevante, facilitando que más usuarios aprovechen la información y su tiempo.

Para concluir, los MLL para la sumarización de artículos científicos no solo apoyan la comprensión y el análisis, sino que también transforman

la forma en que interactuamos con el conocimiento, haciéndolo más accesible, eficiente y relevante para diversos usuarios.

Referencias

Beltrán Beltrán, N. C., & Rodríguez Mojica, E. C. (2021). Procesamiento del lenguaje natural (PLN) - GPT-3.: Aplicación en la Ingeniería de Software. *Tecnol. Investig. Academia TIA*, *8*(1), 18-37.

Briones Viqueira, J. (2023). *Construcción de un modelo de predicción para activos financieros. Creación de una red neuronal recurrente LSTM con machine learning*. Universidade da Coruña. Facultade de Economía e Empresa. doi:http://hdl.handle.net/2183/36522

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Child, R. (2020). Language Models are Few-Shot Learners. arXiv. doi:10.48550/ARXIV.2005.14165

Chintagunta, B., Katariya, N., Amatriain, X., & Kannan, A. (2021). Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization. *Machine Learning for Healthcare Conference*, 354-372.

Devlin, J., & Chang, M.-W. (2 de Noviembre de 2018). Open Sourcing BERT: State-of-the-Art Pretraining for Natural Language Processing. (Google AI Blog) Recuperado el Diciembre de 2024, de https://goo.su/DMgmy

E. J. L. Gómez, M. E. (2011). Fenómenos lingüísticos frecuentes que dificultan la calidad de los textos científicos escritos por profesionales. *Edumecentro*, *3*(1), 73–81.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Iz, B., Downey, D., & Smith, N. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. arXiv. doi:arXiv:2004.10964

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). *The influence of preprocessing on text classification using a bag-of-words representation*. PLOS ONE, 15. doi:10.1371/journal.pone.0232525

Larsen, P., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*, 575-603. doi:http://dx.doi.org/10.1007/s11192-010-0202-z

Ma, L. (2018). *Segmentación automática de palabras en chino*. Universidad Complutense De Madrid Facultad De Filología.

Nasimba Tipan, A. (2023). "Attention is all you need". *Arquitectura Transformers: descripción y aplicaciones*. Universidad Miguel Hernández De Elche.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., . . . Mian, A. (2024). *A Comprehensive Overview of Large Language Models*. arXiv. doi:10.48550/arxiv.2307.06435

Ng, J.-P., & Abrecht, V. (2015). *Better summarization evaluation with word embeddings for ROUGE*. arXiv [cs.CL]. Obtenido de http://arxiv.org/abs/1508.06034https://arxiv.org/abs/1508.06034

Parra-Medina, L. E.-C. (2021). Síndrome de la sobrecarga informativa: una revisión bibliográfica. *Rev Neurol*, *73*(12), 421-428.

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction Tuning with GPT-4. arXiv. doi:10.48550/ARXIV.2304.03277

Ramírez , U. B. (2020). *Método para la generación automática de resúmenes personalizados de textos en español*. Tecnológico Nacional de México.

Talamé, L., Cardoso, A. C., & Amor, M. (2019). *Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python*.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media.

Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2023). Pre-Trained Language Models and Their Applications. *Engineering*, *25*, 51-65. doi:10.1016/j.eng.2022.04.024