Editorial

Evolución del procesamiento natural del lenguaje

Andres Felipe Giraldo Forero[1], Andrés Felipe Orozco Duque[2]

[1] Instituto Tecnológico Metropolitano, Medellín - Colombia, felipegiraldo@itm.edu.co
[2] Universidad de Medellín, Medellín - Colombia, aforozco@udemedellin.edu.co

El lenguaje escrito no ha sido ajeno a los cambios culturales y tecnológicos. Por ejemplo, la invención de la imprenta en el siglo XV, o el desarrollo de los computadores personales y los teléfonos inteligentes en las últimas décadas, fueron hitos que marcaron drásticamente su evolución. De manera paralela al desarrollo del lenguaje, la evolución de la computación y el surgimiento de algoritmos inteligentes, capaces de tomar decisiones han tenido avances impresionantes en los últimos años formando una nueva área de estudio conocida como inteligencia artificial IA, considerada por muchos como la siguiente gran revolución. La integración de las técnicas de IA con el fin de interpretar, manipular y comprender el lenguaje humano da como resultado la rama del procesamiento natural del lenguaje PNL, la cual, teniendo en cuenta los últimos avances en modelos de generación de lenguaje, se puede convertir en el próximo gran hito en cuanto al lenguaje escrito. Los primeros algoritmos de PNL estaban basados en reglas, posteriormente se empezó a utilizar esquemas de clasificación supervisada basados en modelos como la regresión logística, las máquinas de soporte vectorial, los modelos ocultos de Markov, o los árboles aleatorios condicionales, entre otros. El problema de los modelos de clasificación tradicionales es que los mismos están orientados a evaluar cada palabra, o las relaciones entre cada palabra y la palabra anterior, pero no capturan el contexto de las mismas en una frase completa.

Para solucionar este problema, se han desarrollado avances importantes en los últimos 10 años, de los cuales podemos destacar 3 momentos claves: i) El surgimiento de los modelos word embeddings [1] en el 2013, los cuales permiten representar vectorialmente palabras, pero teniendo en cuenta el contexto, de tal forma que palabras sinónimas o relacionadas son representados por vectores con una distancia menor en el espacio vectorial que palabras no relacionadas entre sí. ii) La arquitectura de redes neuronales profundas llamada transformers [2] en el 2017. Esta arquitectura contiene una matriz de atención que evalúa la relación de cada palabra con todas las demás palabras de la sentencia y por lo tanto tiene en cuenta todo el contexto y permite captar el sentido completo de un texto. Adicionalmente, los modelos basados en transformers permiten la implementación de esquemas de transferencia de aprendizaje, de tal forma que se pueden pre-entrenar modelos en grandes cantidades de texto utilizando esquemas auto supervisados (por ejemplo, prediciendo la palabra siguiente en una oración), para después realizar un ajuste fino de parámetros y adaptar el modelo a aplicaciones específicas en las cuales es difícil tener grandes cantidades de datos. iii) El surgimiento de los modelos generativos de lenguaje multipropósito como ChatGPT [3], Bard y LLaMa [4] presentados a finales del 2022 y principios del 2023. Estos modelos tienen una arquitectura con varias capas de transformers lo cual implica que tienen cientos de millones de parámetros que deben ajustarse en el entrenamiento, por lo cual, son modelos que requieren corpus con miles de millones de palabras para su entrenamiento. Estos últimos modelos de lenguaje han marcado un hito importante en la IA y está generando un sinfín de discusiones respecto a su impacto a todo nivel: cultural, educativo, social, entre otros, ya que

sus capacidades en cuanto a la generación de texto son comparables a un humano. Por ejemplo, a continuación, presentamos parte de la respuesta de ChatGPT ante la siguiente entrada: Escribir una editorial sobre procesamiento de lenguaje natural para una revista indexada del área de ingeniería.

"El procesamiento de lenguaje natural (PLN) es un campo interdisciplinario que combina la lingüística, la informática y la inteligencia artificial para hacer posible que las computadoras comprendan, procesen y generen lenguaje humano. En los últimos años, ha habido un gran avance en el PLN gracias a los avances en el aprendizaje profundo y el big data, lo que ha permitido a los investigadores mejorar significativamente la capacidad de las computadoras para procesar el lenguaje natural de manera efectiva…Sin embargo, el PLN todavía tiene muchos desafíos por superar. Por un lado, el lenguaje humano es muy complejo y puede ser ambiguo, lo que hace que sea difícil para las computadoras entenderlo completamente. Además, las variaciones regionales, culturales y generacionales en el lenguaje pueden dificultar aún más el procesamiento del lenguaje natural…"

En general, ChatGPT genera un resumen muy concreto de lo que es el PLN. Es de anotar que chatGPT no busca la información en una base de datos, sino que es un modelo generador de lenguaje y la respuesta la genera a partir de la pregunta y basado en el entrenamiento previo (chatGPT está basado en el modelo GPT-3.5 con aproximadamente 175000 millones de parámetros).

De la respuesta de ChatGPT podemos anotar que contiene datos generales, que dan una visión amplia del tema, pero que no es lo suficientemente concreto como para dar información específica que me permita determinar los avances más importantes en PLN y los retos actuales que sirvan como base a nuevas investigaciones. Sin embargo, podemos seguir interactuando con preguntas más específicas, por ejemplo: ¿Cuáles son los principales avances, retos y brechas del PLN en la actualidad? El lector podrá revisarlo y seguramente obtendrá una lista con cada uno de estos ítems que refleja en términos generales un muy buen resumen de estos puntos.

La pregunta que surge entonces sería la siguiente: ¿Puede ChatGPT reemplazarnos como autores de esta editorial? Nosotros creemos que no. Podríamos decir que ChatGPT sería el equivalente a una persona con muchísima información general en su cerebro, pero siempre será importante la visión del experto temático. Además de la incapacidad de asumir responsabilidad, motivo por el cual las principales editoriales han prohibido que este tipo de IA tengan el rol de coautor de artículos científicos. Por otro lado, es importante indicar que estos modelos multipropósito requieren ser ajustados para trabajar en aplicaciones especializadas. Por ejemplo, el desarrollo de aplicaciones de generación de lenguaje en el área de la salud, se convierten en un campo abierto de investigación ya que modelos multipropósito como ChatGPT pueden generar información poco confiable o datos erróneos que pueden tener consecuencias graves si son utilizados para tomar decisiones clínicas.

Por lo anterior, podemos decir que, para generar soluciones a tareas específicas, aunque se puede utilizar como base los modelos de generación de lenguaje pre-entrenados como GPT o BERT, se deben realizar ajustes finos del modelo con bases de datos propias y realizar las validaciones necesarias para asegurar el desempeño correcto del modelo. Y he aquí uno de los grandes retos que pueden ser abordados desde los grupos de investigación y las pequeñas y medianas empresas que posiblemente no tienen los recursos de las grandes compañías que compiten en la construcción de los nuevos modelos de lenguaje: la creación de bases de datos específicas, con datos suficientes para el entrenamiento o ajuste de los modelos, que sean etiquetadas por varios especialistas en el área de la aplicación y que estén curadas o validadas para evitar sesgos. En la medida en que se generen nuevas bases de datos de calidad, se podrán ver avances en aplicaciones especializadas adaptadas a diferentes contextos, por

ejemplo, a jergas especializadas, o estilos culturales de redacción, que permitan mejorar el desempeño de los modelos en la práctica. En general, implementar modelos de deep learning para NLP desde el punto de vista netamente técnico, puede no ser una tarea tan compleja ya que se dispone de una gran cantidad de herramientas como: Brat o INCEpTION unas interfaces que permite realizar etiquetado de texto, librerías de software como NLTK, Spacy, Stanford CoreNLP, plataformas como Hugging Face que disponen de modelos pre-entrenados, o las API de modelos como GPT4 [3]. El reto está en validar correctamente dichos modelos para evitar el sobreajuste a los datos de entrenamiento, evitar los sesgos, asegurar el poder de generalización del modelo ante datos nuevos, ajustar los modelos a tareas específicas y necesidades puntuales. Por lo anterior invitamos a los investigadores a trabajar en la construcción de nuevas bases de datos y al desarrollo de los esquemas de pre-procesamiento y ajuste, y a compartir estos desarrollos bajo un esquema de ciencia abierta que permita construir soluciones a diferentes problemas específicos a nivel regional y global.

REFERENCIAS

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv*, Oct. 2013, Accessed: Mar. 26, 2023. https://doi.org/10.48550/arXiv.1310.4546

[2] A. Vaswani *et al.*, "Attention Is All You Need," *Neural Information Processing Systems*, Long Beach, CA, USA, Jun. 2017, Accessed: Mar. 26, 2023. https://doi.org/10.48550/arXiv.1706.03762

[3] OpenAI, "GPT-4 Technical Report," Mar. 2023. http://arxiv.org/abs/2303.08774

[4] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv,* Feb. 2023. http://arxiv.org/abs/2302.13971