ALEJANDRO MAURO[1]
CLÍNICA ALEMANA DE SANTIAGO

“¿La inteligencia artificial va a reemplazar a los profesionales de la salud que no usan inteligencia artificial?”

Soy Alejandro Mauro y quiero contarles sobre cómo implementamos modelos de lenguaje en una historia clínica electrónica, en la Clínica Alemana de Santiago, de la que soy jefe de transformación digital.

La medicina tiene una gran asimetría de conocimiento, donde el médico puede decirle algo al paciente y este último, con menos herramientas para evaluarlo, acepta. La informática tiene una condición de asimetría similar. Entonces, la combinación de informática en salud genera capacidades de vender cosas que no son lo que parecen.

Mi idea es contarles nuestra experiencia en grandes modelos de lenguaje y desmitificar la frase: *"la inteligencia artificial va a reemplazar a los profesionales de la salud que no usan Inteligencia Artificial"*. Cualquier tecnología puede ser vendida desde el paradigma del miedo, pero a mí no me gusta porque genera temor de una tecnología que es súper útil. La inteligencia artificial (IA) es espectacular, nos deja subir una montaña, pero la venden como si fuéramos a llegar a la luna. La tecnología no permite eso que se vende.

EL ABC PARA ENTENDER LOS MODELOS DE LENGUAJE

Partamos de cómo una computadora puede hacer lo que queremos. Tenemos dos formas: la programación tradicional, con algoritmos simples donde un humano escribe qué hacer

[1] El Dr. Alejandro Mauro es un médico especializado en Informática Médica, con más de 15 años promoviendo la innovación en salud. Como Jefe de Transformación Digital en la Clínica Alemana de Santiago, ha liderado la implementación de más de 20 algoritmos de inteligencia artificial, situando a la institución a la vanguardia de la medicina digital. Comprometido con la formación, colabora activamente con el Instituto de Ciencias e Innovación en Medicina de la Universidad del Desarrollo. En 2022, fue galardonado con el SNOMED International Award for Excellence por su aporte en el desarrollo de terminologías médicas para Chile y Uruguay que permitieron la implementación de la Receta Electrónica Nacional en el contexto de pandemia. Su capacidad y liderazgo siguen siendo claves en la Transformación Digital en Salud.

en cada caso; o la programación con datos, que es la IA. A su vez, la programación con datos puede realizarse con entrenamiento, que es lo que se viene observando; o sin supervisión, lo que parece mágico. La novedad son los *Transformers*, de los que les voy a contar hoy.

La IA son programas que poseen habilidades antes consideradas exclusivamente humanas. Estas incluyen el ***Machine Learning***, que utiliza algoritmos capaces de reconocer patrones sin necesidad de programación explícita, y el ***Deep Learning***, que emplea redes neuronales para procesar datos. Dentro del ámbito de las redes neuronales, se encuentran dos innovaciones destacadas: los ***Transformers***, que son redes de atención, y las ***redes generativas antagónicas***, que permiten, por ejemplo, la creación de nuevas imágenes.

Actualmente, nos encontramos en lo que se conoce como el "verano de la IA". Este término se refiere a los períodos en los que surge un nuevo tipo de algoritmo, disponemos de más datos para procesar o contamos con una mayor capacidad de cómputo. Durante estos períodos, se genera la expectativa de que se resolverán todos los problemas que no hemos podido solucionar a lo largo de la historia. Sin embargo, este entusiasmo suele ser temporal. Cuando las promesas no se cumplen, se pasa a un "invierno de la IA", una fase en la que, aunque el progreso es más lento, continúa habiendo avances.

Hoy estamos en el tercer "verano de la IA", impulsado por las redes neuronales, las redes de atención y las redes generativas antagónicas. Estas tecnologías permiten trabajar con texto de diversas formas: clasificar, realizar regresiones para predecir valores, anticipar la siguiente palabra, como lo hacen los modelos Transformers, y generar cualquier tipo de multimedia a partir de datos existentes.

Hasta 2023, todo estaba relacionado con imágenes y contexto. GPT no nació de la nada. La versión 2 era muy limitada, pero la versión 3.5 generó un cambio importante porque el modelo mejoró mucho. Ahora, se puede introducir un texto a una máquina que puede predecir una respuesta.

¿Cómo se entrenó ChatGPT? Básicamente, **se le entregaron todos los textos posibles y se procesaron por una *red neuronal de atención*, que aprende cómo los humanos escriben al recibir todos los textos escritos por personas.** Así, la máquina entiende cómo juntamos las palabras para hacer un texto coherente. Después, se la entrena para tareas específicas, como por ejemplo saber hacer un resumen, entregándole textos originales y sus resúmenes escritos por humanos.

OpenAI entrenó a ChatGPT con una base de datos pública llamada Common Crawl Dataset. Con la capacidad y el dinero para pagar un gran data center, se pudo procesar todo el texto de internet, y así lograron armar estas redes neuronales de atención.

Hay un hiperparámetro llamado **temperatura**, que permite ajustar cuánto varía el texto. Al armar la frase, elige una palabra y luego otra que combine con ella. La probabilidad de cada palabra correcta, está relacionada con la temperatura que ajusta la creatividad del modelo. Con una **temperatura alta**, el texto se vuelve más creativo, pero también puede volverse incoherente. Con una **temperatura baja**, el texto es más parecido a los textos con los que fue entrenado y puede parecer escrito por un robot. La temperatura por defecto es 0.7, lo que permite variar y hacer el texto más humano.

GPT, por defecto, varía y cambia su expresión en los textos. ¿Cómo logra GPT generar estos textos? La red de atención tuvo suficientes datos para entender que los humanos varían los textos sobre el mismo tema y no escriben de manera robótica. El modelo fue entrenado con todos los textos disponibles en internet, en todos los idiomas, lo que le permitió identificar relaciones entre las palabras. Por ejemplo, sabe que los humanos suelen escribir "artritis" y "reumatoide" juntos y que, cuando usan "artritis reumatoide", también suelen emplear términos como "autoinmune" e "inflamación". Estas palabras tienen sentido para los humanos cuando se agrupan en un mismo texto. Este conocimiento se aplica a todos los textos que GPT ha procesado en internet.

Así es como funcionan los grandes modelos del lenguaje: son un sistema que puede predecir cuál es la siguiente palabra que tiene que poner para armar un texto coherente.

Lo que revolucionó el funcionamiento de los modelos de lenguaje fue la incorporación de evaluaciones humanas del input y output del modelo durante muchos meses. En plataformas de trabajo freelance, contrataron a personas de diferentes partes del mundo para evaluar las respuestas del modelo. Esto permitió la creación de un segundo modelo enfocado en la evaluación de las respuestas.

Un enfoque eficiente para interactuar con ChatGPT es evitar explicaciones largas. Cuanto más corto y preciso sea el prompt, mejor captará la información importante. No se trata de hablarle a un ser humano, sino de dar órdenes a una computadora que solo necesita algunas palabras clave para generar una respuesta. Palabras como "por favor" y "gracias", así como un texto gramaticalmente bien expresado, son eliminados, y el modelo solo toma las palabras importantes para buscar en su base de conocimiento cómo responder.

Las principales limitaciones de estos modelos de lenguaje son que están diseñados para generar respuestas convincentes, sin importar si son correctas o no. El modelo de lenguaje genera diferentes tipos de respuestas, y otro modelo evalúa cuál de ellas es más probable que reciba la aprobación de un humano. **Estos modelos no saben si lo que dicen es cierto; no piensan, solo predicen el siguiente texto basándose en patrones de datos previos.**

Los modelos pueden ejecutar acciones, pero no son una base de datos y este es un concepto muy importante. Las plataformas de ***Large Language Model*** **(LLM)** más potentes, como las de OpenAI o Google Cloud, tienen programación simple llena de "if". Por ejemplo, si preguntas sobre el tiempo, dice que no tiene información en tiempo real; si preguntas sobre religión, tiene una respuesta específica; si preguntas sobre medicina, te dice que consultes a un médico. Estas son todas cosas que los humanos fueron armando con algoritmos simples para identificar ciertas preguntas y dar respuestas predeterminadas (políticamente correctas). Luego está la programación con datos, lo que el modelo realmente hace: intenta armar un texto.

Es fundamental comprender que los modelos de lenguaje no operan como bases de datos convencionales; en su esencia, intentan generar texto coherente a partir de los textos con los que fueron entrenados. Aunque pueden simular respuestas basadas en patrones de datos previos, es importante recordar que estas respuestas no siempre reflejan información precisa o actual.

EL CASO DE LA HCE DE LA CLÍNICA ALEMANA DE SANTIAGO, CHILE

Este proyecto, denominado **AlemanaGPT**, comenzó con una encuesta que realizamos a médicos, enfermeros y otros profesionales de la salud. Recibimos una gran cantidad de sugerencias, pero en resumen, querían que el modelo pudiera realizar casi cualquier tarea.

Primero, creamos un administrador de prompts. Sabíamos que sería difícil acertar desde el principio, así que construimos un administrador que permite generar en cualquier parte de la historia clínica la posibilidad de llamar a **AlemanaGPT**. Este administrador también permite variar el prompt, limitar que ciertos prompts aparezcan para ciertas especialidades y especificar qué parte de la historia clínica debe utilizarse para cada petición.

¿Qué peticiones armamos? Primero, un **resumen del episodio**: se le envía todo el texto que está escrito en un episodio de atención y el modelo devuelve un resumen. Siempre que se devuelve algo escrito por el modelo, se permite que los profesionales lo editen, lo puedan imprimir, enviar por correo electrónico o descargar.

En el administrador de prompts, tenemos configurados los tipos de resúmenes que entrega por especialidad. Se le ingresa la acción que queremos que el modelo de lenguaje realice (por ejemplo, un resumen) y cómo queremos que lo entregue (como un texto de 700 caracteres). Lo que se hace es enviar la acción que buscamos del modelo, proporcionando el texto escrito en la historia clínica, y el modelo responde en consecuencia.

Por ejemplo, un **informe médico** a partir de lo escrito en la evolución del paciente. El modelo es buenísimo para hacer eso, es excelente para transformar cosas.

Además, tiene un **uso educativo** que inicialmente no esperábamos que fuera tan eficaz, pero que los médicos han encontrado muy útil para trabajar con los residentes. Se le solicita

al modelo que genere preguntas de opción múltiple sobre el paciente visto con el tutor, y luego los residentes responden a estas preguntas.

Otro uso es la **generación de handouts para los pacientes**. El profesional de la salud escribe en terminología médica (con siglas y términos difíciles de comprender) y se le pide al modelo que lo traduzca a un lenguaje accesible para el paciente. También cuando escriben una evolución, pueden pedirle al modelo que les **mejore la redacción** o hasta que **sugiera un plan de tratamiento** para este paciente en base a lo escrito.

Ahora estamos desarrollando un copiloto, que es funcional dentro de una historia clínica. Este copiloto puede leer cualquier cosa que está en la historia clínica, responder preguntas y realizar tareas de **RAG (Retrieval-Augmented Generation**). RAG consiste en una búsqueda inicial utilizando un modelo de lenguaje (LLM) y luego una transformación de los resultados con el mismo LLM. Por ejemplo, puedo pedirle que busque en toda la historia clínica un resultado de laboratorio. El copiloto buscará en las evoluciones y en los resultados de laboratorio. Si solicito los últimos cinco valores de hemoglobina glicosilada, el copiloto puede decir: "En las evoluciones alguien escribió 7.5, y en el laboratorio apareció tal valor".

Además, se puede crear una funcionalidad para generar recetas médicas. El profesional de la salud escribe en texto libre lo que desea que el copiloto haga, y el sistema automáticamente lo formatea y monta en la receta.

Para realizar esto necesitas una institución que te apoye y financie. Innovar es caro, innovar es equivocarse. La Clínica Alemana se reconoce como una institución súper innovadora, y recientemente fue premiada por ello.