TiES Revista de Tecnología e Innovación en Educación Superior

www.ties.unam.mx

De las ideas verdes incoloras hasta ChatGpt: los grandes modelos del lenguaje

Víctor Germán Mijangos de la Cruz

Universidad Nacional Autónoma de México, Facultad de Ciencias, Departamento de Matemáticas, Ciudad de México, México.
ORCID: 0000-0002-8950-2634

Ximena Gutierrez-Vasques

Universidad Nacional Autónoma de México, Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Ciudad de México, México.

ORCID: 0000-0002-1486-2774

Recepción: 21 de abril de 2024.
Aceptación: 17 de mayo de 2024.

Junio 2024 • número de revista 10 • https://doi.org/10.22201/dgtic.26832968e.2024.10.18

De las ideas verdes incoloras hasta ChatGpt: los grandes modelos del lenguaje

Resumen

Los grandes modelos del lenguaje son tecnologías que han mostrado una capacidad notable para producir texto que simula al lenguaje humano escrito; estos modelos están detrás de agentes conversacionales como chatGPT o Gemini. Si bien el impacto y uso de estos modelos se ha extendido a numerosos sectores de la sociedad, no siempre se discuten los fundamentos técnicos y científicos que subyacen a estos desarrollos de la inteligencia artificial. El presente artículo propone dar una introducción al funcionamiento de los modelos del lenguaje, desde las primeras propuestas hasta los grandes modelos actuales. Lo anterior con el fin de incentivar una comprensión más profunda de estas tecnologías y, por lo tanto, ampliar la discusión en torno al origen de algunas de sus limitaciones y potencialidades en diversos ámbitos, por ejemplo, en un marco educativo.

Palabras clave: Modelos del lenguaje, procesamiento del lenguaje natural, inteligencia artificial.

From colorless green ideas to ChatGpt: Large Language Models

Abstract

Large language models are technologies that have shown a remarkable ability to produce text simulating human-written language; these models are behind conversational agents like chatGPT or Gemini. Although the impact and use of these models have extended to numerous sectors of society, the technical and scientific foundations underlying these artificial intelligence tools are not regularly a topic of discussion. This article aims to introduce the functioning of language models: from the early proposals to the current large models. We aim to encourage a deeper understanding of these technologies and thus broaden the discussion around the origins of some of their limitations and potentialities in various fields, for example, in an educational framework

Keywords: *Language models, natural language processing, artificial intelligence.*

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Introducción

"Colorless green ideas sleep furiously" es un famoso ejemplo ideado por el lingüista Noam Chomsky para ilustrar que aunque una oración esté perfectamente bien formada en lo gramatical, no necesariamente tiene coherencia semántica. Podríamos decir que algo similar ocurre con las tecnologías del lenguaje y los avances de la inteligencia artificial. Por muchos años ha predominado la percepción de que, si bien las computadoras logran imitar algunas capacidades del razonamiento y del lenguaje humano, esto es solo superficialmente, pues no logran un entendimiento realmente profundo del lenguaje.

Sin embargo, tras la aparición de los grandes modelos de lenguaje, nuestra noción sobre los límites de estas tecnologías está siendo desafiada cada vez más. En Procesamiento del Lenguaje Natural (PLN) los grandes modelos de lenguaje son tecnologías capaces de generar artificialmente trozos de texto que no sólo exhiben coherencia en lo sintáctico, sino que parecen codificar relaciones más profundas: de tipo semántico, pragmático, conocimiento del mundo, etc. La capacidad de estos modelos para generar texto ha llegado a tales niveles de sofisticación que muchas veces resulta imposible identificar si el texto fue escrito por un humano o producido a partir de una inteligencia artificial.

Quizá los ejemplos más emblemáticos de estos avances son "ChatGPT"[1] de la empresa OpenAI o "Gemini"[2] de Google. Estas herramientas son resultado de la combinación de dos tecnologías principales: 1) un modelo neuronal del lenguaje que ayuda a predecir las secuencias de texto más probables y 2) un chatbot, o agente conversacional, que facilita la interacción con los usuarios mediante un esquema de pregunta-respuesta.

Puesto que tecnologías como ChatGPT y Gemini se han vuelto objeto de discusión en casi todos los ámbitos, así como una herramienta de uso cotidiano para el público general, valdría la pena repasar cómo funcionan los grandes modelos de lenguaje para entender el origen de sus capacidades, así como sus potenciales limitaciones.

Antecedentes: Antes de los grandes modelos

En su concepción más fundamental un modelo de lenguaje busca calcular qué tan probable es que ocurra una determinada oración, o secuencia de texto, en una lengua específica. El poder ponderar qué tan viables resultan las oraciones que fueron generadas artificialmente y elegir las más probables ha sido un componente esencial para el desarrollo de sistemas como la traducción automática, la generación automática del lenguaje, la corrección gramatical, entre otros.

[1] https://chatgpt.com/

[2] https://gemini.google.com

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Los modelos del lenguaje se definen a partir de dos componentes principales: 1) el vocabulario o conjunto de elementos que se pueden combinar para formar unidades de una lengua; y 2) una función de probabilidad para cada una de las posibles combinaciones del vocabulario. Podemos imaginar que, si el vocabulario es el léxico del español, entonces un modelo del lenguaje deberá contar con una función que sea capaz de estimar la probabilidad para cada una de las posibles combinaciones de palabras. Oraciones comunes como "Hola, ¿cómo estás?" tendrán una probabilidad alta, mientras que formaciones sin estructura como "estás hola cómo" tendrán una probabilidad baja o nula.

Sin embargo, la cantidad de cadenas que se pueden formar con un vocabulario finito es infinita. Por tanto, se han propuesto diferentes estrategias para acercarse al cálculo de estas probabilidades. Una de las primeras propuestas para estimar modelos del lenguaje fue la de Claude Edward Shannon, creador de la Teoría de la Información. Para Shannon [1], las probabilidades de cadenas se estiman por medio de lo que el autor llama una aproximación de orden *n*. Esta aproximación factoriza las probabilidades de cadenas en probabilidades condicionales de una palabra dado *n-1* elementos previos (que se suelen llamar historial). El orden *n* del modelo determina qué tanto del historial tomamos en cuenta para predecir el siguiente elemento. Por ejemplo, usando una aproximación de orden 2, la probabilidad de una cadena con los símbolos w_1, w_2...., w_n puede estimarse como:

$$\begin{aligned} p(w_1, ..., w_n) &= p(w_1)p(w_2|w_1)\cdots p(w_n|w_{n-1}) \\ &= p(w_1)\prod_{t=2}^{n} p(w_t|w_{t-1}) \end{aligned}$$

La frase "Colorless green ideas" requerirá que estimemos las probabilidades *p(colorless)*, *p(green|colorless)* y *p(ideas|green)*, a partir de un gran corpus (colección de textos) de la lengua, para obtener la probabilidad total de la cadena. A esto se le conoce como un modelo de lenguaje estadístico basado en n-gramas. Además de asignarle una probabilidad a una oración de una lengua, los modelos de lenguaje tienen capacidad generativa, es decir, permiten generar nuevas cadenas a partir de un historial o un contexto previo. Para esto, podemos buscar la palabra *w* que maximice la probabilidad $p(w|w_1...w_n)$. Si incorporamos esta nueva palabra predicha a la cadena previa, $w_1...w_nw$, podemos repetir el procedimiento y buscar otra que maximice la probabilidad dado este nuevo contexto $p(w'|w_1...w_nw)$. Repitiendo este proceso generamos una nueva oración del lenguaje (véase [2]).

Este tipo de probabilidades condicionales se estiman a partir de simples conteos en un corpus de entrenamiento con textos de alguna lengua. Sin embargo, estos modelos tradicionales del lenguaje exhiben algunas limitaciones: el tamaño del contexto es siempre fijo y limitado (generalmente una, dos o tres palabras previas), dificultando que los modelos de lenguaje puedan hacer predicciones que dependan de un contexto de tamaño variable, o que puedan generar texto condicionado a dependencias muy distantes. Por otro lado, estos modelos tienden a generar texto formado por oraciones sintácticamente correctas pero muchas veces carentes de coherencia semántica; por lo que la adopción de este tipo de modelos para realizar agentes conversacionales no era tan común en la industria, comparada con los sistemas de IA generativa hoy en día.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Los primeros modelos neuronales

Los modelos de n-gramas evidenciaron el potencial de usar métodos estadísticos para procesar el lenguaje. Una nueva ola de modelos partió de esta base, sin embargo, ahora utilizaron redes neuronales artificiales como herramienta para estimar las probabilidades. A principios de los años 2000, Yoshua Bengio, premio Turing por sus contribuciones en inteligencia artificial, propuso un modelo del lenguaje basado en redes neuronales artificiales. La propuesta de Bengio [3] era utilizar una red neuronal multicapa para estimar las probabilidades de una palabra dado su historial.

En la Figura 1 se muestra la arquitectura de la red neuronal. La salida es simplemente la probabilidad de la *i*-ésima palabra, dado el contexto (esta probabilidad se calcula utilizando una función llamada *softmax*). En los pasos intermedios del procesamiento que realiza la red se utilizan lo que se conoce como capas ocultas, componente fundamental del aprendizaje profundo (entre más capas ocultas mayor profundidad). Lo interesante, además de la capacidad de las redes neuronales para predecir palabras dado un contexto, es cómo estas palabras se representan para que las procese la red neuronal. Las redes neuronales toman como entrada vectores (arreglos numéricos) pues no son capaces de trabajar directamente con cadenas de texto. Para solventar esto, Bengio [3] propone *embeddings* (a veces traducido como "encajes") de las palabras en vectores. En la Figura 1, estos *embeddings* se denotan como $C(w_t)$; los valores de estos vectores se aprenden automáticamente como parte del proceso de entrenar una red neuronal para que prediga una palabra dado un contexto.

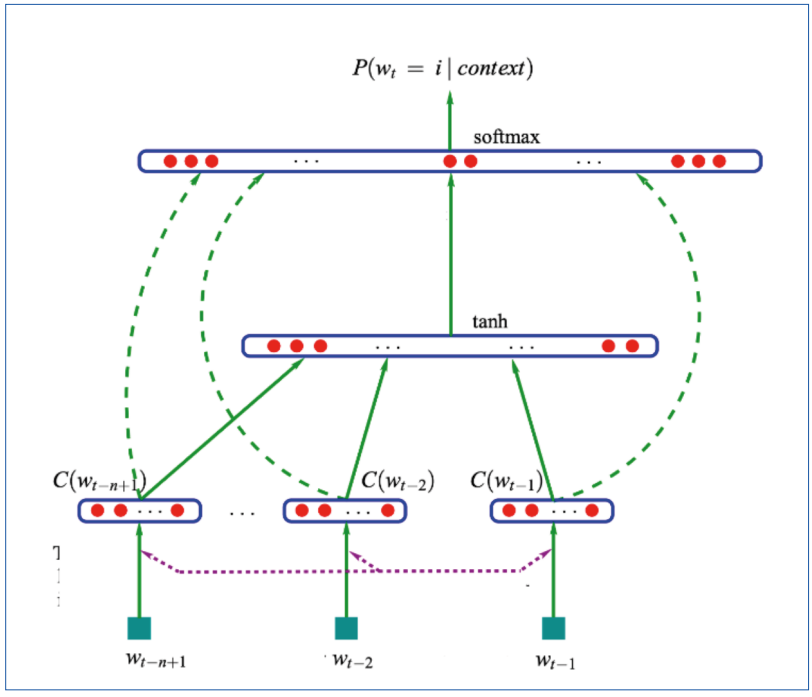

Figura 1. Estructura de la red neuronal para el modelo del lenguaje (adaptado de Jurafsky [2]).
Fuente: elaboración propia.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Cuando la red termina de entrenarse, se obtienen *embeddings* que convergen a valores numéricos que capturan la semántica de las palabras. Como se muestra en la Figura 2, cada *embedding* es un vector asociado a una palabra del vocabulario. Estos vectores tratan de capturar las relaciones semánticas, de tal forma que vectores que correspondan a palabras semánticamente relacionadas estarán cercanas ("gato" y "minino" en la Figura 2) y alejadas de aquellas con significados distintos. El fundamento que permite a la red aprender *embeddings* es la hipótesis distribucional [4]: palabras con sentidos similares comparten contextos similares. A partir de observar miles o millones de contextos en un gran corpus, la red neuronal puede estimar los vectores de palabras.

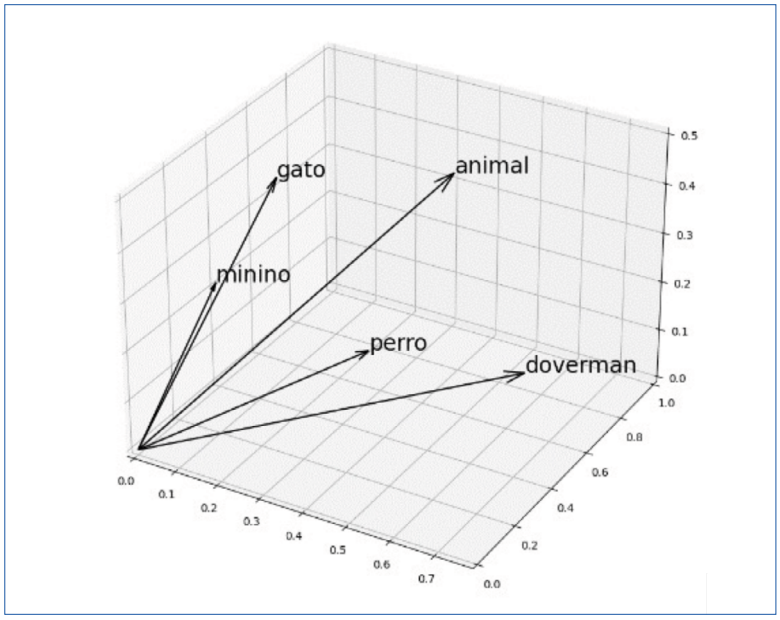

Figura 2. Visualización de los embeddings como vectores. Fuente: elaboración propia.

Esto fue una idea novedosa, de gran impacto; más allá de predecir secuencias, los modelos del lenguaje incorporaban capacidades semánticas mediante vectores. Así surgieron varias propuestas neuronales para crear vectores de palabra o *word embeddings*, la más popular fue Word2Vec [5], y le siguieron modelos como FastText [6], GloVe [6], Ida2Vec, Node2Vec [7], etc. Este cambio de los métodos estadísticos a los neuronales representó una revolución en el área de PLN e IA, la cual no sólo fue posible por los métodos nuevos sino porque el poder de los sistemas de cómputo ya permitía procesar arquitecturas neuronales con una vasta cantidad de datos.

Otra contribución importante llegó en 2017, cuando se propuso que estos vectores no sólo codificaran una noción estática de la "semántica" de las palabras aisladas, sino que su representación vectorial se transformara dinámicamente dependiendo del contexto. Por ejemplo, en la oración "él no nada nada" esperaríamos tener dos representaciones vectoriales distintas de la palabra "nada", la primera relacionada a la semántica del verbo nadar y la segunda tendría que ver con un adverbio de negación. En otras palabras,

la representación vectorial asociada a una palabra cambia de acuerdo con el contexto en que aparece. En este sentido, la idea de representaciones contextualizadas representó un cambio en la forma de aproximar el procesamiento del lenguaje natural. Métodos como CoVe (*Contextualized Vectors*), y ELMo [9] introdujeron los *embeddings* contextualizados y fueron precursores de los grandes modelos del lenguaje que se usan ahora.

Las entrañas de un gran modelo de lenguaje

Modelos como ELMo hacían uso de redes neuronales recurrentes. Estas arquitecturas resultan más flexibles para modelar el lenguaje, permiten tomar el contexto previo y el consecuente (bi-direccional) para modelar el *embedding* de una palabra, así como incorporar contextos variables de gran tamaño en la predicción de texto.

Las redes neuronales recurrentes dieron paso a los modelos *encoder-decoder o seq2seq* (*sequence to sequence*). En este tipo de paradigmas se utiliza una red neuronal para codificar una secuencia de entrada (*encoder*) y a partir de esta codificación se utiliza otra red neuronal (*decoder*) que aprende a decodificar generando otra secuencia. Muchas tareas de PLN pueden modelarse así, por ejemplo: a) traducción automática (entrada del *encoder*: texto en una lengua fuente, salida del *decoder*: texto en una lengua destino); b) resumen automático (entrada del *encoder*: una secuencia larga de texto, salida del *decoder*: una secuencia breve de texto que corresponde al resumen); c) generación de paráfrasis (entrada del *encoder*: una secuencia de texto, salida del *decoder*: una secuencia que corresponde a una paráfrasis del texto de entrada); d) generación de código (entrada del *encoder*: una secuencia de texto en lenguaje natural, salida del *decoder*: una secuencia de código de computadora).

Son tantas las aplicaciones que pueden abordarse como la conversión de una secuencia a otra que hoy en día representa el enfoque neuronal predominante para resolver tareas de PLN.

Sin embargo, las redes neuronales recurrentes demandan un alto costo computacional y tienen limitaciones cuando deben modelar contextos de gran tamaño. Para solventar lo anterior, Vaswani et.al. [10] introdujeron un modelo innovador llamado Transformer, que se basa en un mecanismo de las redes neuronales profundas llamado atención. La atención, como su nombre lo dice, permite poner mayor "atención" en ciertas palabras del contexto, así como "ignorar" otras, dependiendo de cuáles resulten más útiles para predecir una palabra dado un contexto largo. Este mecanismo se puede aplicar a muchas tareas de PLN, por ejemplo, en la traducción automática neuronal el mecanismo de atención permite que el *decoder* se enfoque en ciertas palabras de la secuencia de entrada (*encoder*) que tienen mayor relevancia para generar o predecir la secuencia de salida, es decir, qué palabras tienen mayor influencia para traducir una oración de una lengua a otra.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

La atención tiene un principio sencillo que se basa en ponderar qué tanto influyen las palabras de una secuencia de entrada (*encoder*) en la generación o predicción de las palabras en una secuencia de salida (*decoder*). Por ejemplo, en una tarea de traducción si se tiene una oración como "el niño se cayó" y se quiere traducir al inglés, el proceso de generación de la palabra "the" le asignará un score alto de atención a "el" y bajo a las demás palabras de la entrada, mientras que "fell" tendría que asignarle un *score* alto tanto a "se" como a "cayó". Estos *scores* se basan en probabilidades: las palabras que influyen más tienen mayor probabilidad de pasar su información al proceso de generación de la secuencia de salida (traducción). En la Figura 3 se muestra gráficamente la atención. En este ejemplo, se entrenó un Transformer para poder traducir sentencias del español al otomí del Mezquital. Si el sistema recibe la frase de entrada "Pase usted a descansar" (que se traduce en otomí "thogi gi zi tsaya"), los cuadros en blanco señalan dónde hay una mayor atención del modelo: para obtener la traducción de "thogi gi" se pone mayor atención a "pase". Para obtener "tsaya" la atención de la red neuronal se pone en su correspondiente español "descansar", mientras que para el reverencial "zi" se pone atención en "usted". En la oración original, "a" parece no requerirse para ninguna traducción (no se le pone atención), pues el otomí no cuenta con este tipo de preposición.

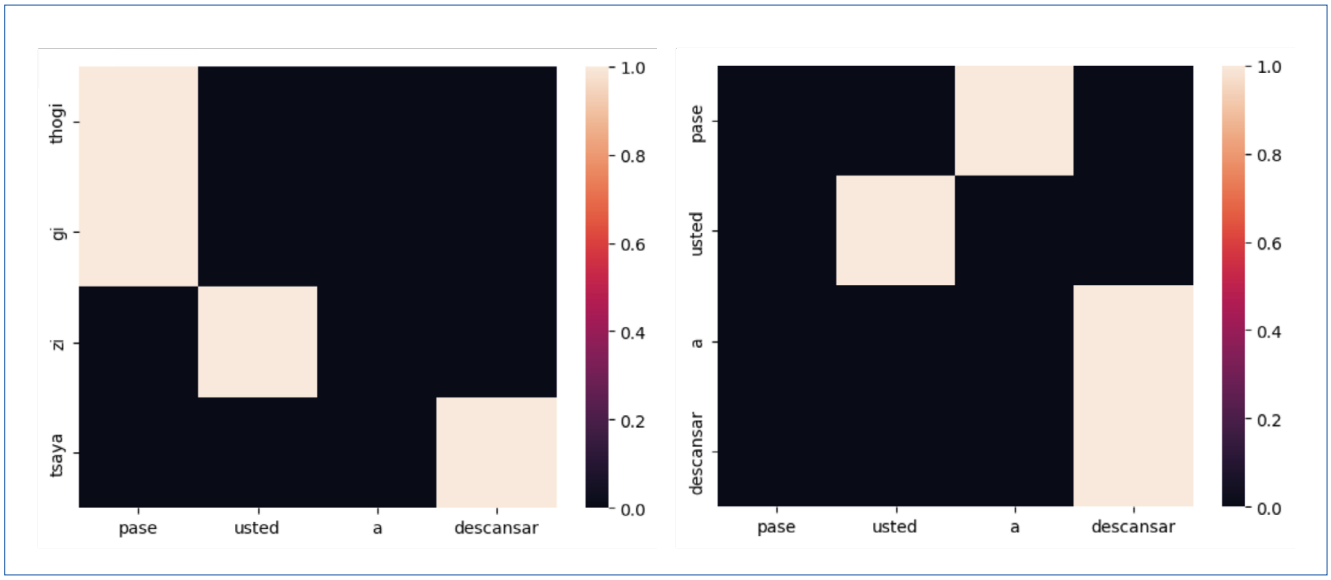

Figura 3. Matrices de atención en un problema de traducción de español a otomí. Fuente: elaboración propia.

Los valores representados con colores más claros son los más cercanos a una probabilidad 1, mientras que los más oscuros se acercan a probabilidad 0. En la derecha de la Figura 3 se observa otra matriz de atención. Esta se conoce como autoatención, pues se enfoca únicamente en los elementos de entrada; es decir, relaciona a la sentencia en español consigo misma, es una manera de obtener una codificación informativa y compacta de la estructura y las relaciones de una oración de entrada. Por ejemplo, en la Figura 3 vemos que el verbo "pase" está asociado a la preposición "a". La autoatención ha sido una innovación de

suma importancia para los modelos generativos. En la autoatención, los modelos capturan la estructura relevante de las secuencias de entrada (*encoder*) y a partir de esta representación pueden generar mejores salidas (*decoder*), ya sea en traducción automática u otras tareas de generación del lenguaje natural. Los Transformers son la arquitectura neuronal que está detrás de ChatGPT, LLaMA [11], T5 [12], BERT [13] y de muchos de modelos de lenguaje actuales.

La capacidad generativa de los modelos actuales se sigue basando en concepciones fundamentales como las de los modelos de n-gramas presentados más arriba: buscamos estimar la probabilidad $p(w|w_1,...,w_n)$ para generar, palabra por palabra, una secuencia que sea probable dada una entrada o contexto previo.

Como vimos, Claude Shannon, padre de la teoría de la información, ya había establecido las bases para calcular la probabilidad de ocurrencia de una palabra dado cierto contexto. La influencia de la estadística y el aprendizaje de máquina sigue siendo relevante para entender los modelos recientes.

Gracias a las innovaciones recientes (redes neuronales recurrentes, mecanismo de atención) se aprovecha un historial grande de palabras precedentes y ya no se limita sólo a unas cuantas. La familia de modelos del lenguaje que predicen la "mejor" palabra condicionada a un historial previo es conocida como modelos autorregresivos. Por ejemplo, si se introduce una cadena como "el gato persiguió al...", el modelo buscará, entre un vocabulario, cuál es la palabra que mejor se adapta al contexto previo; en este caso, el modelo podría elegir la palabra "ratón".

Estas estimaciones de probabilidad se obtienen de las distribuciones de palabras y oraciones de una lengua a partir de cantidades masivas de corpus textuales de entrenamiento. El uso de cantidades inmensas de texto combinado con el aumento de poder de cómputo y la sofisticación en las redes neuronales artificiales dieron pie a lo que ahora se conoce como grandes modelos del lenguaje (*Large Language Models* o LLMs por sus siglas en inglés). Como hemos discutido previamente, ese tipo de modelos neuronales tiene la capacidad de generar representaciones vectoriales de la lengua (*embeddings*) así como hacer predicciones de secuencias como parte de un mismo proceso de aprendizaje automático. Los grandes modelos del lenguaje han aumentado considerablemente su complejidad, así como sus capacidades, abriendo una ventana de posibilidades, pero al mismo tiempo generando nuevos retos y problemáticas.

Los límites actuales y el futuro

Debido a su naturaleza probabilística, estos modelos pueden entrenarse para que aprendan a realizar una gran diversidad de tareas. Por ejemplo, sistemas pregunta-respuesta, resumidores automáticos, generación automática de código de programación, entre muchas otras aplicaciones. Además, poseen cierta

capacidad "creativa", si bien no podemos decir que estos modelos son creativos en el sentido humano sí son capaces de generar textos novedosos y de responder de manera distinta a una misma pregunta por su base probabilística.

Hoy en día resulta difícil que un modelo de lenguaje produzca pedazos de texto que contengan oraciones agramaticales o con anomalías semánticas, como es la propuesta por Chomsky: "Colorless green ideas sleep furiously". Estos modelos parecen haber dominado la habilidad de reproducir muchas de las estructuras típicas subyacentes a una lengua natural. Los modelos actuales pueden producir estructuras formales adecuadas que además están enriquecidas con un conocimiento de la semántica y pragmática de las palabras con base en las representaciones vectoriales de éstas.

Aun así, los modelos del lenguaje tienen el potencial de generar información falsa muy elaborada y perfectamente bien escrita. En PLN, a este fenómeno se le ha denominado "alucinación" y constituye uno de los grandes retos actuales en la investigación en IA (puede revisarse a Xu & Kankanhalli [14] para profundizar).

Al ser tecnologías que entran en contacto con cada vez más usuarios, se debe ser especialmente cuidadoso con la veracidad y pertinencia del tipo de información que estos modelos son capaces de generar. En otras palabras, si un usuario utiliza un modelo generativo como asistente de escritura o para búsqueda de información, el usuario debe tener cierto conocimiento del tema para verificar que el texto generado no contenga estas "alucinaciones" y estar consciente de que, aún con todos los avances modernos, estos modelos basados en redes neuronales artificiales tienen la capacidad de producir texto con información falsa.

Si bien en el contexto actual se suele antropomorfizar a estos sistemas, es decir, les conferimos cualidades humanas pues interactuamos con ellas por medio de nuestro lenguaje natural, es importante no perder de vista de dónde viene el conocimiento de estas inteligencias artificiales. Estos modelos construyen relaciones complejas y una representación del mundo resultado de procesar millones de documentos extraídos de la web y de muchas otras fuentes. Los humanos no necesitamos leer toda la Wikipedia para aproximar el significado de las palabras. Los mecanismos de construcción del conocimiento son distintos ya que, entre otras cosas, los humanos tenemos acceso a mucha información que está fuera del texto (estímulos visuales, sonoros, etc.), mientras que los modelos del lenguaje suelen estar restringidos a capturar información interna al texto.

Como los corpus de entrenamiento determinan el tipo de conocimiento que adquieren estas tecnologías, es relevante pensar en el origen y calidad de los datos que alimentan a estos modelos. Es común observar que los modelos del lenguaje reproducen sesgos sociodemográficos de diversos tipos (raza, género, orientación sexual). La atenuación de estos sesgos es un aspecto relevante que se debe considerar en el desarrollo e implementación de estas tecnologías de inteligencia artificial con el fin de no perpetuar estereotipos ni

discursos de odio o discriminatorios (puede revisarse a Liang et al [15]). Sobre todo, si se planea que estas herramientas se integren plenamente a entornos educativos, gubernamentales y comerciales.

Por otra parte, mientras que el uso de cantidades exacerbadas de datos de entrenamiento es clave para el éxito de estos modelos, también constituye parte de su maldición. No todas las lenguas del mundo poseen los recursos textuales que se requieren para construir este tipo de tecnologías. A veces podemos tener la percepción de que la IA generativa se encuentra en un estado de desarrollo muy avanzado y disponible, pero esto es solo cierto para un puñado de lenguas. En la actualidad existen modelos de lenguaje multilingües pero sigue habiendo una distancia grande entre la calidad de predicciones que se obtienen para lenguas dominantes, como el inglés, y el desempeño que se obtiene cuando se prueban otras lenguas. Por más sofisticados que sean estos modelos, es un problema abierto el lograr que funcionen con una cantidad más modesta de datos y abarcar así la gran diversidad lingüística que existe en el mundo.

A pesar de las limitaciones, es una realidad que los grandes modelos de lenguaje son el componente principal de las tecnologías de inteligencia artificial con las que muchos interactúan en la actualidad. Estas tecnologías del lenguaje han provocado que cambie la forma en cómo realizamos búsquedas de información y resolvemos problemas. El ámbito educativo, por ejemplo, es uno de los que más se está transformando con la introducción de las tecnologías generativas, esto impacta las prácticas de los alumnos y de los profesores [16], [17].

Nuestra postura no es desincentivar el uso de los grandes modelos del lenguaje ni de las herramientas derivadas de la IA, sino fomentar una visión más transparente sobre sus limitaciones para seguir avanzando en su desarrollo y perfeccionamiento.

Más allá de negar sus capacidades, quizá la discusión deba centrarse en cómo incorporar estas tecnologías a la vida diaria de manera informada, segura y regulada. Sobre todo, considerando que las principales herramientas comerciales utilizan modelos de lenguaje que son “cerrados”. Esto implica que tanto el código fuente como los parámetros de la red neuronal y el corpus de origen no están disponibles públicamente, dificultando la transparencia, accesibilidad y reproducibilidad de estos desarrollos tecnológicos.

Referencias

[1] C. E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948.

[2] D. Jurafsky y J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2009.

[3] Y. Bengio, R. Ducharme, y P. Vincent, “A neural probabilistic language model,” en Advances in Neural Information Processing Systems, vol. 13, 2000.

[4] M. Sahlgren, "The distributional hypothesis," en Italian Journal of Linguistics, vol. 20, pp. 33-53, 2008.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, vol. 26, 2013.

[6] P. Bojanowski, E. Grave, A. Joulin, y T. Mikolov, "Enriching word vectors with subword information," en Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[7] J. Pennington, R. Socher, y C. D. Manning, "Glove: Global vectors for word representation," en Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.

[8] A. Grover y J. Leskovec, "node2vec: Scalable feature learning for networks," en Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855-864, 2016.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, y L. Zettlemoyer, "Deep contextualized word representations," en Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237, 2018.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, y I. Polosukhin, "Attention is all you need," en Advances in Neural Information Processing Systems, vol. 30, 2017.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, y G. Lample, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971*, 2023.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, y P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," en Proceedings of NAACL-HLT, pp. 4171-4186, 2019.

[14] Z. Xu, S. Jain, y M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models", *arXiv preprint arXiv:2401.11817*, 2024.

[15] P. P. Liang, C. Wu, L. P. Morency, y R. Salakhutdinov, "Towards understanding and mitigating social biases in language models", en *International Conference on Machine Learning*, pp. 6565-6576, 2021.

[16] L. Codina y C. Garde, "Uso de ChatGPT en la docencia universitaria: fundamentos y propuestas", *repositori.upf.edu*. [En línea]. Disponible: https://repositori.upf.edu/handle/10230/57015.

[17] L. J. Linares, J. A. L. Gómez, J. Á. M. Baos, F. P. R. Chicharro, y J. S. Guerrero, "ChatGPT: reflexiones sobre la irrupción de la inteligencia artificial generativa en la docencia universitaria", en *Actas de las Jornadas sobre la Enseñanza Universitaria de la Informática (JENUI)*, vol. 8, pp. 113-120, 2023.