El lenguaje humano y la informática*

Dina Wonsever

Señor presidente de la Academia Nacional de Letras, colegas académicos, queridos amigos, amigas y familia.

En primer lugar, quiero expresar mi agradecimiento a la Academia Nacional de Letras de Uruguay por el honor de haberme elegido miembro de número. Es mi intención honrar este nombramiento asumiendo con entusiasmo las actividades que implique. Agradezco también a Marisa Malcuori por las muy elogiosas palabras que me ha dedicado, en un cálido recibimiento. Con Marisa, así como con otros integrantes del Instituto de Lingüística de la Facultad de Humanidades y Ciencias de la Educación y del grupo Procesamiento de Lenguaje Natural de la Facultad de Ingeniería hemos compartido una fructífera etapa de trabajo conjunto y colaboración interdisciplinaria. Esta colaboración se materializó en varios proyectos de investigación financiados, colaboración en distintos cursos de grado y posgrado y codirección de tesis. Me da gran alegría esta nueva oportunidad de encuentros e intercambios en el ámbito de la Academia.

Ocupo el sillón de nombre Julio Herrera y Reissig, un destacado poeta uruguayo que murió muy joven y que vivió durante un período de su vida en la casa que es sede de la Academia. Las referencias a su obra incluyen muchas veces la mención a esa casa, que dispone de un altillo en la azotea que ofrece una vista panorámica al Río de la Plata, en la Rambla Sur de Montevideo. A pesar de su corta vida, Herrera y Reissig (1875-1910) tuvo una obra significativa y es considerado parte importante del patrimonio poético del Uruguay.

La Academia Nacional de Letras tiene una integración que incluye tanto expertos en lingüística, escritores, poetas, críticos literarios como historiadores, médicos, arquitectos, juristas.

Mi integración a esta corporación, siendo especialista en informática, significa para mí, y para mi grupo de investigación en la Facultad de Ingeniería, un gran orgullo y, a la vez, una gran responsabilidad. Un factor que creo que explica la honrosa distinción de

* Discurso de ingreso a la Academia Nacional de Letras, 24 de junio de 2024.

la que fui objeto es que dentro de la informática trabajo en un área «bisagra» entre la informática y la lingüística, que es el procesamiento del lenguaje natural. Otro factor para considerar es el particular momento que estamos viviendo en el desarrollo del procesamiento automático del lenguaje y, más en general, de la inteligencia artificial. Un ejemplo paradigmático es la existencia y la popularidad de nuevas herramientas conversacionales multilingües. Para decir un nombre que a esta altura es muy popular, herramientas del tipo Chat GPT (Open AI, 2024), con muy evidentes logros y también muy evidentes fallas.

En este contexto, creo que es pertinente dedicar mi ponencia a presentar, en primer lugar, una definición del área y una rápida recorrida, desde los orígenes hasta el momento actual, del procesamiento de lenguaje natural. Observar su evolución en sus más de setenta años de vida puede dar pautas sobre las posibilidades que ofrece y los problemas que enfrenta actualmente. Referido a la actualidad, hablaré sobre un aporte muy reciente: los muy grandes modelos de lenguaje (LLM*, Large Language Models*), que parecen encapsular toda una lengua.

Qué es el procesamiento de lenguaje natural. Definición y términos equivalentes

Una definición habitual del término *procesamiento de lenguaje natural,* de ahora en más PLN, es que es una subdisciplina de la inteligencia artificial que se enfoca en cómo lograr que las computadoras se comuniquen usando el lenguaje humano. Su objetivo principal es permitir que las computadoras comprendan, interpreten y respondan al lenguaje humano, ya sea hablado o escrito.

El PLN combina aspectos de la lingüística, la informática y otras disciplinas asociadas, siempre con una orientación concreta, buscando crear herramientas informáticas con buen desempeño. Se han usado como equivalentes los términos *tecnologías del lenguaje humano* e *ingeniería lingüística.* Estos dan una idea clara del carácter «ingenieril» del PLN que, si bien se apoya en fundamentos teóricos (lingüística, lógica, matemática), es generalmente en relación con la construcción de una herramienta informática.

Lingüística computacional y procesamiento de lenguaje natural

El término *lingüística computacional* se ha utilizado como equivalente a *procesamiento de lenguaje natural.* La definición que aparece en la página de la Association for Computational Linguistics (ACL)[1] no contribuye a separar nítidamente ambas áreas:

> La lingüística computacional es el estudio científico del lenguaje desde una perspectiva computacional. Los lingüistas computacionales están interesados en proporcionar modelos computacionales de diversos tipos de fenómenos lingüísticos. Estos modelos pueden estar «basados en el conocimiento» («hechos a mano») o «basados en datos» (estadísticos o empíricos) (ACL, 2024).

Es interesante señalar que el evento científico más importante para el PLN es justamente la conferencia anual de la ACL.

Se puede decir que el procesamiento de lenguaje natural y la lingüística computacional, si bien comparten el campo temático, tienen objetivos diferentes. El PLN está orientado a tareas de lenguaje, o sea a lograr que una computadora realice tareas tales como contestar preguntas, traducir, corregir la gramática de un texto, entre muchas otras. La lingüística computacional es una rama de la lingüística y, como tal, tiene objetivos diferentes, que pasan por explicar el funcionamiento del lenguaje.

La lingüística computacional funciona en gran conexión con el PLN, ya sea utilizando las herramientas de PLN para construir objetos de interés en lingüística, tales como árboles de derivación, o para determinar la similitud semántica entre dos términos, ya sea proponiendo modelos lingüísticos formalizados y aptos, por lo tanto, para una implementación informática que será base de desarrollos en PLN. Lo cierto es que ambas disciplinas han evolucionado de modo similar, hasta casi llegar a confundirse. En el momento de las teorías lingüísticas altamente formalizadas, son estos modelos los que toma como base el PLN, construyendo representaciones como paso intermedio para lograr sus objetivos. Cuando el pensamiento sobre el lenguaje asigna un rol central a los datos, o sea al conjunto

[1] Principal asociación dedicada a la lingüística computacional y al procesamiento de lenguaje natural, creada en el año 1962.

concreto de enunciados proferidos, o sea a la *performance* en la distinción competencia/*performance* propuesta por Chomsky (1965), la lingüística pasa a apoyarse en grandes corpus, las investigaciones sobre distintos fenómenos pasan a hablar de significación estadística y la lingüística computacional utiliza como apoyo para el análisis los procesos de tratamiento de texto que suministra el PLN. Un ejemplo paradigmático lo constituyen la anotación y desambiguación de categoría gramatical *(tagging)* y la generación de árboles sintácticos *(parsing)* cuando ambos procesos se realizan de modo automático.

El PLN en la vida cotidiana

Vivimos inmersos en una sociedad tecnológica que utiliza profusamente las Tecnologías de la Información y las Comunicación (TIC). Tenemos sistemas de comunicación que nos acompañan continuamente, y que nos conectan entre nosotros y con grandes repositorios de información centralizados a través de programas informáticos. La comunicación con las computadoras se hizo de modo clásico con lenguajes formales específicos, pero muchas de las tareas de gestión de la información son tareas de lenguaje. Escribir con función de autocompletado, acceder a corrección ortográfica y gramatical, buscar información, traducir de una lengua natural a otra son tareas de alcance general que se hacen en el ambiente informático. Para estas tareas habituales de manejo del lenguaje existen aplicaciones cada vez de mejor calidad, que suelen venir preinstaladas, o al menos accesibles, en el equipamiento básico de una computadora personal o de un teléfono inteligente. Una aplicación que llegó muy recientemente a un gran nivel de calidad, aunque aún no está disponible de modo integrado en herramientas usuales de procesamiento de texto, es el resumen automático.

El PLN y la inteligencia artificial

La inteligencia artificial (IA) es un concepto amplio que aparece referenciado desde los inicios de la computación. La podemos definir como la disciplina que estudia técnicas y algoritmos para resolver problemas que típicamente se realizan mediante el razonamiento humano.

Es claro que muchas de las capacidades típicamente humanas no se consideran comprendidas actualmente en lo que se entiende por

inteligencia artificial. Por ejemplo, la capacidad de cálculo mental superrápida siempre se asoció a la inteligencia humana, pero cuando se empezó a hablar de inteligencia artificial no se consideraba este aspecto, que ya estaba incluido en las primeras computadoras existentes. Otro hito de la inteligencia humana, el saber jugar bien a juegos basados en reglas como el ajedrez o el Go, sería uno de los primeros objetivos logrados por la inteligencia artificial. Actualmente son programas informáticos como Deep Blue (IBM, 2024; Campbell et al., 2002) o AlphaGo (Google, 2024) los vencedores en estos juegos, según narra apasionadamente Benjamín Labatut en su libro *Maniac* (2023), y ya no es más un desafío el lograr vencer el desempeño humano en estos juegos de estrategia (tal vez vuelva a serlo si aparecen otros jugadores más competentes). Deep Blue fue de los últimos exponentes de la inteligencia artificial tradicional, basado en algoritmia de «fuerza bruta»: métodos exhaustivos de búsqueda, con algunos refinamientos. A la parte de búsqueda se le agregaron heurísticas potentes, apoyadas muchas veces en una base de datos de partidas reales. Los mejores programas actuales se basan en redes neuronales y aprenden jugando contra sí mismos. Esto incluye a AlphaGo, vencedor frente al campeón de Go (Labatut, 2023).

La victoria en «juegos de inteligencia» es un objetivo ya logrado y no un hito a superar. Los puntos fundamentales que componen el *mainstream* de la IA actualmente son el procesamiento de lenguaje natural, el procesamiento de imágenes y el uso de métodos de aprendizaje automático, concentrados casi exclusivamente en redes neuronales profundas *(deep learning),* aplicados a muy diversos dominios científicos, tecnológicos y empresariales. Por ejemplo, en el dominio de las predicciones meteorológicas, el aprendizaje usando redes neuronales es el estado del arte actualmente, superando a los complejos modelos matemáticos que se estaban utilizando. Es interesante señalar que tanto el manejo del lenguaje como la percepción y reconocimiento visual y auditivo son áreas que han resultado muy difíciles para las computadoras, mientras que son completamente naturales y no requieren ningún tipo de aprendizaje por parte de los humanos.

Últimamente, además, aparecen nuevos modos de usar el término *inteligencia artificial* o la abreviación IA:

- Particularizándola a una entidad concreta, como si fuera un nombre contable. Este uso no es en principio adecuado, ya

que «inteligencia artificial» es un concepto abstracto. Pero aparecen variados ejemplos, del tipo «una IA de seguridad», «una IA para detectar *spam*». Se trata simplemente de programas especializados en alguna tarea concreta, que utilizan ya sea métodos de aprendizaje o alguna función de reconocimiento de imágenes o de habla/texto escrito. Se utiliza a veces la sigla en inglés ANI, por *Artificial Narrow Intelligence.* El uso de estos términos está asociado al *marketing,* en el que aparece también el término *hype,* que indica algo que genera muy fuertes expectativas, muchas veces exageradas por la publicidad. Esta utilización «marketinera» de la IA está asociada a los desprestigios que ha sufrido (y que seguramente vuelva a sufrir) en círculos académicos.

- En la inteligencia artificial general (IAG): se utiliza este término para lo que sería «la inteligencia en serio», la inteligencia al estilo humano, general, en máquinas, y a la que todavía no se ha llegado.
- También se ha hablado de la IAS, esta sería la superinteligencia (*s* de súper), y corresponde a la superioridad de máquinas sobre humanos. Aquí ya estamos en el terreno de la ciencia ficción.

Breve historia del PLN

En la década de los cincuenta del siglo pasado ocurrieron eventos relevantes en la informática y en el procesamiento automático del lenguaje humano. Por una parte, el nacimiento efectivo de la primera computadora «física» de programa almacenado (ya existe una antecesora abstracta que es la máquina universal de Turing [1937]). Se trata de la máquina de Von Neumann, con una arquitectura que se habría de mantener a lo largo de muchos años. En paralelo con lo que estaba ocurriendo con la inteligencia artificial (en el reducido círculo que manejaba estos temas), existía gran optimismo en los posibles logros de la nueva máquina.

Se presenta a continuación una reseña histórica de lo que pasó con el procesamiento de lenguaje en sus más de setenta años de desarrollo, a través de una selección de algunos hitos significativos, tanto en los logros tecnológicos como en los fundamentos teóricos.

Década del cincuenta

1950: Alan Turing publica *Computing Machinery and Intelligence*, proponiendo el test de Turing como criterio para determinar la inteligencia de las máquinas. El *Juego de la imitación* es el nombre que le dio Turing a su propuesta: consistía en un experimento en el que un jurado se comunicaba con dos individuos a los que no veía ni escuchaba. Uno de los individuos en realidad era una computadora, y esta computadora pasaba el test si el jurado no distinguía entre la persona y la máquina. Es significativo que se haya elegido como representante de la inteligencia humana la capacidad de lenguaje. Esta elección tal vez no haya sido la más afortunada, como parecen indicar los actuales grandes modelos de lenguaje. El test de Turing sería superado en varias oportunidades, y actualmente los grandes modelos generativos brindan ejemplos de diálogos fluidos, con perfecto dominio del lenguaje, pero muchas veces sin apego a la verdad y con errores severos cuando se les plantean pequeños problemas de razonamiento. La conclusión sería (y esto tal vez no es una novedad) que se puede hablar muy bien y pensar muy mal.

1954: Una tarea prioritaria que se le plantea a la novel computadora a comienzos de la década del cincuenta, y en épocas del inicio de la Guerra Fría, es la traducción automática ruso-inglés. En un experimento conjunto entre la Universidad de Georgetown y la IBM (Hutchins, 2004), en 1954, se presenta una demostración con la traducción de 60 oraciones del ruso al inglés en una máquina IBM 701. Se trataba de un conjunto de ejemplos simples y muy controlados: en total, había 250 palabras en juego, las oraciones habían sido cuidadosamente elegidas, se habían escrito a mano algunas reglas de traducción que bastaban para los ejemplos seleccionados. El experimento fue visto como un éxito y esto motivó que fluyera el financiamiento de distintas fuentes a proyectos de traducción automática.

1956: Noam Chomsky publica un artículo con la así llamada *jerarquía de Chomsky,* una clasificación de lenguajes formales definidos por un paradigma generativo y a los que se asocia un autómata reconocedor. Hay una definición concisa de lo que es un lenguaje generado por una gramática y una clasificación en cuatro niveles, según su complejidad, de lenguajes formales. Los tipos 2 y 3 de esta clasificación, conocidos respectivamente como lenguajes independientes de contexto y lenguajes regulares, han sido de vital importancia en el desarrollo de los lenguajes de programación. Los

lenguajes tipo 3 han sido ampliamente utilizados en la descripción de la morfología de lenguas naturales.

1957: Noam Chomsky publica *Syntactic Structures,* donde se presentan varios aspectos centrales para la teoría lingüística conocida como gramática generativa, tales como el modelo de reglas de reescritura tipo 2 (combinado con transformaciones, que luego se abandonaron) para la descripción de la sintaxis, la autonomía de la sintaxis respecto a los significados y el rechazo a los modelos probabilistas en el estudio del lenguaje.

1957: Frank Rosenblatt (1958) propone un modelo —perceptrón— orientado al reconocimiento de patrones e inspirado en las neuronas que componen el cerebro. Es la primera red neuronal artificial, inicio de una línea de trabajo que se mantuvo poco conocida por décadas y que constituye actualmente la base de modelos que revolucionaron el estado del arte.

Década del sesenta

1965: ELIZA es uno de los primeros sistemas de diálogo reportados. Con propósitos de demostración, Joseph Weizenbaum (1966) genera un programa que imita a un psicoterapeuta rogeriano que interactúa con su interlocutor, invitándolo a expresarse e interviniendo lo menos posible. Está basado en reglas y utiliza estrategias simples como asociar a palabras significativas algunos esquemas de frase, rellenando huecos según la situación. Pensado como un entretenimiento tuvo consecuencias inesperadas, ya que resultó creíble y los usuarios lo visualizaban como un interlocutor a quien confiarle sus problemas.

1966: El informe ALPAC (Automatic Language Processing Advisory Committee [Hutchins, 1996]) declara que los logros de la traducción automática a ese momento desaconsejan su uso, incluso en combinación con un traductor humano, y propone no continuar invirtiendo. Esto inicia un período de congelamiento de la investigación

Década del setenta

1970: Se desarrolla SHRDLU, un programa de comprensión del lenguaje natural creado por Terry Winograd en el MIT, que interactúa

con un mundo de bloques virtual. Es un sistema de diálogo de comprensión completa sobre un dominio muy acotado. Hay objetos de distintos tipos (cubos, prismas, esferas, etcétera) que tienen asociado un color y una ubicación (sobre el piso, sobre una mesa, sobre otro objeto). Hay un usuario que puede interrogar sobre el estado del sistema, preguntando dónde está cada objeto. El usuario puede también realizar acciones, cambiando la ubicación de un objeto. Se trata entonces de un mundo supersimplificado y se espera «comprensión» completa. Obviamente, el rango de formas sintácticas y palabras que este sistema puede comprender es muy limitado, pero para las formas adecuadas hay interacción verdadera. Es, en algún sentido, lo opuesto a ELIZA, que da la impresión de comprender todo, pero se maneja simplemente por patrones de texto.

1972: PARRY fue presentado en 1972 por el psiquiatra Kenneth Colby. Mientras que ELIZA era una simulación de un terapeuta, PARRY simulaba a una persona con esquizofrenia paranoide. De hecho, PARRY y ELIZA interactuaron en algunas oportunidades. Superó el test de Turing en una instancia de este adaptada a la esquizofrenia, con psiquiatras como jueces y pacientes reales como humanos participando además de PARRY.

1972: Se introduce el sistema de respuesta a preguntas LUNAR, desarrollado por William Woods (1977) para consultas sobre un conjunto de datos geológicos de la Luna recogidos por la misión Apolo 11. La información está gerenciada por un manejador de bases de datos y la consulta del usuario se traduce a una consulta en el lenguaje específico del manejador. Es un sistema de comprensión completa, dominio acotado y lenguaje obligatoriamente restringido, en la línea de SHRDLU, pero con un nivel de complejidad mayor. Se desarrollaron muchas aplicaciones de dominio específico y lenguaje restringido como esta, algunas estuvieron operativas mucho tiempo. Los problemas esenciales de sistemas de este tipo fueron el gran tiempo de desarrollo y la fragilidad lingüística, asociada al modelado determinista de la gramática que describe los formatos de las preguntas. La incorporación en PLN de modelos probabilistas brindaría un enfoque mucho más robusto frente a las variantes lingüísticas de la pregunta.

Década del ochenta

1980: Los métodos de gramática basados en reglas como *Lexical-Functional Grammar* (LFG) y *Head-driven Phrase Structure Grammar* (HPSG) se desarrollan y ganan popularidad. La década del ochenta marca el apogeo de los modelos simbólicos.

1982: Se publica *Frame Semantics,* de Charles Fillmore. Se propone una semántica basada en situaciones prototípicas que incluye conocimiento enciclopédico. Es una alternativa a la semántica veritativo-condicional asociada a los sistemas simbólicos. Es la base de un recurso de representación semántica ampliamente utilizado, FrameNet, que ofrece recursos descriptivos de libre acceso, tanto en entidades léxicas como en corpus anotados.

1986: Algoritmo de descenso por gradiente *(backpropagation)* para el entrenamiento de redes neuronales multicapa. En 1986 se publica un artículo de Rumelhart et al., en la revista *Nature,* que incluye una evaluación experimental del método. Este método es esencial para el desarrollo del paradigma de redes neuronales. Se reporta que este método se manejó de modo previo e independiente en varios lados y por distintos autores, pero se suele ubicar el artículo de 1986 como referencia. La eficacia de este método es fundamental para el desarrollo práctico de redes neuronales de gran porte.

Década del noventa

1990: Los modelos ocultos de Markov (HMM) y los métodos estadísticos empiezan a dominar en el reconocimiento del habla y el etiquetado de partes del habla. También se utilizan para el etiquetado automático con categoría gramatical (POS *tagging*), que logran una muy buena *performance* y habilitan una etapa siguiente de análisis sintáctico probabilístico.

1992: Se lanza la primera versión del Penn Treebank, corpus del inglés anotado con árboles sintácticos. Se inicia un paradigma que corre en paralelo con la aplicación de métodos de aprendizaje automático en PLN. Los métodos de aprendizaje supervisado se apoyan en los corpus anotados para el entrenamiento de modelos. La anotación se hace en primer lugar con árboles, pero más adelante se construirán esquemas de anotación para fenómenos variados, por ejemplo semántica temporal o relaciones discursivas.

1992: IBM desarrolla el modelo de alineación de oraciones y frases para la traducción automática estadística como parte del proyecto Candide (Brown et al., 1990). Se propone un sistema estadístico que se apoya en corpus paralelos para deducir probabilidades a partir de ejemplos. Se transforma el problema de la traducción en un problema de alineamiento de frases en dos idiomas, para lo que se propone el método del canal ruidoso, basado en teoría de la información. El modelo se complementa con dos modelos de lenguaje, el del lenguaje destino y un modelo condicional de transferencia.

1998: Proyecto FrameNet (Baker et al., 1998) en la Universidad de Berkeley, Estados Unidos. FrameNet implementa una versión de semántica por roles argumentales de predicados verbales, con nombres de roles locales al marco de definición. Se produjeron descripciones semánticas de varios miles de elementos léxicos del inglés y un corpus anotado.

1998: Se introduce WordNet (Fellbaum, 1998), una base de datos léxica del inglés que agrupa palabras en conjuntos de sinónimos (*synset,* en la terminología de WordNet). Sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos, cada uno de los cuales expresa un concepto distinto. Los *synsets* están interconectados mediante relaciones semánticas como la sinonimia, hiperonimia o antonimia. A partir del proyecto inicial en inglés, se construyeron versiones en varios idiomas, tales como español, francés, portugués, entre otros, utilizando el esquema original. Existen actualmente versiones en más de cien lenguas y una organización (Global WordNet) que promueve interacciones y conferencias específicas.

Primera década del siglo XXI

2006: Google Translate lanza su servicio abierto de traducción automática estadística basado en grandes corpus bilingües en múltiples pares de lenguas. No tiene un funcionamiento demasiado bueno y hay reportados distintos tipos de problemas. Uno de ellos es la baja de la *performance* cuando el inglés no es ninguna de las lenguas en cuestión, ya que en ese caso la traducción pasa primero por una traducción intermedia al inglés, lo que suma errores en el proceso.

2007: Primera versión del corpus Ancora (Taulé, 2008), corpus del español y del catalán, con unas 500.000 palabras cada uno,

anotados sintácticamente. Este ha sido de gran relevancia para la comunidad de habla hispana, ya que se utilizó para generar analizadores sintácticos abiertos. El corpus está anotado en formato de constituyentes, pero se tradujo al formato de dependencias y se dispone de analizadores en ese formato.

2008: Algoritmos para análisis sintáctico por dependencias, Joachin Nivre (2008). El análisis de dependencias utiliza un corpus anotado y un algoritmo de aprendizaje que aprende la próxima elección óptima, en un análisis basado en memoria que toma decisiones locales. El análisis de dependencias pasa a ser un estándar, y disponiendo de un corpus anotado es muy sencillo generar analizadores sintácticos para distintos lenguajes.

Segunda década del siglo XXI

2011: Apple lanza Siri, una asistente virtual basada en reconocimiento del habla y la comprensión del lenguaje natural.

2013: Se publica el modelo Word2Vec de Google, que permite la representación de palabras en vectores de alta dimensión.

2014: El modelo seq2seq *(sequence-to-sequence)* de Google introduce el uso de redes neuronales recurrentes para traducción automática.

2017: Se presenta el modelo Transformer en el artículo «Attention is All You Need», revolucionando el PLN con el uso de mecanismos de atención.

2018: Open AI y Google desarrollan modelos de lenguaje de gran escala como GPT (Generative Pre-trained Transformer) y BERT (Bidirectional Encoder Representations from Transformers).

Tercera década del siglo XXI

2020: Open AI lanza GPT-3, un modelo de lenguaje con 175 mil millones de parámetros, demostrando capacidades avanzadas en generación de texto y comprensión del lenguaje natural.

2021: Se lanzan modelos como T5 (Text-To-Text Transfer Transformer) y BART (Bidirectional and Auto-Regressive Transformers) que mejoran aún más las tareas de generación y comprensión del lenguaje natural.

2023: Open AI lanza GPT-4, mejorando significativamente las capacidades de generación de texto y comprensión en una amplia variedad de tareas de PLN.

2024: Conviven distintos modelos de lenguaje, de distintas empresas, en la expectativa tecnológica dominante del momento. Se proponen diversas técnicas: *prompting, chain of thought, fine tuning* para tratar de explotar de modo seguro las capacidades de los LLM; pasan a ser multimodales (texto e imagen). Además de texto común, se genera texto en lenguajes de programación. O sea, se generan programas, los modelos pasan a ser parte integrante del proceso de desarrollo de *software.*

El paradigma actual: grandes modelos de lenguaje

En poco más de setenta años se desarrollaron innovaciones prodigiosas que cambiaron el estilo de vida de gran parte de la humanidad. La tecnología nos permite vivir integrados, con posibilidad real de comunicarnos instantáneamente con cualquier parte de la tierra y con acceso a una enorme cantidad de información centralizada, en formato texto, audio, video, en diversos idiomas o con posibilidad casi inmediata de hacer las traducciones requeridas.

El estado del arte actual en PLN pasa por disponer de herramientas informáticas con casi increíble capacidad de generar lenguaje fluido. Están disponibles en gran variedad de idiomas; la construcción de una herramienta de este tipo para una lengua dada requiere de una gran cantidad de texto en esa lengua. El nombre original en inglés es *Large Language Models,* que traduciríamos literalmente por grandes modelos de lenguaje.

Son las grandes estrellas del *boom* de 2020 de la IA, aunque no necesariamente están presentes en todas las facilidades o aplicaciones que se publicitan como inteligencia artificial incorporada al producto. Considerando su diseño técnico, son modelos de lenguaje basados en redes neuronales, de tipo generativo. La referencia al tamaño tiene que ver con la red neuronal: cuántas variables, cuántos parámetros.

Los primeros modelos probabilísticos del lenguaje fueron los modelos de n-gramas. Los n-gramas son secuencias de n palabras consecutivas en un texto. Ha sido usual trabajar con bigramas y

trigramas, o sea, secuencias de dos o de tres palabras. Estos modelos expresan las probabilidades de tiras de dos o tres palabras como la probabilidad condicional de una palabra dada la palabra previa o las dos previas. Se cumple, además, que conociendo la probabilidad de secuencias de dos o tres palabras, y haciendo una hipótesis de independencia respecto a palabras previas a la anterior, o a las dos anteriores, según el caso, podemos calcular la probabilidad de una oración. Esto puede ser útil, por ejemplo, en traducción. Entre varias opciones de oración traducida en el lenguaje objetivo, podemos elegir la más probable. Es interesante notar que este criterio es «observable»: ante una duda, podría exponer, por ejemplo, las mejores opciones entre las que eligió el algoritmo. Los modelos de n-gramas, con probabilidades computadas sobre corpus grandes, permitieron un avance muy grande en la traducción automática.

Los modelos de lenguaje instancian la hipótesis distribucional que enuncia que conocemos una palabra por los contextos en los que aparece. En la práctica, lo que se hace es asignar probabilidades a palabras considerando todas sus ocurrencias en muy grandes corpus de textos. El esquema algorítmico es el de redes neuronales multicapa (se habla de *deep learning* o redes profundas) con un muy gran número de parámetros, del orden de miles de millones, y con una arquitectura que permite cómputos en paralelo y considerar porciones del texto relativamente alejadas en los cálculos. Por ejemplo, la última versión de Chat GPT, GPT-4, tiene más de 200 mil millones de parámetros. Sin entrar en el detalle del funcionamiento interno de la red, es claro que no es posible analizar los cálculos internos para intentar explicar la computación de la red ante una entrada determinada. Dado que los LLM alternan entre resultados muy buenos, incluso para tareas para las que no fueron entrenados, y errores evidentes (a veces dan información falsa, inventada).

Algunas características de los LLM:

- Son costosos en procesadores y en cantidad de texto de entrada requerido para entrenar.
- Presentan conducta de caja negra: no dan explicaciones, no pueden justificar porque llegan a determinadas conclusiones. No mencionan fuentes de las que obtienen la información.
- Pueden presentar «alucinaciones». Directamente, inventan información fáctica.

- Pueden presentar sesgos raciales, de género o de otro tipo. Ejemplos de esto es asociar hombre con médico y mujer con enfermero en una situación en la que hay un hombre y una mujer, y se sabe que alguien ejerce la medicina y alguien ejerce la enfermería. Esto corresponde claramente a sesgos que se presentan en los datos consumidos en el entrenamiento.

Habiendo marcado todos los puntos débiles, es necesario reconocer el tremendo poder y capacidad de estos modelos. La fluidez en el lenguaje, tan arduamente perseguida en los setenta años de historia del PLN, se logra en estos modelos. Son como una especie de cápsula que contiene el lenguaje, y dado un LLM de una lengua determinada se pueden desarrollar muy diversas aplicaciones. Se están desarrollando estrategias para integrarlos a diversos tipos de aplicaciones. Estas estrategias pasan por «instruir» específicamente al modelo sobre la tarea a realizar, ya sea sin presentar ejemplos (*zero-shot)* o presentando algunos ejemplos *(few-shot).* Es muy importante el tamaño del contexto que se puede manejar en la interacción con el LLM, ya que este contexto se puede usar para transferir ejemplos o datos propios de los que extraer una respuesta. Todos estos puntos están en pleno desarrollo y cambio.

Referencias bibliográficas

Association for Computational Linguistics, 2024. Recuperado de <https://www.aclweb.org/portal/>.

Baker, Collin; Charles J. Fillmore, y John B. Lowe. «The Berkeley FrameNet Project», en *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,* vol. 1. Montreal, Quebec: Association for Computational Linguistics, 1998, pp. 86-90.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer y Paul S. Roossin. «A Statistical Approach to Machine Translation», en *Computational Linguistics,* vol. 16, n.º 2, 1990, pp. 79-85.

Campbell, Murray, Joseph Hoane y Feng-hsiung Hsu. «Deep Blue», en *Artificial Intelligence,* vol. 134, n.os 1-2, 2002, pp. 57-83. Recuperado de <https://www.sciencedirect.com/science/article/pii/S0004370201001291?via%3Dihub>.

Chomsky, Noam. «Three models for the description of language», en *IRE Transactions on Information Theory,* vol. 2, n.º 3, 1956, pp. 113-124.

—. *Aspects of the Theory of Syntax*. Cambridge: MIT Press, 1965.

Fellbaum, Christiane. *WordNet: An Electronic Lexical Database.* Cambridge: MIT Press, 1998.

—. WordNet and wordnets. En Brown, Keith et al. (Eds.), *Encyclopedia of Language and Linguistics.* Oxford: Elsevier, 2005, pp. 665-670.

Fillmore, Charles. «Frame Semantics», en *Linguistics in the Morning Calm.* Seúl: Hanshin Publishing Company, 1982.

Google. *AlphaGo,* 2024. Recuperado de <https://deepmind.google/technologies/alphago/>.

—. «MetNet-3: A state-of-the-art neural weather model available in Google products». *Google Research,* 2023. Recuperado de <https://research.google/blog/metnet-3-a-state-of-the-art-neural-weather-model-available-in-google-products/>.

Hutchins, John. «ALPAC, the (in)famous report», en *MT News International*, n.º 14, 1996, pp. 9-12. Recuperado de <https://web.archive.org/web/20071006133016/http://www.hutchinsweb.me.uk/ALPAC-1996.pdf>,

—. «The Georgetown-IBM experiment demonstrated in January 1954», en *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas: Technical Papers,* Washington, 2004, pp. 102-114. Recuperado de <https://link.springer.com/chapter/10.1007/978-3-540-30194-3_12>.

ibm. ‹Deep Blue», en *IBM*, 2024. Recuperado de <https://www.ibm.com/history/deep-blue>.

Labatut, Benjamin. *Maniac*. Barcelona: Anagrama, 2023.

Nivre, Joakim. (2008). «Algorithms for Deterministic Incremental Dependency Parsing», en *Computational Linguistics,* vol. 34, n.º 4, 2008, pp. 513-553.

Open ai. Chat gpt (gpt-4) [Large language Model], 2024. <https://chat.openai.com/>

Rosenblatt, Frank. «The perceptron: a probabilistic model for information storage and organization in the brain», en *Psychological Review,* vol. 65, n.º 6, 1958, pp. 386-408.

Rumelhart, David E., Geoffrey E. Hinton y Ronald J. Williams. «Learning representations by back-propagating errors», en *Nature,* n.º 323, 1986, pp. 533-536.

Saphra, Naomi, Eve Flesig, Kyunghyun Cho y Adam Lopez. «First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models», en *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* Ciudad de México, 2024, pp. 2310-2326. Recuperado de <https://aclanthology.org/2024.naacl-long.128>.

Taulé, Mariona, M. Antònia Martí y Marta Recasens. «AnCora: Multilevel Annotated Corpora for Catalan and Spanish», en *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (lrec'08). Marrakech: European Language Resources Association (ELRA), 2008. Recuperado de <http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf>.

Thompson, Clive. «What the history of AI tells us abouts its future». *MIT Technology Review,* 2022. Recuperado de <https://www.technologyreview.com/2022/02/18/1044709/ibm-deep-blue-ai-history/>.

Turing, Alan M. «On Computable Numbers, with an Application to the Entscheidungsproblem», en *Proceedings of the London Mathematical Society 2,* vol. 42, n.os 1, 1937, pp. 230-265. <https://doi.org/10.1112/plms/s2-42.1.230>.

—. «Computing Machinery and Intelligence», en *Mind,* n.os 49, 1950, pp. 433-460.

Woods, William. «Lunar rocks in natural English: Explorations in natural language question answering», en Zampolli A. (Ed.), *Linguistic Structures Processing.* Ámsterdam: North-Holland, 1977, pp. 521-569.

Weizenbaum, Joseph. «eliza, a computer program for the study of natural language communication between manand machine», en *Communications of the ACM,* vol. 9, n.º 1, Cambridge, 1966, pp. 36-45.