ELENA GONZÁLEZ-BLANCO

Inteligencia artificial y tecnologías del lenguaje

EN EL CORAZÓN DE LA ENCRUCIJADA

La inteligencia artificial es una etiqueta extensamente utilizada hoy día que abarca numerosos y complejos conceptos. Este artículo explica cómo una de las partes más significativas de esta disciplina son las tecnologías del lenguaje, que han avanzado notablemente en los últimos años si bien aún están lejos de superar la capacidad comunicativa del ser humano. Repasamos los principales avances y retos en este ámbito, así como su impacto en el futuro cercano en numerosas aplicaciones.

Palabras clave:
inteligencia artificial, tecnologías del lenguaje, procesamiento del lenguaje natural, asistentes virtuales, *machine learning*

Artificial intelligence and language technologies
IN THE HEART OF THE CROSSROADS

Artificial intelligence is a label widely used today that includes many complex concepts. This article explains how one of the most significant parts of this subject are the language technologies, that have advance notably in recent years although they are still far from surpassing the communicative ability of the human being. We review the main advances and challenges in this field, as well as their impact in the near future in several applications.

Keywords: *artificial intelligence, language technologies, natural language processing, virtual assistants, machine learning*

La inteligencia artificial es hoy día una etiqueta compleja que tiende a generalizarse para aludir a cualquier capacidad mediante la que las máquinas pueden realizar tareas consideradas propiamente humanas. Se trata de un sintagma poliédrico en el que se suman fenómenos como el reconocimiento, la transcripción y la reproducción de la voz humana y los sonidos, del procesamiento del lenguaje y su generación, de la visión artificial y del reconocimiento automatizado de imágenes o de la voz, entre otros. A todos ellos se añaden conceptos que salpican los artículos e informes de últimas tendencias, como el *machine learning* o aprendizaje automático, y el *deep learning* o aprendizaje profundo basado en la emulación de los sistemas del cerebro y la construcción de redes neuronales.

Por lo general, estos términos confluyen mezclados con los datos a través múltiples formas: *big data*, *data analytics*, *data visualization*, *data science* y *business intelligence*, que hacen dificultoso, en muchos casos, vislumbrar sus usos y posibles aplicaciones en el mundo real y especialmente aclarar todos estos conceptos a la sociedad, poco versada en tecnología pero muy preocupada por cuestiones éticas, regulatorias, de privacidad y de transformación digital en todos los ámbitos.

Aunque el concepto de inteligencia artificial suena novedoso y reciente, la realidad es que las tecnologías subyacentes a esta etiqueta llevan ya más de seis décadas en desarrollo. Hemos de remitirnos a 1950 para recordar la figura que realmente transformó este primer escenario de la programación hacia la inteligencia: Alan Turing, que en su artículo *Computing Machinery and Intelligence*[1], planteaba la pregunta: "¿Pueden las máquinas hablar como los hombres?".

Este fue el origen del llamado *Test de Turing*, que sentó las bases del juego de la imitación de la máquina al hombre y pretendía analizar cuándo la máquina confundiría al ser humano emulando sus capacidades lingüísticas. Sobre los avances realizados en los años 50 de la mano de científicos como Marvin Minsky —fundador del Laboratorio de Inteligencia Artificial del Massachusetts Institute of Technology (MIT)— se han construido y mejorado muchos de los algoritmos que hoy día están en la base de nuestros sistemas de procesamiento de datos.

Sin embargo, la historia de la inteligencia artificial y de la tecnología aplicada al lenguaje ha estado llena de altibajos y momentos de auge: tras la revolución de Turing, los años dorados de la primera etapa se extendieron hasta 1975, época en la que los sistemas de procesamiento se basaban en algoritmos de reglas fundamentados sobre lógica, para pasar después a un invierno de silencio provocado por los límites del *hardware*, que volvería a gozar de auge a partir de los años 80 con la introducción del concepto de sistemas expertos, —que combinaban estos primeros algoritmos con bases de datos con las que enlazar y almacenar el conocimiento—.

Un segundo invierno llegó a finales de los años 80 y no ha vuelto a despertar hasta comienzos del presente siglo, pues con el cambio de milenio parece haber resucitado la moda de la inteligencia artificial con un furor que esta vez se queda para cambiar el mundo y no esfumarse más. ¿Por qué?

Realmente no hay una sola razón sino la conjunción de varios factores que hacen que, para muchos, el actual sea el momento propicio de invertir, desarrollar y transformar la industria gracias a la transformación digital.

El primer revulsivo es, indudablemente, la propia tecnología, pues nos encontramos en un momento en que las exponenciales mejoras, tanto a nivel de *software* —potencia y variedad de algoritmos, cantidad de código abierto u *open-source*, comunidades amplias de desarrollo…—, como de *hardware* —creación de máquinas potentes con unidades de procesamiento capaces de asumir la potencia de las multiplicidad de procesos en paralelo que requieren las redes neuronales, como las unidades de procesamiento gráfico o GPU desarrolladas por empresas como Intel o Nvidia—, han hecho posible que los procesos de analítica de datos que antes duraban horas, incluso días, arrojen resultados a tiempo real, utilizando espacios minúsculos y a muy bajo coste al alcance de cualquier usuario y desarrollador.

El segundo factor que ha potenciado este auge tecnológico es la cantidad de datos masivos generados exponencialmente, de los cuales se calcula que un porcentaje de entre el 80 por ciento y el 90 por ciento no están estructurados. En 1992 el tráfico diario mundial de Internet era de 100 Gigabit/día y en 2015 ha pasado a ser de 15.000 millones de GB por día. Para 2020 se esperan alcanzar unos 44 zetabytes de datos diarios y, sin embargo, la realidad es que la mayoría de los datos que se producen no se analizan y los no estructurados —como el lenguaje— no se procesan.

Las tecnologías del Procesamiento del Lenguaje Natural (PLN) podrían, en muchos casos, utilizarse para transformar estos datos no

Nuestra lengua, el español, es uno de los grandes activos que puede ser catalizador de nuestra competitividad en inteligencia artificial

1 **Turing, A.** (1950): "Computing Machinery and Intelligence" en Mind, número 49, páginas 433-460.

estructurados de tipo lingüístico en conocimiento y obtener valor añadido gracias a la clasificación, extracción y entendimiento de la información, que permitirán alcanzar las expectativas que el fenómeno del *big data* comenzó a prometer hace unos años.

Hay que añadir además un tercer factor, que es la proliferación de artefactos digitales, las denominadas "nuevas plataformas IoT (*internet of things* o internet de las cosas)", que permiten a los usuarios interactuar constantemente con sus teléfonos inteligentes (*smartphones*) u ordenadores, pero mediante interfaces que en muchos casos van más allá de las tradicionales pantallas y se activan mediante la voz, como la tecnología ponible (*wearables*).

Todas estas circunstancias han provocado que nos encontremos ante una verdadera revolución de importantes consecuencias, de las que apenas estamos viendo la punta del iceberg, no solamente encaminadas a transformar y mejorar los actuales procesos mediante la reducción de costes y la automatización, sino a crear nuevos modelos de negocio y nuevas líneas de ingresos gracias a la amplitud de posibilidades, la generación de nuevos *insights* y las mejoras en el análisis de datos tanto en volumen como en velocidad a tiempo real.

Tecnologías del lenguaje

Las tecnologías del lenguaje se están convirtiendo, pues, en una de las áreas de mayor potencial dentro de la inteligencia artificial, gracias a su combinación con los sistemas tradicionales de Procesamiento del Lenguaje Natural basado en reglas, con las últimas tecnologías de *machine learning* y *deep learning*.

Los algoritmos de PLN permiten, en primera instancia, lograr que la máquina interprete el texto más allá de una secuencia de caracteres binarios, convirtiéndolos en palabras, mediante procedimientos de *lematización* y *stemming* (agrupación de palabras de una misma raíz eliminando variantes de singular, femenino, tiempos verbales…), detección de estructuras sintácticas y funcionalidad de las palabras en la frase (POS o *Part of Speech*), desambiguación e identificación de referencias anteriores en el texto (en demostrativos, pronombres relativos, etcétera), y clasificación semántica utilizando diccionarios especializados (*wordnets*).

Para que estos funcionen, es necesario acompañarlos de una serie de librerías, gramáticas y diccionarios digitales asociados a cada lengua, que permiten que rápidamente el ordenador pueda codificar los términos existentes en un texto. Así, por ejemplo, una de las operaciones más sencillas y útiles que se realizan utilizando el PLN es la extracción automática de entidades nombradas o NER (*Name Entity Recognition*), que permite extraer entidades de los documentos como nombres propios, lugares geográficos, direcciones de email, números, números de identificación fiscal… sin necesidad de leer el texto en su totalidad.

Entre los diferentes tipos de diccionarios, caben destacar también los denominados "corpus de polaridad", que definen cada una de las palabras asociándolas a su carga semántica positiva o negativa, operación básica para poder detectar el proceso conocido como "análisis de sentimiento" que tan frecuentemente se utiliza en análisis de redes sociales —por ejemplo, para hacer minería de opinión en los debates políticos— o para medir la satisfacción de los clientes tras recibir un servicio. Estos sistemas suelen funcionar por entrenamiento, es decir, combinan las reglas iniciales establecidas con el etiquetado manual de los textos y la iteración sobre los modelos iniciales hasta conseguir un nivel satisfactorio de análisis automatizado.

Además de las reglas, otra de las técnicas que se ha empleado ya desde los años 80 es la aplicación de estadística al análisis lingüístico digital para detectar patrones y realizar inferencias a partir de los mismos. Estos sistemas funcionan en el momento que hay suficientes datos que permitan visibilizar la repetición de patrones mediante técnicas sencillas como el cómputo por frecuencia de palabras. Basados en esta metodología se han creado sistemas de clasificación de textos utilizando *clusters* o grupos por repetición de patrones en las frases.

Las reglas y la estadística se han visto superadas y mejoradas con la irrupción de los sistemas de aprendizaje automático, que también han revolucionado el panorama de las tecnologías lingüísticas, introduciendo conceptos como los *word embeddings* o vectorización de relaciones de palabras en un plano tridimensional, que servirán como base posterior a la construcción de redes neuronales convolucionales (RNN) y a las LSTM, que se basan en modelos semi-supervisados y no supervisados de aprendizaje automático.

La combinación de todas estas tecnologías ha mejorado la exactitud de los algoritmos y la velocidad de procesamiento, pero su adecuación y correcto funcionamiento en cuanto a exactitud de resultados (se mide en porcen-

El español, el segundo idioma más hablado del mundo, solo representa el treinta por ciento de la facturación de tecnologías del lenguaje basadas en PLN

Las tecnologías del lenguaje son una de las áreas de mayor potencial dentro de la inteligencia artificial

taje de aciertos) depende enormemente del contexto en que estos algoritmos se sitúen, del volumen de datos disponible y del entrenamiento en el ámbito en que se haya producido el análisis. Por esta razón, no hay una solución perfecta para el análisis lingüístico por ordenador, sino que los mejores resultados suelen obtenerse mediante la combinación de varias de las operaciones descritas anteriormente unidas a un buen conjunto de datos correctamente entrenado en los sistemas y en la lengua correspondiente, junto a clasificaciones de datos estructurados y modelados que se articulan en torno a ontologías y modelos semánticamente organizados.

Traducción automática

Un caso ilustrativo de esta situación es la evolución de la traducción automática, disciplina especialmente relevante en Europa por la propia idiosincrasia lingüística de nuestro continente. La propia Comisión Europea pasó de los sistemas de reglas basado en ontologías y gramáticas (*Apertium*, por ejemplo) a sistemas estadísticos (*Moses*) para recientemente incorporar el concepto de *Neural Machine Translation* (*Open NMT*) a sus traducciones parlamentarias. Hay que decir, sin embargo, que esta evolución no siempre es la más adecuada para lenguas minoritarias, que siguen funcionando de forma más correcta con reglas y diccionarios al no haber suficiente volumen de datos para entrenarlas.

Nos encontramos ante un momento muy relevante en el que la inteligencia artificial y el lenguaje confluyen en el epicentro del debate, pues la eclosión de interfaces de voz en dispositivos, como los coches autónomos o los asistentes virtuales, ha creado la necesidad de mejora y de incorporación de los mismos a nuestro día a día, cuya funcionalidad es entender el lenguaje humano, procesarlo e interactuar en forma de voz y chatbots y asistentes virtuales como Alexa, Echo, Siri o Cortana.

Pero, ¿para qué sirven realmente todos estos robots que hablan y cómo mejoran o dónde se pueden aplicar realmente estos motores cognitivos o plataformas de inteligencia artificial a nuestros mercados, a las diferentes industrias y, concretamente, a las empresas que trabajan con datos lingüísticos y textuales?

Si pensamos que estos sistemas son capaces de extraer la información más relevante de un texto, de hacer un resumen breve, de agilizar un sistema de preguntas y respuestas frecuentes (*FQ&A*) dentro de una gran base de datos, de interactuar en forma de chat con preguntas y respuestas sobre un conjunto de textos, de detectar y analizar el sentimiento de un cliente que llama a un *call center* y de realizar una labor de vigilancia o escucha activa en un sistema multicanal en el que las redes sociales, las publicaciones en prensa o los datos abiertos públicos se cruzan con los intereses de las industrias en análisis de gran complejidad, nos damos cuenta del gran potencial que pueden alcanzar estos sistemas, que aún apenas están siendo tímidamente aplicados.

Por poner algunos ejemplos concretos de casos de uso ya existentes, encontramos que existen abogados digitales que buscan, analizan y procesan sentencias; sistemas automáticos de clasificación y análisis de historiales médicos en el sector sanitario; de sistemas de generación automática de informes a partir de los datos proporcionados por las redes inteligentes de distribución eléctrica (*smartgrid*); los asesores robóticos (*roboadvisors*) que ya interactúan en lenguaje natural con humanos para las recomendaciones financieras; o asistentes virtuales que permiten agilizar el proceso de consulta de pólizas en las compañías de seguros, o de sistemas que miden y clasifican poemas automáticamente en diferentes lenguas buscando patrones comunes, podemos apreciar el gran potencial que aún nos queda por explorar tecnológicamente.

Una gran brecha

Aunque a día de hoy, el mercado de la inteligencia artificial y específicamente de las tecnologías del lenguaje está dominado por las grandes empresas procedentes del mundo angloparlante —en el que las GAFAs: Google, Amazon, Facebook y Apple se baten con la competencia asiática para ostentar el liderazgo, y esto sin contar la gran competencia que se nos acerca por el sudeste asiático—, observamos que existe una gran brecha entre las soluciones angloparlantes y el resto de las lenguas, en un ámbito en el que en español son comparativamente mucho más débiles —menos de un 30 por ciento de facturación a pesar de ser el segundo idioma más hablado del mundo— y con un mercado muy fragmentado, tanto en empresas tecnológicas como en soluciones específicas aplicadas a la industria.

Nos encontramos, pues, en un momento de oportunidad, en el que España y otros países de América Latina cuentan con grupos científicamente muy potentes que han desarrollado y están desarrollando soluciones competitivas en tecnologías del lenguaje de relevancia mundial, pero cuya implementación aún tendrá que realizar una importante transferencia al tejido empresarial para poder competir con uno de los grandes activos que puede ser el catalizador de nuestra competitividad en el ámbito de la inteligencia artificial: nuestra lengua, el español.

Bibliografía

Beccue, M. y Kaul, A.: "Natural Language ProcessingEnterprise Applications for Natural Language Technologies (Processing, Understanding, Generation) Software and Systems: Market Analysis and Forecasts" en *Tractica Research Report*, 2017.

Jurafsky, D. y Martin, J. (2018): *Speech and Language*. Stanford , Universidad de Stanford. Disponible en https://web.stanford.edu/~jurafsky/slp3

Ransbotham, S.; Gerbert, P.; Reeves, M.; Kiron, D. y Spira, M.: "Artificial Intelligence in Business Gets Real" en *MITSloan Management Review Research Report*, 2018.

Russell, S. y Norvig, P. (2016): *Artificial Intelligence: A Modern Approach*. Londres, Pearson.