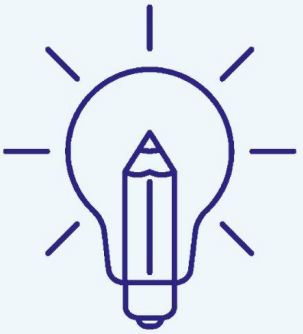
RETOS I+D+i 2022

El español en la inteligencia artificial,
de la tecnología a la aplicación social



La comunidad hispanohablante ronda los 600 millones de personas en todo el mundo. Pese a ello, la mayoría de los sistemas de inteligencia artificial –en especial, los relacionados con el procesamiento del lenguaje natural, como los asistentes virtuales– son entrenados en inglés o en chino. Una vez entrenados en esos idiomas, las grandes tecnológicas simplemente traducen los resultados a las distintas lenguas de los mercados en los que operan.

Con ello, se pierde una riqueza en la comunicación muy importante que no se ve reflejada en estos modelos inteligentes. Por lo pronto, cabe recordar que el inglés posee una variedad gramatical más limitada que, con un menor uso de preposiciones o artículos, lo que puede hacer que los sistemas de inteligencia artificial no ofrezcan la misma precisión al establecer relaciones con un ciudadano en este idioma que en el inglés con que han sido entrenados.

Además, al estar los algoritmos entrenados con personas anglosajonas, tampoco recogen las variantes ni particularidades regionales del español. De hecho, el castellano apenas representa el 30% del mercado mundial de las tecnologías de procesamiento de lenguaje natural.

"Ahora ya no sólo nos enfrentamos al inglés, la lengua predominante hasta ahora; también a otra que cada vez ocupa más espacio: el chino, que es una lengua tipológicamente muy distinta del inglés", advierte **María Victoria Pavón**, investigadora del grupo **Gramática, Léxico, Discurso e Historia** (Departamento de Humanidades - Filosofía, Lenguaje y Literatura), especializada en sintaxis del español.

Puede parecer anecdótico que un asistente virtual cometa un error de vez en cuando por no comprender bien el idioma. Pero lo cierto es que las consecuencias, escaladas a mercados como el 'contact ' o el entretenimiento, pueden ser millonarias en forma de ventas perdidas o clientes insatisfechos. Máxime cuando se tiene a la automatización en el contacto con los consumidores de manera masiva.

Jerónimo Arenas, investigador del **grupo Machine Learning for Data Science (ML4DS)**, del Departamento de Teoría de la Señal y Comunicaciones, así lo corrobora, en referencia a uno de los proyectos en los que trabaja actualmente, IntelComp, una plataforma para ayudar a las políticas públicas en Ciencia, Tecnología e Innovación: *"La aproximación al uso del español es muy pobre porque la inmensa mayoría de los recursos disponibles están en inglés, así que recurrimos a la traducción automática y el core del procesamiento automático del lenguaje va a ser con los textos ya en inglés"*, reconoce.

"Nosotros ya estamos entrenando con diferentes bases de datos en español, como la que recoge las transcripciones de las actas de la Eurocámara", apunta **Belén Ruiz**, del grupo **Human Language and Accessibility Technologies Group (HULAT)** perteneciente al Departamento de Informática.

En ese sentido, en la Estrategia Nacional de Inteligencia Artificial –recogida en el componente 16 del Plan de Recuperación– ya se contempla una partida de 28 millones de euros en tres años para un Plan de Tecnologías de Lenguaje Natural.

El objetivo de este plan, que incluye la creación de un centro de inteligencia artificial en español, es claro: dotar de recursos (principalmente a través de la apertura de bases de datos y corpus lingüísticos) a empresas y startups que pudieran investigar y explotar algoritmos en español y promover nuestro idioma en el ámbito de la IA.

"También es fundamental contar con una buena base de datos bien etiquetada, lingüistas en los equipos de trabajo para poderlo hacer y abordar casos de uso concretos. Por ejemplo, uno de los colectivos olvidados es de las personas con discapacidad intelectual. Aplicar técnicas de lenguaje claro y de simplificación textual para su posterior lectura fácil sería una iniciativa con una gran proyección", defiende Belén Ruiz.

Y sobre este punto de partida, el pasado uno de marzo, el Gobierno de España aprobó el proyecto de recuperación, financiado con fondos europeos (PERTE) ligado al español cuyo objetivo es desarrollar las oportunidades que presenta nuestro idioma como un activo para impulsar la economía. El PERTE 'Nueva Economía de la Lengua' para maximizar el valor del español y las lenguas cooficiales en la transformación digital se estructura en siete objetivos: inteligencia artificial, aprendizaje del español en el mundo, turismo de la lengua, industrias culturales, español global, lenguas cooficiales y ciencia en español.

Sobre este último objetivo, el fomento del español en la ciencia, Pavón señala que *"muchos problemas de traducción se resolverían si en los congresos científicos se hablara en español en un porcentaje más alto de lo que se hace ahora. Detectar las peculiaridades del lenguaje científico es fundamental a la hora de anotar los textos y relevante para extraer la información sobre las relaciones entre las distintas partes del texto o cómo dar prioridad a los contenidos".*

Esto en lo que concierne a la actividad pública, pero lo cierto es que también existe un enorme interés en estas lides por parte del sector privado. Buena prueba de ello es el proyecto LEIA, impulsado por la RAE con el apoyo de las 'big tech' (Telefónica, Google, Amazon, Microsoft, Facebook o Twitter, entre otras).

En este marco, las tecnológicas se han comprometido a utilizar los materiales de la RAE (diccionarios, gramática, ortografía...) en el desarrollo de sus asistentes de voz, procesadores de texto, buscadores, chatbots, sistemas de mensajería instantánea, redes sociales y cualquier otro recurso, así como a seguir los criterios sobre buen uso del idioma aprobados por la Real Academia Española.

El otro gran proyecto ya existente en nuestro país es conocido como MarIA, en esta ocasión fruto de una alianza entre el Barcelona Supercomputing Centre, IBM, el gobierno central y la Biblioteca Nacional. Con sus

modelos está trabajando ya el equipo de investigadores del que forma parte Arenas en un proyecto europeo de contratación pública, para el que tienen en cuenta tanto textos en español como las cooficiales. *"Algunos de los partners que tenemos, como la Generalitat, nos proporcionan textos en catalán y en bilingüe, y aquí sí vamos a trabajar el NLP sobre las lenguas cooficiales. MarIA ahora mismo sólo está en español, pero su idea es incorporar ese conjunto de modelos para las lenguas cooficiales",* adelanta el investigador.

Uno de los problemas con los que se encuentran los investigadores es que las bases de datos en español no están bien anotadas. *"Para encontrar, por ejemplo, la relación entre palabras antes hay que realizar un trabajo manual para luego inferir en un* algoritmo. *Es una tarea que requiere un importante volumen de trabajo que se realiza de forma manual", explica Arenas.*

Eso en lo que atañe al idioma, pero lo cierto es que España y Europa están embarcadas en una regulación mucho más amplia en torno a la inteligencia artificial que siente las bases de uso y los límites para las empresas en esta tecnología de vanguardia.

Así pues, en 2021 la Unión Europea presentó su propia propuesta de normativa en materia de inteligencia artificial. La norma, que da respuesta a las limitaciones éticas de esta tecnología y a su futuro desarrollo, es clave en el momento actual: estamos en pleno debate sobre los sesgos y las 'black boxes' en muchos algoritmos, sobre los problemas de la vigilancia masiva y el reconocimiento facial, acerca de la brecha racial aumentada por la tecnología...

Asimismo, Europa necesita extraordinariamente un impulso a su tejido innovador en IA para recuperar empuje frente a China –con el importante motor público– y Estados Unidos –hogar de las todopoderosas hasta el extremo multinacionales de Silicon Valley–.

Finalmente, la norma establecía la prohibición de prácticas como la vigilancia indiscriminada de forma generalizada o los sistemas de puntuación (scoring) realizados a partir de la Inteligencia Artificial, definición y control de los sistemas de IA considerados de "alto riesgo", creación de autoridades nacionales y europeas de supervisión y multas de hasta el 4% de los ingresos anuales para aquellos que no respeten la regulación.

En este sentido, **Jesús R. Mercader**, perteneciente a los grupos de investigación de **Derecho del Trabajo, Cambios Económicos y Nueva Sociedad**, y **Seguridad Social y Prevención de Riesgos Laborales**, *"prácticamente todos los sectores de actividad en los que los trabajos se puedan mecanizar se van a ver afectados, mientras que el efecto va a ser menor en aquellos con un mayor factor de creatividad. No será con carácter inmediato, pero en muchos ámbitos ya estamos asistiendo a esos procesos de robotización y en todos ellos hay que contemplar la variable jurídica".*

De cómo se articule el diseño final de la regulación en materia de inteligencia artificial y cómo se materialicen los actuales proyectos en este terreno dependerá gran parte de las capacidades tecnológicas y la soberanía digital de Europa. Un asunto candente, que todavía requiere de un debate amplio y multidisciplinar, que arroje luz a muchos de los puntos aún no demasiado definidos en el uso de la IA, especialmente en el ámbito laboral, sanitario o desde la perspectiva de la Administración Pública.

Más información de interés para innovar juntos:

Grupos de Investigación participantes en la validación de este reto:

- Gramática, Léxico, Discurso e Historia
- Machine Learning for Data Science (ML4DS)
- Human Language and Accessibility Technologies (HULAT)
- Derecho del Trabajo, Cambios Económicos y Nueva Sociedad
- Seguridad Social y Prevención de Riesgos Laborales

Startups/Spinoffs del programa de Incubación de la UC3M

- Evidence-Based Behavior S.L.
- Aptent soluciones, S.L.