Inteligencia artificial en español: una hoja de ruta global

Adrián González Sánchez

Un asistente al congreso probando una experiencia de RV con los auriculares Meta Oculus Quest 2 en el stand de SK telecom durante el Mobile World Congress (MWC) la mayor feria del sector centrada en dispositivos móviles, 5G, IOT, AI y big data, celebrado, el 3 de marzo de 2022 en Barcelona, España. (Foto de Joan Cros/NurPhoto vía Getty Images

El futuro de la IA en español pasa por la creación de plataformas y espacios colaborativos de lenguas, para llevar a cabo acciones sincronizadas con impacto escalable a largo plazo, en el que la suma de las iniciativas en su conjunto sea superior al valor de las mismas por separado.

Imaginen un mundo en el que las tecnologías de inteligencia artificial (IA) más avanzadas puedan pensar y comunicar en español, con el mismo nivel de calidad y rendimiento que cualquier otra aplicación en el predominante idioma inglés, y adaptadas a la diversidad y belleza del lenguaje en todas sus variantes y componentes geográficas dentro del contexto hispanohablante. Imaginen un ecosistema de innovación que permita una colaboración directa

entre entidades públicas, grandes empresas, universidades y *startups* innovadoras. Imaginen un gran *cerebro* conectado que combine capacidades de computación avanzadas, modelos lingüísticos, e ingentes cantidades de datos en forma de corpus de referencia que permitan continuar avanzando la investigación e innovación tanto a nivel académico como en los entornos profesionales.

Las tecnologías del lenguaje español

El lenguaje, como elemento clave en cualquier sociedad e instrumento de comunicación e interacción entre personas, ha guiado el desarrollo tecnológico de las últimas décadas para facilitar la creación de herramientas equipadas con habilidades lingüísticas similares a las de los seres humanos. Por ejemplo: buscadores de Internet que entienden nuestra intención, asistentes personales que nos escuchan, servicio al cliente automatizado, autocorrectores de escritura… todos ellos basados en tecnologías que, en mayor o menor medida, comienzan a estar disponibles a gran escala y en varios idiomas. Una suerte de IA lingüística que impacta el desarrollo de nuevas soluciones innovadoras.

Lejos quedan los tiempos en los que dicha innovación estaba exclusivamente ligada al inglés —como consecuencia directa del origen geográfico de los primeros y mayores centros de innovación—, pero no es menos cierto que la predominancia del idioma de Shakespeare continúa presente, y las tecnologías ligadas a lenguas como el español o el francés siguen su camino para desarrollarse y desplegarse a gran escala.

Prueba de la importancia del lenguaje español y su desarrollo tecnológico es el denominado PERTE (Proyectos Estratégicos para la Recuperación y Transformación Económica) Nueva economía de la lengua en España, un proyecto estratégico aprobado en 2022 dentro del Plan de Recuperación, Transformación y Resiliencia, el cual canaliza los fondos europeos para la recuperación en la era pospandemia. No sólo sitúa al castellano y las lenguas cooficiales (catalán, gallego, euskera y valenciano) como una de las prioridades del Estado español, sino que confirma la importancia de las mismas, su relación con las tecnologías de inteligencia artificial y de procesamiento del lenguaje natural (PLN), y el potencial económico de una IA en español fuerte y madura.

Por supuesto, y siendo el idioma español un activo clave y nexo de unión a ambos lados del Atlántico (sin olvidar otros países con tradición hispanohablante como Guinea Ecuatorial o Filipinas), el potencial de una IA en español tiene alcance internacional, especialmente alimentado por la riqueza y variedad de vocabulario, acentos, expresiones, literatura, etcétera. Sin ir más lejos, una de las primeras innovaciones de modelos lingüísticos en español fue BETO , el cual no surgió en España sino que fue desarrollado por la Universidad de Chile en 2019

como respuesta a los predominantes en inglés.

Innovación y un plan global

Por supuesto, no todo son modelos cuando hablamos de una inteligencia artificial en español. La innovación de la IA combina varios factores de éxito y las tecnologías orientadas a la lengua no son una excepción. En el caso de la IA en español, dicha innovación se ha basado en una serie de acciones espontáneas y asíncronas de entidades públicas, privadas, y de la comunidad técnica. Concretamente, las acciones se han enfocado en la creación de los primeros corpus (datos de referencia), modelos de inteligencia artificial y, por supuesto, grandes recursos de computación para manejar esa magnitud de datos y los modelos que los consumen.

Si tuviéramos que imaginar la hoja de ruta pragmática para lograr una evolución exponencial de la IA en español, podríamos imaginar un plan en tres fases:

Inicios de la IA en español (hasta 2022). Dentro del conjunto de iniciativas espontáneas, se han visto todo tipo de innovaciones loables que marcan el camino para los próximos años.

Concretamente, después del previamente mencionado BETO, surgieron otros modelos en español como RigoBERTa, del Instituto de Ingeniería del Conocimiento (IIC) de la Universidad Autónoma de Madrid y basado en la arquitectura del modelo original ??DeBERTa, o BERTIN, que sigue la arquitectura de RoBERTa. El nombre de estos modelos deriva del original “BERT” (Bidirectional Encoder Representations from Transformers), publicado por Google AI en 2018. Este post en Linkedin de Alejandro Vaca Serrano resume la secuencia de acciones.

Por otro lado, y como no podría ser de otra manera, la Real Academia Española (RAE) se posicionó en 2019 como actor de referencia para contribuir al desarrollo de la IA en español mediante la iniciativa LEIA, la cual surgió con el objetivo de velar por un buen uso del español por parte de humanos y máquinas. A dicha iniciativa se unieron las grandes empresas como Telefónica, Google, Amazon, Microsoft, Twitter y Meta, en una muestra clara del interés suscitado. Además, se ha anunciado recientemente una subvención de 5 millones de euros para seguir impulsando dicho proyecto.

Mención especial merece MarIA, el primer modelo masivo de IA en español basado en el

corpus de la Biblioteca Nacional de España (BNE) y en la potencia de computación del Barcelona Supercomputing Center (BSC). Según los autores, MarIA incluye capacidades de comprensión y generación de texto en español, y está disponible de manera abierta para que pueda ser utilizado por desarrolladores de aplicaciones.

A nivel de comunidad y *datasets* abiertos, cabe destacar el *subset* del conjunto de datos multilingüe "mC4", o el recientemente anunciado "esCorpius", que ha utilizado capacidad en la nube para escanear contenido web en español mediante técnicas de *crawling*.

Otras tecnologías como la famosa OpenAI GPT-3, propulsada por Elon Musk con un fuerte apoyo de Microsoft, los servicios lingüísticos de la nube y librerías de las grandes empresas (Microsoft, Google, Amazon, IBM, y Meta), o tecnologías de generación de texto como BLOOM, han incorporado soporte en castellano con un rendimiento significativo, y esto ilustra la importancia de la lengua de Cervantes (como no podría ser de otra manera, dada la magnitud del mercado de usuarios hispanohablantes).

Por último, pero no por ello menos importante, otras lenguas cooficiales han seguido una evolución similar con proyectos similares, como el Proyecto AINA en Catalunya, con la campaña "La nostra llengua és la teva veu" que nace con misión de generar datos en catalán (incluyendo todas sus variedades de acentos) y de permitir a las empresas utilizar asistentes de voz, traductores automáticos o agentes conversacionales. Igualmente, Euskadi ha desarrollado traductores neuronales y capacidades de "texto a voz" potentes con sus iniciativas Batua e Itzuli API.

Fase de impulso inicial (entre 2022 y 2024). Ahora bien, imaginemos plataformas y espacios en los que la suma del impacto de las acciones individuales sea exponencialmente mayor que el impacto de dichas acciones por sí solas. Es decir, si creamos modelos o corpus de datos que puedan ser aprovechados por otros muchos actores del ecosistema de innovación, más allá de las barreras empresariales o geográficas.

Con la aparición del previamente mencionado PERTE de la lengua, el cual incentivará la colaboración público-privada a través de la Alianza para la Nueva Economía de la Lengua con el apoyo de instituciones como el Instituto Cervantes, la Biblioteca Nacional de España, la

Secretaría General de Estados Iberoamericanos y la Organización de Estados Iberoamericanos, se abre una fase de impulso público inicial en el que los fondos públicos deberán sentar la base para el desarrollo futuro de la IA en español.

Ese contexto único, así como la participación de actores internacionales, suponen una oportunidad para esta hoja de ruta global. Concretamente, la posibilidad de crear una plataforma de colaboración que permita no sólo compartir e intercambiar los elementos de la fase 1 (datos, modelos y capacidad de computación) a través de un modelo tipo repertorio o *marketplace*, sino la capacidad de generar comunidades de práctica entre los actores de innovación que trabajen para desarrollar y adoptar las tecnologías de la IA en español.

Más allá de las opciones clásicas de innovación abierta en las que los datos y los modelos son compartidos de manera pública por el bien mayor de la innovación (una opción que a día de hoy choca frontalmente con los intereses de muchas organizaciones públicas y privadas), hay dos técnicas que pueden suponer un incentivo a la innovación colaborativa sin renunciar a la propiedad intelectual de las empresas: por un lado, las técnicas de APIficación de datos que facilitan el acceso a datos a través de interfaces sencillas de programación, en la que las organizaciones pueden facturar por niveles de acceso e incluso habilitar compartición de datos gratuita bajo ciertas condiciones y, por otro, técnicas de IA federadas, en la que los modelos de IA están disponibles de manera *remota*, eliminando la necesidad de mover datos que además puedan incluir información personal. Proyectos como FLUTE de la rama de investigación de Microsoft habilitan este tipo de innovación.

Escalabilidad ecosistémica (a partir de 2024). Es decir, cómo amplificar el impacto del apoyo público inicial con iniciativas público-privadas en las que primen las “3 Cs” (coordinación, combinación y colaboración).

En este sentido, hay dos factores clave que marcarán el éxito de esta iniciativa, y que permita a los actores del ecosistema aprovechar la base de lo creado en la segunda fase de la hoja de ruta. En primer lugar, la voluntad de colaborar y crear arquitecturas modulares que permitan optimizar esfuerzos, implementar nuevas soluciones de manera pragmática y sobre todo evitar *reinventar la rueda*. Todo ello como una manera de reducir costes y aumentar eficiencias. Y, en segundo lugar, la capacidad para encontrar modelos de cooperación originales en los que las organizaciones puedan obtener valor económico y organizacional, ya sea a través de acuerdos específicos entre pares, o con modelos de contratos en línea que faciliten y automaticen dichos automatismos de cooperación.

La escalabilidad geográfica será clave para generar masa crítica de desarrolladores y usuarios de plataformas que utilicen la IA en español, dado el tamaño del mercado LATAM en

comparación con —relativamente reducido—mercado español. A tal fin, la participación de la Secretaría General de Estados Iberoamericanos y la Organización de Estados Iberoamericanos ayudará a crear los mecanismos y los puentes necesarios para incentivar la colaboración a ambos lados del Atlántico.

Asimismo, aunque la creación de corpus y modelos de los últimos años ha aportado una base muy buena para los primeros pasos de la IA en español, la liberación (tipo código y datos abiertos) de los mismos podría generar un impacto exponencial a nivel de ecosistema. Dada la complejidad de dicho escenario, es posible que los fondos europeos puedan ayudar a incentivar este tipo de prácticas. El pragmatismo para combinar tecnologías en español y en inglés ayudará también a cerrar la brecha de innovación entre ambas lenguas.

En resumen, pese a que el dominio del inglés ha sido la norma a lo largo de los años de la innovación tecnológica, existe una oportunidad única para poner en valor un activo creciente y relevante como el idioma español. La escalabilidad de una IA en español llegará por la vía público-privada, y solamente mediante un enfoque pragmático, colaborador y sobre todo en el que se optimicen esfuerzos y en el que prime el intercambio de capacidades entre los actores del ecosistema se logrará reducir la distancia con respecto a la IA en inglés, y se generará una verdadera economía de la lengua española y de su inteligencia artificial.

Fecha de creación
26 agosto, 2022