Published on *BSC-CNS* (https://www.bsc.es)

El primer sistema masivo de Inteligencia Artificial de la lengua española, MarIA, empieza a resumir y generar textos

Cinco meses después de su lanzamiento, el sistema expande sus capacidades para utilizar el lenguaje.

MarIA ha sido creado en el Barcelona Supercomputing Center, entrenado con más de 135 mil millones de palabras del archivo web de la Biblioteca Nacional e impulsado por la Secretaría de Estado de Digitalización e Inteligencia Artificial, dentro de los objetivos de la Estrategia Nacional de Inteligencia Artificial y del Plan de Recuperación

Por el volumen y capacidades de MarIA, la lengua española se sitúa en el tercer puesto de los idiomas que disponen de modelos masivos de acceso abierto, después del inglés y el mandarín

Se publica en abierto para que los desarrolladores de aplicaciones lo puedan utilizar en infinidad de usos

Las aplicaciones creativas y empresariales, y aquellas relacionadas con la digitalización de la Administración Pública aumentan

El proyecto MarIA, el sistema de modelos de lengua creado en el Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS), a partir de los archivos web de la Biblioteca Nacional de España (BNE) y enmarcado y financiando con el Plan de Tecnologías del Lenguaje de la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA), ha avanzado en su desarrollo y su nueva versión permite resumir textos existentes y crear nuevos textos a partir de titulares o de palabras.

El proyecto MarIA es el primer sistema de inteligencia artificial masivo y experto en comprender y escribir en lengua española. Por su volumen y capacidades, ha situado a la lengua española en el tercer puesto de los idiomas que disponen de modelos masivos de acceso abierto, después del inglés y el mandarín. Se ha construido a partir del patrimonio documental digital de la Biblioteca Nacional de España, que rastrea y archiva las webs elaboradas en español y se ha entrenado con el superordenador MareNostrum 4. Y se publica en abierto para que los desarrolladores de aplicaciones, compañías, grupos de investigación y la sociedad en general lo puedan utilizar en infinidad de usos.

Los últimos avances de MarIA constituyen un hito en la consecución de objetivos de la Estrategia Nacional de Inteligencia Artificial y del Plan de Recuperación, Transformación y Resiliencia, con los que España pretende liderar a nivel mundial el desarrollo de herramientas, tecnologías y aplicaciones para la proyección y uso de la lengua española en los ámbitos de aplicación de la IA. En concreto, el Plan Nacional de Tecnologías del Lenguaje en el que se enmarca este proyecto tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales en lengua española y lenguas coficiales.

Modelos para comprender la lengua y modelos para generar textos

Un modelo de lenguaje es un sistema de inteligencia artificial formado por conjunto de redes neuronales profundas que han sido entrenadas para adquirir una comprensión de la lengua, de su léxico y de sus mecanismos para expresar el significado y escribir a nivel experto. Estos modelos estadísticos complejos, que relacionan palabras en textos de modo sistemático y masivo, son capaces de “entender” no sólo conceptos abstractos, sino también el contexto de los mismos. Con estos modelos, los desarrolladores de diferentes aplicaciones pueden crear herramientas para múltiples usos, como clasificar documentos o crear correctores o herramientas de traducción.

La primera versión de MarIA fue elaborada con RoBERTa, una tecnología que crea modelos del lenguaje del tipo “codificadores”. Este tipo de modelos, dada una secuencia de texto, generan una interpretación que puede servir para, por ejemplo, clasificar documentos, responder a preguntas tipo test, encontrar similitudes semánticas en diferentes redactados o detectar los sentimientos que se expresan en ellos.

La nueva versión ha sido creada con GPT-2, una tecnología más avanzada que crea modelos generativos decodificadores y añade prestaciones al sistema. Los modelos decodificadores, dada una secuencia de texto, pueden generar nuevos textos. Con ello, pueden servir, por ejemplo, para hacer resúmenes automáticos, simplificar redactados complicados a la medida de diferentes perfiles de usuario, generar preguntas y respuestas, mantener diálogos complejos con los usuarios e incluso redactar textos completos (que podrían parecer escritos por humanos) a partir de un titular o de un pequeño número de palabras.

Estas nuevas capacidades convierten a MarIA en una herramienta que, con entrenamientos “ad hoc” adaptados a tareas específicas, puede ser de gran utilidad para desarrolladores de aplicaciones, empresas y administraciones públicas. Por ejemplo, los modelos que hasta ahora se han desarrollado en inglés se utilizan para generar sugerencias de texto en aplicaciones de escritura, para resumir contratos o los complicados documentos que detallan las prestaciones de un producto, en función de lo que quiere saber cada usuario, y para buscar informaciones concretas dentro de grandes bases de datos de texto y relacionarlas con otras informaciones relevantes.

“Con proyectos como MarIA, que se verán incorporados al PERTE para el desarrollo de una economía digital en español,’ damos pasos firmes hacia una inteligencia artificial que piense en español, lo que multiplicará las oportunidades económicas para las empresas y la industria tecnológica española. Porque la lengua es mucho más que un medio de comunicación. Es una proyección de la forma que tenemos de ver el mundo, también en la nueva realidad digital”, señala la secretaria de Estado de Digitalización e Inteligencia Artificial, Carme Artigas.

“Como institución responsable del depósito legal electrónico, la Biblioteca Nacional de España (BNE) conserva millones de sitios web, millones de palabras que se repiten en un contexto determinado y que son producto de muchas recolecciones de la web española, tanto de dominio.es como selectivas, realizadas desde hace años por los equipos de la BNE, lo que conforma el gran corpus del español que hoy se habla en nuestro país — Explica Ana Santos, directora de la BNE—. Para nosotros es una gran satisfacción que estos archivos resulten de utilidad para este proyecto pionero, basado en tecnologías de inteligencia artificial, que va a permitir que las máquinas puedan comprender y escribir en lengua española, lo que supone un hito en el campo del procesamiento del lenguaje natural”.

“Agradecemos la iniciativa de la SEDIA de impulsar temas de futuro, como la potenciación del idioma español en el mundo digital y el entorno de la IA — afirma el director del BSC-CNS, Mateo Valero—. Estamos encantados de poner nuestros expertos en lenguaje natural e inteligencia artificial y la capacidad de cálculo de nuestras infraestructuras al servicio de los retos relevantes para la sociedad, como la que da respuesta esta iniciativa”.

Por su parte, la directora de la División de Procesos y Servicios Digitales de la BNE, Mar Pérez Morillo, ha destacado que “*en las recolecciones ponemos el foco en eventos que han influido o marcado la sociedad y su lenguaje*”. Igualmente, la BNE coopera de forma activa con los centros de recopilación autonómicos que utilizan las herramientas que la BNE pone a su disposición. “*Llevamos una carrera contra el tiempo, desarrollando estrategias y herramientas que luchen contra la que llaman la edad oscura digital*”, ha explicado Morillo.

Entrenada con más de 135 mil millones de palabras y 9,7 trillones de operaciones

En los modelos del lenguaje, el número de parámetros con los que se entrena el sistema es el elemento que les aporta mayor capacidad de generalización y, por tanto, inteligencia. Los datos de la Biblioteca Nacional con los que se ha entrenado MarIA están constituidos por más de 135 mil millones de palabras (135.733.450.668, concretamente), que ocupan un total de 570 Gigabytes.

Para crear y entrenar a MarIA se ha utilizado el superordenador MareNostrum del BSC y ha sido necesaria una potencia de cálculo de 9,7 trillones de operaciones (969.exaflops). Un flop (operación de coma flotante) es la unidad de medida con que se expresa la capacidad de cálculo de un superordenador por segundo y exa es el prefijo que expresa 10^{18}, es decir, un trillón.

De estos 969 exaflops, 201 fueron necesarios para procesar los datos procedentes de la Biblioteca Nacional, eliminar todo aquello que no fuera texto bien formado (números de páginas, gráficos, oraciones que no terminan, codificaciones erróneas, oraciones duplicadas, otros idiomas, etc.) y guardar solamente los textos correctos en lengua española, tal y como es realmente utilizada. Los restantes 768 exaflops se utilizaron para entrenar las redes neuronales del modelo GPT-2.

La versión actual de MarIA dará ahora lugar a versiones especializadas en distintas áreas de aplicación, incluyendo biomedicina y legal, y evolucionará para resolver los problemas específicos mencionados anteriormente.

En paralelo el PlanTL continuara expandiendo MarIA para: adaptarse a los nuevos desarrollos tecnológicos en procesamiento del lenguaje natural (modelos más complejos que el GP-T2 ahora implementado) entrenados con mayor cantidad de datos, crear espacios de trabajo para facilitar el uso de MarIA por compañías y grupos de investigación en los entornos computaciones adecuados y embeberlos en sistemas de evaluación y certificación de la calidad de los sistemas desarrollados en distintos dominios.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---