Los desarrolladores de aplicaciones ya disponen de un sistema de inteligencia artificial experto en comprender y escribir la lengua española

El modelo ha sido creado en el BSC-CNS y se ha entrenado en el superordenador MareNostrum con archivos de datos de la Biblioteca Nacional.

El proyecto se ha financiado con fondos del Plan de Tecnologías del Lenguaje del Ministerio de Asuntos Económicos y Agenda Digital y del Future Computing Center, una iniciativa del BSC-CNS e IBM.

MarIA, que es el nombre del sistema, está disponible en abierto para que cualquier desarrollador, empresa o entidad pueda utilizarlo sin coste. Sus posibles aplicaciones van desde los correctores o predictores del lenguaje, hasta las aplicaciones de resúmenes automáticos, chatbots, búsquedas inteligentes, motores de traducción y subtitulación automática, entre otros. Los ficheros de datos que han servido para entrenar a MarIA no están en dominio público y por lo tanto no están accesibles en internet. Son los WARC resultantes del rastreo y archivado de la web española, que la Biblioteca Nacional de España conserva, en virtud de la ley de depósito legal, como patrimonio documental. El BSC-CNS ha podido utilizarlos para entrenar al sistema gracias a la participación de ambas instituciones en el Plan de Tecnologías del Lenguaje.

El primer modelo de IA masivo de la lengua española

MarIA es un conjunto de modelos del lenguaje o, dicho de otro modo, redes neuronales profundas que han sido entrenadas para adquirir una comprensión de la lengua, su léxico y sus mecanismos para expresar el significado y escribir a nivel experto. Logran trabajar con interdependencias cortas y largas y son capaces de entender, no sólo conceptos abstractos, sino también el contexto de los mismos.

El primer paso para crear un modelo de la lengua es elaborar un corpus de palabras y frases que será la base sobre la que se entrenará el sistema.

Para crear el corpus de MarIA, se utilizaron 59 terabytes (equivale a 59.000 gigabytes) del archivo web de la Biblioteca Nacional. Posteriormente, estos archivos se procesaron para 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 se guardaron solamente los textos bien formados en la lengua española, tal y como es realmente utilizada. Para este cribado y su posterior compilación fueron necesarias 6.910.000 horas de procesadores del superordenador MareNostrum y los resultados fueron 201.080.084 documentos limpios que ocupan un total de 570 gigabytes de texto limpio y sin duplicidades.

Este corpus supera en varias órdenes de magnitud el tamaño y la calidad de los corpus disponibles en la actualidad. Se trata de un corpus que enriquecerá el patrimonio digital del español y del propio archivo de la BNE y que podrá servir para múltiples aplicaciones en el futuro, como tener una imagen temporal que permita analizar la evolución de la lengua, comprender la sociedad digital en su conjunto y, por supuesto, el entreno de nuevos modelos.

Una vez creado el corpus, los investigadores del BSC-CNS utilizaron una tecnología de redes neuronales (basada en la arquitectura Transformer), que ha demostrado excelentes resultados en el inglés y que se entrenó para aprender a utilizar la lengua. Las redes neuronales multicapa son una tecnología de Inteligencia Artificial y los entrenamientos consisten, entre otras técnicas, en presentar a la red textos con palabras ocultas, para que aprenda a adivinar cuál es la palabra ocultada dado su contexto.

Para este entrenamiento han sido necesarias 184.000 horas de procesador y más de 18.000 horas de GPU. Los modelos liberados hasta ahora tienen 125 millones y 355 millones de parámetros respectivamente.

Marta Villegas, responsable del proyecto y líder del grupo de minería de textos del BSC-CNS, explica la importancia de poder implementar las nuevas tecnologías de Inteligencia Artificial, “que están transformando completamente el campo del procesamiento del lenguaje natural. Con este proyecto contribuimos a que el país se incorpore a esta revolución científico-técnica y se posicione como actor de pleno derecho en el tratamiento computacional del español”.

Por su parte, Alfonso Valencia, director del departamento de Ciencias de la Vida del BSC-CNS, argumenta que “la infraestructura de Computación de Altas Prestaciones del BSC-CNS ha demostrado ser esencial para este tipo de grandes proyectos que requieren tanto de mucha computación como de grandes cantidades de datos. Para nosotros, es muy satisfactorio poner capacidades técnicas y conocimiento experto al servicio de un proyecto con tantas repercusiones para la posición del español en la sociedad digital”.

La Biblioteca Nacional de España, como establece su ley reguladora, tiene entre sus funciones “impulsar y apoyar programas de investigación tendentes a la generación de conocimiento sobre sus colecciones, estableciendo espacios de diálogo con centros de investigación”. Con este proyecto, enmarcado en el Plan de Tecnologías del Lenguaje, la BNE explora nuevas vías de explotación de los datos y las colecciones que conserva, y busca impulsar la reutilización, nuevos proyectos de investigación y mejorar el acceso de los ciudadanos a la información.

Próximos pasos

Después de lanzar los modelos generales, el equipo minería de textos del BSC-CNS está trabajando en la ampliación del corpus, con nuevas fuentes de archivos que aportarán textos con particularidades diferentes a los que se encuentran en los entornos web, como por ejemplo publicaciones científicas del CSIC.

También está prevista la generación de modelos entrenados con textos de diferentes lenguas: castellano, catalán, gallego, euskera, portugués y español de Hispanoamérica.

El BSC y el Plan-TL

El BSC-CNS es la oficina técnica del Plan de las Tecnologías del Lenguaje (Plan-TL) de la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA). Como tal, su misión es facilitar el desarrollo de sistemas del lenguaje más competitivos a la sociedad, compañías y grupos de investigación, haciendo públicos modelos de lenguaje tanto generales como específicos -para dominios como la biomedicina o la legal- y liberando conjuntos de texto para entrenar y evaluar nuevos modelos.

Información del Plan-TL: https://plantl.mineco.gob.es/Paginas/index.aspx

Modelo RoBERTa-base: https://huggingface.co/BSC-TeMU/roberta-base-bne

Modelo RoBERTa-large: https://huggingface.co/BSC-TeMU/roberta-large-bne

Repositorio de información: https://github.com/PlanTL-SANIDAD/lm-spanish

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---