MareNostrum generará un modelo del lenguaje en español a partir de millones de contenidos digitales de la Biblioteca Nacional de España

La generación de modelos de lenguaje es vital para integrar el conocimiento lingüístico y del mundo a la inteligencia artificial.

El proyecto forma parte del encargo de la Secretaría de Estado para el Avance Digital al BSC, en el marco del Plan de Impulso de las Tecnologías del Lenguaje

El supercomputador MareNostrum ya ha empezado a recibir la ingente cantidad de datos provenientes del Archivo Web de la Biblioteca Nacional de España y que será la base para generar un modelo del lenguaje del español y de otras lenguas del estado. El Archivo de la Web Española es la colección formada por los sitios web con dominio .es (incluidos blogs, foros, documentos, imágenes, vídeos, etc.) más todos aquellos considerados patrimonio documental incluidos en otros dominios que se recolectan con el fin de preservar el patrimonio documental español en Internet y asegurar el acceso al mismo. El responsable de realizar esta tarea es el Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC) por encargo de la Secretaría de Estado para el Avance Digital (SEAD), en el marco del Plan de Impulso de las Tecnologías del Lenguaje.

La labor encargada al BSC es doble: el transporte de los datos al supercomputador y su procesado para generar el modelo del lenguaje. Desde hace unos meses MareNostrum ha iniciado el almacenaje de los contenidos, tras el desarrollo de un proceso de extracción de los datos textuales del archivo web de la biblioteca, de modo que ha sido posible transferir los contenidos rápidamente al BSC. Y es que el transporte de esta ingente cantidad de datos suponía uno de los principales retos de la iniciativa. En estos momentos el supercomputador tiene almacenado 45 terabytes.

El siguiente paso será el procesamiento de estos datos para generar modelos del lenguaje a través de las tecnologías del procesamiento del lenguaje natural. Este recurso ya existe para el inglés, siendo el más conocido Google Bert, que ha supuesto un antes y un después en el procesamiento del lenguaje natural. El modelo en el que trabaja el BSC destaca de otras iniciativas de modelos del español por la cantidad, calidad y variedad de los datos, lo que hace que sea más preciso y de uso más transversal.

Los modelos del lenguaje y la inteligencia artificial

Los modelos del lenguaje reproducen el uso de la lengua y permiten conocer el significado real de las palabras, incluso de las frases enteras, ya que los datos están contextualizados y tienen más información, más sentido. Esto permite desambiguar el sentido de las palabras (por ejemplo, distinguir el sentido de *brutal* en *un brutal asesinato* y *la serie te gustará. Es brutal*). También permite interpretar el sesgo ideológico, y abre la puerta a abordar la ironía, el sentido figurado y enriquecer los sistemas de inteligencia artificial con sentido común.

Quim Moré, investigador del departamento de CASE del BSC, y David Vicente, jefe de equipo del grupo de Operaciones, son los responsables de este proyecto en el centro. Quim Moré asegura que *"la generación de modelos de lenguaje es vital para la inteligencia artificial. La aplicación computacional de un modelo del lenguaje desambiguado y con un contexto fundamentado en nuestro conocimiento del mundo supone un gran avance en la generación de sistemas cada vez más inteligentes y, a la vez, más cercanos."*

Las aplicaciones de este modelo son múltiples, desde la traducción automática, a la ciberseguridad, hasta la descripción del contenido de un cuadro del siglo XV hecha por un robot. Ahora bien, modelos capaces de generar esta revolución requieren de unos recursos computacionales y de datos que sólo unas pocos centros y compañías, como Google o Facebook, tienen.

En este sentido, Moré destaca que *"tenemos la gran suerte de tener en el MareNostrum la capacidad computacional necesaria y, por otro lado, tenemos la ingente cantidad de datos lingüísticos revisados y de calidad aportados por la Biblioteca Nacional. Tenemos una oportunidad importantísima de estar al nivel de los grandes centros de inteligencia artificial y de aportar una aplicación computacional del conocimiento lingüístico a la cultura"*.

El Archivo de la Web Española

El Archivo de la Web Española es la colección formada por los sitios web con dominio .es y otros (incluidos blogs, foros, documentos, imágenes, vídeos, etc.) que se recolectan con el fin de preservar el patrimonio documental español en Internet y asegurar el acceso al mismo. En diciembre de 2019 se cumplieron 10 años del lanzamiento del proyecto de archivado de la web española. Desde entonces, la Biblioteca Nacional de España ha consolidado su infraestructura, las políticas y los procesos para llevar a cabo esta tarea de preservación del patrimonio en línea, como llevan haciendo desde hace años las bibliotecas nacionales más importantes del mundo.

Más información aquí.

Ver vídeo de la Jornada con motivo del 10º aniversario del Archivo de la Web Española:
https://www.youtube.com/watch?v=oySUYJdiDwY&feature=youtu.be

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---