MareNostrum generarà un model del llenguatge en espanyol a partir de milions de continguts digitals de la Biblioteca Nacional d?Espanya

La generació de models de llenguatge és vital per integrar el coneixement lingüístic i del món a la intel·ligència artificial.

El projecte forma part de l'encàrrec de la Secretaria d'Estat de Digitalització i Intel·ligència Artificial al BSC, en el marc de el Pla d'Impuls de les Tecnologies del Llenguatge

El supercomputador MareNostrum ja ha començat a rebre la ingent quantitat de dades provinents de l'Arxiu Web de la Biblioteca Nacional d'Espanya i que serà la base per generar un model del llenguatge de l'espanyol i d'altres llengües de l'estat. L'Arxiu de la Web Espanyola és la col·lecció formada pels llocs web amb domini .es (inclosos blocs, fòrums, documents, imatges, vídeos, etc.) més tots aquells considerats patrimoni documental inclosos en altres dominis que es recullen amb la finalitat de preservar el patrimoni documental espanyol a internet i assegurar l'accés a aquest. El responsable de realitzar aquesta tasca és el Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC) per encàrrec de la Secretaria d'Estat de Digitalització i Intel·ligència Artificial (SEDIA), en el marc del Pla d'Impuls de les Tecnologies del Llenguatge.

La tasca encarregada al BSC és doble: el transport de les dades al supercomputador i el seu processament per generar el model del llenguatge. Des de fa uns mesos MareNostrum ha iniciat l'emmagatzematge dels continguts, després del desenvolupament d'un procés d'extracció de les dades textuals de l'arxiu web de la biblioteca, de manera que ha estat possible transferir els continguts ràpidament al BSC. I és que el transport d'aquesta ingent quantitat de dades suposava un dels principals reptes de la iniciativa. En aquests moments el supercomputador té emmagatzemats 45 terabytes.

El següent pas serà el processament d'aquestes dades per generar models del llenguatge a través de les tecnologies del processament del llenguatge natural. Aquest recurs ja existeix per a l'anglès, sent el més conegut Google Bert, que ha suposat un abans i un després en el processament del llenguatge natural. El model en el qual treballa el BSC destaca d'altres iniciatives de models de l'espanyol per la quantitat, qualitat i varietat de les dades, el que fa que sigui més precís i d'ús més transversal.

Els models del llenguatge i la intel·ligència artificial

Els models del llenguatge reprodueixen l'ús de la llengua i permeten conèixer el significat real de les paraules, fins i tot de les frases senceres, ja que les dades estan contextualitzats i tenen més informació, més sentit. Això permet desambiguar el sentit de les paraules (per exemple, distingir el sentit de brutal en un brutal assassinat i la sèrie t'agradarà. És brutal). També permet interpretar el biaix ideològic i obre la porta a abordar la ironia, el sentit figurat i enriquir els sistemes d'intel·ligència artificial amb sentit comú.

Quim Moré, investigador de departament de CASE del BSC, i David Vicente, cap d'equip del grup d'Operacions, són els responsables d'aquest projecte en el centre. Quim Moré assegura que "la generació de models de llenguatge és vital per a la intel·ligència artificial. L'aplicació computacional d'un model del llenguatge desambiguat i amb un context fonamentat en el nostre coneixement del món suposa un gran avanç en la generació de sistemes cada vegada més intel·ligents i, alhora, més propers". Les aplicacions d'aquest model són múltiples, des de la traducció automàtica, a la ciberseguretat, fins a la descripció del contingut d'un quadre de segle XV feta per un robot. Ara bé, models capaços de generar aquesta revolució requereixen d'uns recursos computacionals i de dades que només uns pocs centres i companyies, com Google o Facebook, tenen.

En aquest sentit, Moré destaca que "tenim la gran sort de tenir a MareNostrum la capacitat computacional necessària i, d'altra banda, tenim la ingent quantitat de dades lingüístiques revisats i de qualitat aportats per la Biblioteca Nacional. Tenim una oportunitat importantíssima d'estar al nivell dels grans centres d'intel·ligència artificial i d'aportar una aplicació computacional del coneixement lingüístic a la cultura ".

L'Arxiu de la web Espanyola

L'Arxiu de la web Espanyola és la col·lecció formada pels llocs web amb domini .es i altres (inclosos blocs, fòrums, documents, imatges, vídeos, etc.) que es recullen amb la finalitat de preservar el patrimoni documental espanyol a Internet i assegurar el accés a aquest. Al desembre de 2019 es van complir 10 anys del llançament del projecte d'arxivat de la web espanyola. Des de llavors, la Biblioteca Nacional d'Espanya

ha consolidat la seva infraestructura, les polítiques i els processos per dur a terme aquesta tasca de preservació del patrimoni en línia, com porten fent des de fa anys les biblioteques nacionals més importants del món.

Més informació aquí.

Veure vídeo de la Jornada amb motiu del 10 aniversari de l'Arxiu de la web Espanyola: https://www.youtube.com/watch?v=oySUYJdiDwY&feature=youtu.be

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---

Source URL (retrieved on *24 abr 2024 - 13:37*): https://www.bsc.es/ca/noticies/noticies-del-bsc/marenostrum-generar%C3%A0-un-model-del-llenguatge-en-espanyol-partir-de-milions-de-continguts-digitals-de