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

El primer sistema massiu d´Intel·ligència Artificial de la llengua espanyola, MarIA, comença a resumir i generar textos

Cinc mesos després del llançament, el sistema expandeix les seves capacitats per utilitzar el llenguatge.

MarIA ha estat creat al Barcelona Supercomputing Center, entrenat amb més de 135 mil milions de paraules de l'arxiu web de la Biblioteca Nacional i impulsat per la Secretaria d'Estat de Digitalització i Intel·ligència Artificial, dins dels objectius de l'Estratègia Nacional d'Intel·ligència Artificial i del Pla de Recuperació

Pel volum i les capacitats de MarIA, la llengua espanyola se situa al tercer lloc dels idiomes que disposen de models massius d'accés obert, després de l'anglès i del mandarí

Es publica en obert perquè els desenvolupadors d'aplicacions puguin utilitzar-lo en infinitat d'usos

Les aplicacions creatives i empresarials i les relacionades amb la digitalització de l'Administració pública augmenten

El projecte MarIA, el sistema de models de llengua creat al Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS), a partir dels arxius web de la Biblioteca Nacional d'Espanya (BNE) i emmarcat i finançant amb el Pla de Tecnologies del Llenguatge de la Secretaria d'Estat de Digitalització i Intel·ligència Artificial (SEDIA), ha avançat en el desenvolupament i la nova versió permet resumir textos existents i crear-ne de nous a partir de titulars o de paraules.

El projecte MarIA és el primer sistema d’intel·ligència artificial massiu i expert a comprendre i escriure en llengua espanyola. Pel seu volum i capacitats, ha situat la llengua espanyola al tercer lloc dels idiomes que disposen de models massius d'accés obert, després de l'anglès i el mandarí. S'ha construït a partir del patrimoni documental digital de la Biblioteca Nacional d'Espanya, que rastreja i arxiva les webs elaborades en espanyol, i s'ha entrenat amb el superordinador MareNostrum 4. Es publica en obert perquè els desenvolupadors d'aplicacions, companyies, grups de recerca i la societat en general el puguin fer servir en infinitat d'usos.

Els darrers avanços de MarIA constitueixen una fita en la consecució d'objectius de l'Estratègia Nacional d'Intel·ligència Artificial i del Pla de Recuperació, Transformació i Resiliència, amb què Espanya pretén liderar a nivell mundial el desenvolupament d'eines, tecnologies i aplicacions per a la projecció i ús de la llengua espanyola als àmbits d'aplicació de la IA. En concret, el Pla Nacional de Tecnologies del Llenguatge en què s'emmarca aquest projecte té com a objectiu fomentar el desenvolupament del processament del llenguatge natural, la traducció automàtica i els sistemes conversacionals en llengua espanyola i llengües cooficials.

Models per comprendre la llengua i models per generar textos

Un model de llenguatge és un sistema d'intel·ligència artificial format per conjunt de xarxes neuronals profundes que han estat entrenades per adquirir una comprensió de la llengua, el seu lèxic i els seus mecanismes per expressar el significat i escriure a nivell expert. Aquests models estadístics complexos que relacionen paraules en textos de manera sistemàtica i massiva són capaços d’entendre” no només conceptes abstractes, sinó també el context dels mateixos. Amb aquests models, els desenvolupadors de diferents aplicacions poden crear eines per a múltiples usos, com ara classificar documents o crear correctors o eines de traducció.

La primera versió de MarIA va ser elaborada amb RoBERTa, una tecnologia que crea models del llenguatge del tipus “codificadors”. Aquest tipus de models, atesa una seqüència de text, generen una interpretació que pot servir per, per exemple, classificar documents, respondre preguntes tipus test, trobar similituds semàntiques en diferents redactats o detectar els sentiments que s'hi expressen.

La nova versió ha estat creada amb GPT-2, una tecnologia més avançada que crea models generatius descodificadors i afegeix prestacions al sistema. Els models descodificadors, donada una seqüència de text, poden generar nous textos. Amb això, poden servir, per exemple, per fer resums automàtics, simplificar redactats complicats a mida de diferents perfils d'usuari, generar preguntes i respostes, mantenir diàlegs complexos amb els usuaris i fins i tot redactar textos complets (que podrien semblar escrits per humans), a partir d'un titular o un petit nombre de paraules.

Aquestes noves capacitats converteixen MarIA en una eina que, amb entrenaments ad hoc adaptats a tasques específiques, pot ser de gran utilitat per a desenvolupadors d'aplicacions, empreses i administracions públiques. Per exemple, els models que s'han desenvolupat fins ara en anglès s'utilitzen per generar suggeriments de text en aplicacions d'escriptura, per resumir contractes o els complicats documents que detallen les prestacions d'un producte, en funció del que vol saber cada usuari, i per cercar informacions concretes dins de grans bases de dades de text i relacionar-les amb altres informacions rellevants.

“Amb projectes com MarIA, que es veuran incorporats al 'PERTE per al desenvolupament d'una economia digital en espanyol,' fem passos ferms cap a una intel·ligència artificial que pensi en espanyol, cosa que multiplicarà les oportunitats econòmiques per a les empreses i la indústria tecnològica espanyola. Perquè la llengua és molt més que un mitjà de comunicació. És una projecció de la manera que tenim de veure el món, també a la nova realitat digital”, assenyala la secretària d'Estat de Digitalització i Intel·ligència Artificial, Carme Artigas.

“Com a institució responsable del dipòsit legal electrònic, la Biblioteca Nacional d'Espanya (BNE) conserva milions de llocs web, milions de paraules que es repeteixen en un context determinat i que són producte de moltes recol·leccions del web espanyol, tant de domini.es com a selectives, realitzades des de fa anys pels equips de la BNE, cosa que conforma el gran corpus de l'espanyol que avui es parla al nostre país —explica Ana Santos, directora de la BNE—. Per a nosaltres és una gran satisfacció que aquests arxius resultin d'utilitat per a aquest projecte pioner, basat en tecnologies d'intel·ligència artificial, que permetrà que les màquines puguin comprendre i escriure en llengua espanyola, fet que suposa una fita en el camp del processament del llenguatge natural”.

“Agraïm la iniciativa de la SEDIA ??d'impulsar temes de futur, com ara la potenciació de l'idioma espanyol al món digital i l'entorn de la IA —afirma el director del BSC-CNS, Mateo Valero—. Estem encantats de posar els nostres experts en llenguatge natural i intel·ligència artificial i la capacitat de càlcul de les nostres infraestructures al servei dels reptes rellevants per a la societat, com al que dóna resposta aquesta iniciativa”.

Per la seva banda, la directora de la Divisió de Processos i Serveis Digitals de la BNE, Mar Pérez Morillo, ha destacat que “*a les recol·leccions posem el focus en esdeveniments que han influït o marcat la societat i el seu llenguatge*”. Igualment, la BNE coopera de manera activa amb els centres de recopilació autonòmics que utilitzen les eines que la BNE posa a la vostra disposició. "*Portem una carrera contra el temps, desenvolupant estratègies i eines que lluitin contra la que anomenen l'edat fosca digital*", ha explicat Morillo.

Entrenada amb més de 135 mil milions de paraules i 9,7 trilions d'operacions

En els models del llenguatge, el nombre de paràmetres amb què s'entrena el sistema és l'element que els aporta més capacitat de generalització i, per tant, intel·ligència. Les dades de la Biblioteca Nacional amb què s'ha entrenat MarIA estan constituïdes per més de 135 mil milions de paraules (135.733.450.668, concretament), que ocupen un total de 570 Gigabytes.

Per crear i entrenar MarIA s'ha utilitzat el superordinador MareNostrum del BSC i ha calgut una potència de càlcul de 9,7 trilions d'operacions (969.exaflops). Un flop (operació de coma flotant) és la unitat de mesura amb què s'expressa la capacitat de càlcul d'un superordinador per segon i exa és el prefix que expressa 10^{18} , és a dir, un trilió.

D'aquests 969 exaflops, 201 van ser necessaris per processar les dades procedents de la Biblioteca Nacional, eliminar tot allò que no fos text ben format (números de pàgines, gràfics, oracions que no acaben, codificacions errònies, oracions duplicades, altres idiomes, etc.) i guardar només els textos correctes en llengua espanyola, tal com és realment utilitzada. La resta de 768 exaflops es van utilitzar per entrenar les xarxes neuronals del model GPT-2.

La versió actual de MarIA donarà ara lloc a versions especialitzades en diferents àrees d'aplicació, incloent-hi biomedicina i legal, i evolucionarà per resoldre els problemes específics esmentats anteriorment.

En paral·lel, el PlanTL continuarà expandint MarIA per: adaptar-se als nous desenvolupaments tecnològics en processament del llenguatge natural (models més complexos que el GP-T2 ara implementat) entrenats amb més quantitat de dades, crear espais de treball per facilitar l'ús de lMarIA per companyies i grups de recerca als entorns computacions adequats i embeure'ls en sistemes d'avaluació i certificació de la qualitat dels sistemes desenvolupats en diferents dominis.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---