Els desenvolupadors d'aplicacions ja disposen d'un sistema d'intel·ligència artificial expert en comprendre i escriure la llengua espanyola

El model ha estat creat al BSC-CNS i s'ha entrenat al superordinador MareNostrum amb arxius de dades de la Biblioteca Nacional de España (BNE).

El projecte s'ha finançat amb fons del Pla de Tecnologies del Llenguatge del Ministeri d'Afers Econòmics i Agenda Digital i del Future Computing Center, una iniciativa del BSC i IBM.

MarIA, que és el nom de sistema, està disponible en obert perquè qualsevol desenvolupador, empresa o entitat pugui utilitzar-lo sense cost. Les seves possibles aplicacions van des dels correctors o predictors del llenguatge, fins a les aplicacions de resums automàtics, *chatbots*, recerques intel·ligents, motors de traducció i subtitulació automàtica, entre d'altres. Els fitxers de dades que han servit per entrenar MarIA no estan en domini públic i per tant no estan accessibles a internet. Són els WARC resultants de l'rastreig i arxivat de la web espanyola, que la Biblioteca Nacional de España conserva, en virtut de la llei de dipòsit legal, com a patrimoni documental. El BSC ha pogut utilitzar-los per entrenar el sistema gràcies a la participació de les dues institucions en el Pla de Tecnologies del Llenguatge.

El primer model d'IA massiu de la llengua espanyola

MarIA és un conjunt de models del llenguatge o, dit d'una altra manera, 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. Aconsegueixen treballar amb interdependències curtes i llargues i són capaços d'entendre, no només conceptes abstractes, sinó també el seu context.

El primer pas per crear un model de la llengua és elaborar un corpus de paraules i frases, que serà la base sobre la qual s'entrenarà el sistema.

Per crear el corpus de MarIA, es van utilitzar 59 terabytes (equival a 59.000 gigabytes) de l'arxiu web de la Biblioteca Nacional. Posteriorment, aquests arxius es van processar per eliminar tot allò que no fos text ben format (com números de pàgines, gràfics, oracions que no acaben, codificacions errònies, oracions duplicades, altres idiomes, etc.) i es van guardar només els textos ben formats en la llengua espanyola, tal com és realment utilitzada. Per a aquest cribratge i la seva posterior compilació van ser necessàries 6.910.000 hores de processadors del superordinador MareNostrum i els resultats van ser 201.080.084 documents nets, que ocupen un total de 570 gigabytes de text net i sense duplicitats.

Aquest corpus supera en diverses ordres de magnitud la mida i la qualitat dels corpus disponibles en l'actualitat. Es tracta d'un corpus que enriquirà el patrimoni digital de l'espanyol i de l'arxiu de la BNE i que podrà servir per a múltiples aplicacions en el futur, com tenir una imatge temporal que permeti analitzar l'evolució de la llengua, comprendre la societat digital en el seu conjunt i, per descomptat, l'entrenament de nous models.

Un cop creat el corpus, els investigadors del BSC van utilitzar una tecnologia de xarxes neuronals (basada en l'arquitectura Transformer), que ha demostrat excel·lents resultats en l'anglès i que es va entrenar per aprendre a utilitzar la llengua. Les xarxes neuronals multicapa són una tecnologia d'intel·ligència artificial i els entrenaments consisteixen, entre d'altres tècniques, a presentar a la xarxa textos amb paraules ocultes, perquè aprengui a endevinar quina és la paraula amagada donat el seu context.

Per a aquest entrenament han estat necessàries 184.000 hores de processador i més de 18.000 hores de GPU. Els models alliberats fins ara tenen 125 milions i 355 milions de paràmetres respectivament.

Marta Villegas, responsable de el projecte i líder del grup de mineria de textos del BSC-CNS, explica la importància de poder implementar les noves tecnologies d'intel·ligència artificial, "que estan transformant completament el camp del processament del llenguatge natural. Amb aquest projecte contribuïm al fet que el país s'incorpori a aquesta revolució científico-tècnica i es posicioni com a actor de ple dret en el tractament computacional de l'espanyol ".

Per la seva banda, Alfonso Valencia, director del departament de Ciències de la Vida del BSC-CNS, argumenta que "la infraestructura de Computació d'Altes Prestacions del BSC ha demostrat ser essencial per a aquest tipus de grans projectes que necessiten tant molta computació com grans quantitats de dades. Per a nosaltres, és molt satisfactori posar capacitats tècniques i coneixement expert al servei d'un projecte amb tantes repercussions per a la posició de l'espanyol en la societat digital ".

La Biblioteca Nacional de España, com estableix la seva llei reguladora, té entre les seves funcions "impulsar i donar suport a programes d'investigació tendents a la generació de coneixement sobre les seves col·leccions, establint espais de diàleg amb centres de recerca". Amb aquest projecte, emmarcat en el Pla de Tecnologies del Llenguatge, la BNE explora noves vies d'explotació de les dades i les col·leccions que conserva, i busca impulsar-ne la reutilització, nous projectes de recerca i millorar l'accés dels ciutadans a la informació.

Següents passos

Després de publicar els models generals, l'equip de mineria de textos del BSC està treballant en l'ampliació del corpus, amb noves fonts d'arxius que aportaran textos amb particularitats diferents als que es troben en els entorns web, com ara publicacions científiques del CSIC.

També està prevista la generació de models entrenats amb textos de diferents llengües: castellà, català, gallec, euskera, portuguès i espanyol de Hispanoamèrica.

El BSC i el Pla-TL

El BSC és l'oficina tècnica del Pla de les Tecnologies del Llenguatge (Pla-TL) de la Secretaria d'Estat de Digitalització i Intel·ligència Artificial (SEDIA). Com a tal, la seva missió és facilitar el desenvolupament de sistemes del llenguatge més competitius a la societat, companyies i grups de recerca, fent públics models de llenguatge tant generals com específics -per a dominis com la biomedicina o la legal- i alliberant conjunts de text per entrenar i avaluar nous models.

Informació del Pla-TL: https://plantl.mineco.gob.es/Paginas/index.aspx

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

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

Repositori d'informació: https://github.com/PlanTL-SANIDAD/lm-spanish

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---