El projecte AINA, la IA i les tecnologies del llenguatge

Marta Villegas Montserrat
Barcelona Supercomputing Center
ORCID: 0000-0003-0711-0029
marta.villegas@bsc.es

Marta Villegas fa més de 25 anys que treballa com a investigadora en el camp del processament del llenguatge natural. Actualment és la responsable de la Unitat de Tecnologies de la Llengua al Barcelona Supercomputing Center - Centro Nacional de Supercomputación, on dirigeix els treballs per al desenvolupament de models lingüístics. La Unitat ha compilat recentment el corpus espanyol i català més gran mai creat i ha desenvolupat models de referència *transformers* que han tingut un gran impacte, tant en el món acadèmic com en la indústria. Coordina el projecte AINA i és responsable de diversos projectes nacionals i europeus.

Resum

Una de les àrees més rellevants de la IA és el processament del llenguatge natural (PLN). En aquest àmbit, tot i que actualment la majoria dels grans models de llenguatge ja són multilingües, hi ha una diferència substancial entre les capacitats dels models pel que fa a l'anglès i a la resta de llengües. En aquest sentit, el projecte AINA té per objectiu desenvolupar la infraestructura necessària per què la inclusió del català a les aplicacions d'IA sigui prou atractiva i viable. Aquest article presenta els objectius del projecte i n'explica les característiques generals.

Paraules clau: *IA; tecnologies del llenguatge; català; PLN*

Abstract

The AINA Project, Artificial Intelligence, and Language Technologies

One of the most relevant areas of AI is Natural Language Processing (NLP). In this area, even though most of the large language models are currently multilingual, there is an important difference between the capabilities of English models and the other languages. Thus, the AINA project aims at developing the necessary infrastructure so that the inclusion of Catalan in AI applications becomes appealing and feasible. This article presents the objectives of the project and explains its main characteristics.

Keywords: *AI; language technologies; Catalan; NLP*

Terminàlia 27 (2023): 80-84 · DOI: 10.2436/20.2503.01.189
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

Una mica de context

En els darrers anys, la intel·ligència artificial (IA) ha estat causant sensació en el món de la tecnologia i el seu impacte en la nostra societat és cada vegada més i més gran. Sense cap mena de dubte, una de les àrees més rellevants de la IA és el processament del llenguatge natural (PLN), una disciplina que preten ensenyar les màquines a entendre i utilitzar el llenguatge humà.

Fins als anys 80, la majoria de sistemes de PLN es basaven en regles. A finals d'aquesta dècada, els algorismes d'aprenentatge automàtic es van fer cada vegada més populars i, aviat, els models estadístics van reemplaçar les regles, cosa que va suposar una millora substancial. La següent revolució va arribar el 2013 amb la introducció de les representacions vectorials de paraules (*embeddings*) com Word2vec,[1] GloVe[2] i FastText.[3] Els *embeddings* encapsulen el «significat» d'una paraula en un vector després de llegir quantitats massives de text i analitzar els contextos en què apareix cada paraula. La idea és que paraules similars tenen contextos similars. Aquests primers models s'entrenaven amb una xarxa neuronal petita i tenien resultats prou bons, però no suficients, ja que cada paraula tenia un únic vector, quan en realitat una paraula pot tenir múltiples significats.

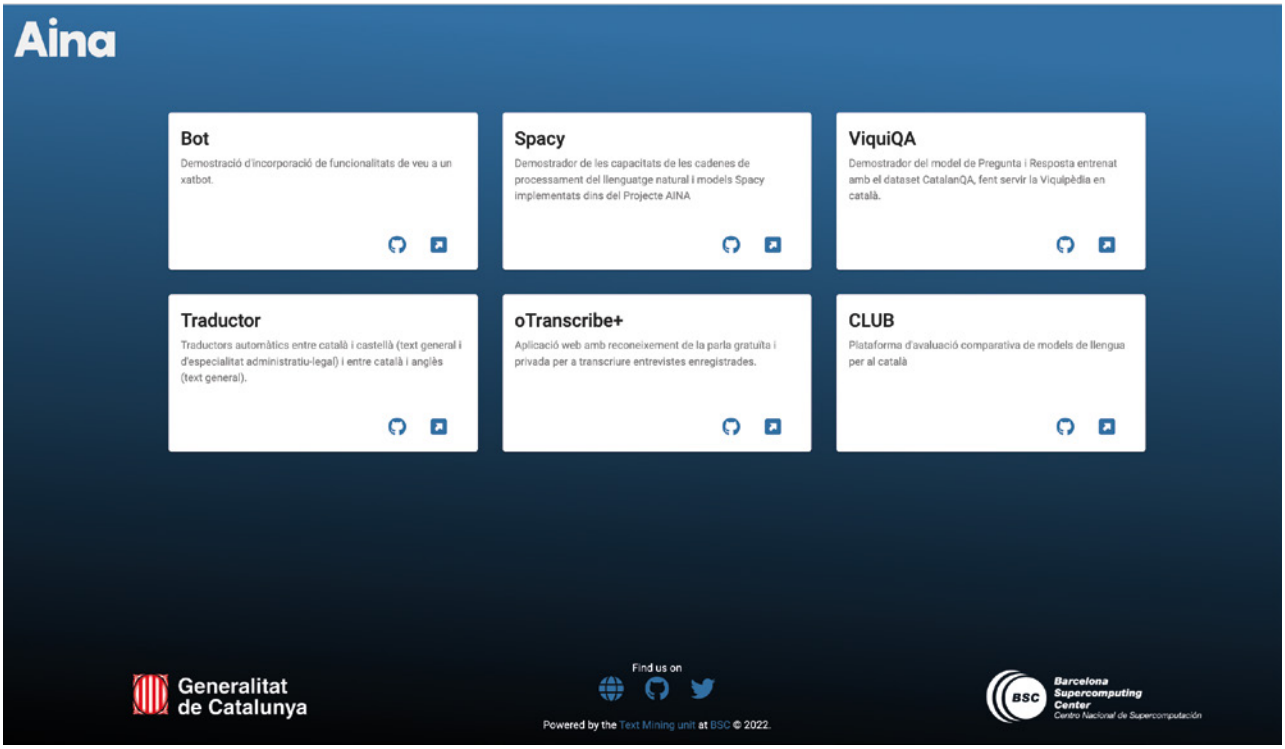
A finals del 2018, Google va publicar els models BERT introduint l'arquitectura *transformers*[4] i provocant una autèntica revolució en el camp de les tecnologies del llenguatge. Des del primer moment, els *transformers* van suposar una millora espectacular en qualsevol de les tasques típiques del processament del llenguatge. Prova d'això són la quantitat i qualitat dels sistemes i aplicacions de PLN que van sorgir, com ara sistemes de traducció automàtica, assistents virtuals com Siri, Alexa o Google Assistant, sistemes de correcció i autocompletat de text, classificadors de documents, anàlisi de sentiments, entre molts d'altres. Google i les grans empreses tecnològiques nord-americanes van començar la cursa dels anomenats grans models de llenguatge (coneguts també com models fundacionals) en què la mida del model (mesurada en paràmetres), la quantitat de dades utilitzades per entrenar-lo i la quantitat de còmput necessari són cada vegada més grans. Cal recordar que per entrenar el model GPT-3 d'OpenAI,[5] que conté 175 mil milions de paràmetres, es van utilitzar 300 mil milions de *tokens* i es van necessitar 3.14e23 *flops*.[6]

Quan parlem de models del llenguatge, ens referim a models que aprenen la probabilitat d'ocurrència de les paraules, basant-se en exemples que han estat vistos. Disposar de bons models és crucial per a les tecnologies de la llengua (TL), ja que permet obtenir millors resultats i rendiments en qualsevol tasca de processament del llenguatge. Com ja hem vist, per a obtenir bons models cal disposar de grans quantitats de dades (*big data*) i capacitat de càlcul per poder processar-les.

Recentment, el xatbot ChatGPT, llançat per l'empresa nord-americana OpenAI al novembre del 2022, ha demostrat l'extrema capacitat disruptiva d'aquesta tecnologia, que canvia paradigmes i és capaç de millorar amb nous llançaments, com el model GPT4, també desenvolupat per OpenAI. La successió de models, arquitectures i propostes no para de créixer; així, al març del 2023, Google va publicar PaLM-E, «un model de llenguatge multimodal personificat».[7] Poc després, Google va publicar Bard, la seva versió d'un xatbot amb IA. Gairebé al mateix temps, el gegant xinès dels motors de cerca Baidu va presentar el seu model Ernie.[8]

Sense cap mena de dubte, els xatbots impulsats per els grans models de llenguatge estan revolucionant la tecnologia de la informació tal com la coneixem, amb implicacions i impacte substancials per a la societat, la investigació i la indústria en general. ChatGPT i tecnologies similars desenvolupades per altres gegants tecnològics nord-americans o asiàtics es poden utilitzar com a assistents d'escriptura, ajudants personals, solucionadors de problemes en general, robots de text i, en general, *sparring* per a tasques, reptes i situacions quotidianes de la nostra vida personal o professional. Poden aplicar-se pràcticament a tots els àmbits, des de l'atenció al client fins a l'assistència sanitària, passant per la mobilitat, l'educació, les finances, les assegurances, el comerç electrònic i molts altres. Des del llançament de ChatGPT, milers de casos d'ús professional han estat ràpidament creats, implementats i validats per tercers, molts ja utilitzats en entorns de producció.

Font: AINA. Polítiques digitals



Font: aina.bsc.es

El cas és que la majoria d'aquests models, inicialment, es van generar per a l'anglès, és a dir, es van entrenar amb dades exclusivamente, o molt majoritàriament, en anglès. Aviat, la necessitat de disposar de models per a altres llengües i la manca de dades suficients en aquestes van propiciar l'aparició dels primers models multilingües, models entrenats amb dades en diferents llengües que mostraven capacitats multilingües.

Tot i que podem afirmar que la majoria dels grans models de llenguatge són multilingües, el cert és que hi ha una diferència substancial entre les capacitats dels models pel que fa a l'anglès i a la resta de llengües. Els models son clarament anglocèntrics senzillament perquè s'han entrenat amb moltes dades en anglès i molt poques en altres llengües. L'escassa presència de la majoria de llengües en aplicacions de PLN provoca que moltes llengües mundials i europees puguin desaparèixer relativament aviat de l'ecosistema digital (primer pas per a la seva extinció en el món real).[9] Això va alertar el Parlament Europeu, que el setembre de 2018 va aprovar una resolució per garantir l'equitat de les llengües en l'era digital.[10]

El projecte AINA

AINA és un projecte impulsat pel Departament d'Empresa i Treball que, amb la col·laboració del Barcelona Supercomputing Center, ha de permetre al català fer un salt qualitatiu i quantitatiu en l'ecosistema digital. AINA neix en un moment especialment delicat, en què els avenços tecnològics en el camp del PLN són alhora un una oportunitat i un repte per al català. D'una banda, les possibilitats que ofereixen les TL són cada vegada més extraordinàries i tenen un impacte més gran en les nostres vides. Però de l'altra, clarament es focalitzen en les llengües globals, deixant la resta en una situació de desavantatge.

AINA, doncs, té per objectiu desenvolupar la infraestructura necessària perquè la inclusió del català a les aplicacions d'IA sigui prou atractiva i viable, tant per a les grans companyies tecnològiques com per a la indústria local. Per tant, el projecte té com a objectiu el desenvolupament de recursos per a la IA/TL a fi d'aconseguir el següent:

- Proveir el català de la infraestructura necessària per al desenvolupament d'aplicacions basades en IA/TL (assistents de veu, traductors automàtics, agents conversacionals, etc.);
- Garantir que la inclusió del català a les aplicacions de IA/TL sigui rendible i atractiva per a les empreses del sector, tant a escala local com global;
- Impulsar el sector de la IA/TL a Catalunya millorant-ne la competitivitat;
- Aconseguir que el ciutadà de Catalunya pugui participar en català en el món digital al mateix nivell que els parlants d'una llengua global;
- Fomentar la investigació capdavantera en IA/TL per tal que Catalunya estigui ben preparada per respondre al repte digital i de la societat basada en el coneixement.

AINA és essencialment infraestructura lingüística en què el valor de les dades és cabdal: la tecnologia avança molt ràpidament, però les dades són persistents. Disposar de dades de qualitat suficients és un actiu segur i de futur que garanteix l'actualització de la tecnologia. D'altra banda, la IA/TL és sens dubte un àmbit en expansió que genera un immens volum de negoci i llocs de treball qualificats. Es calcula que al 2020 el mercat mundial de la IA/TL es va situar en 16.400 milions de dòlars i es preveu que arribi als 59.000 milions el 2027, amb un creixement anual del 20,1%. És, per tant, necessari garantir la competitivitat de Catalunya en aquest àmbit, mitjançant la vigilància tecnològica i sectorial i la recerca més puntera, amb la participació d'empreses i grups de recerca del sector, garantint l'enfortiment de tot l'ecosistema.

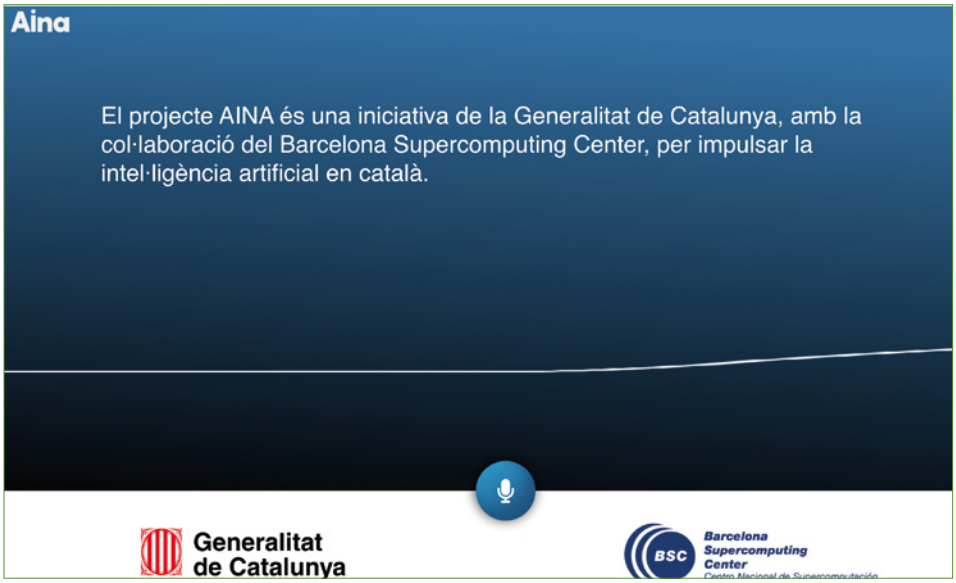
Quan parlem de dades lingüístiques ens referim a qualsevol dada de tipus textual o sonora (o audiovisual) produïda mitjançant el llenguatge humà. Les dades textuals serviran per entrenar models de llengua i desenvolupar aplicacions lingüístiques diverses, incloent-hi motors de traducció automàtica, i les dades de parla permetran crear sistemes de reconeixement i de síntesi de veu. Per tal de garantir el subministrament continu de dades, AINA preveu establir acords que defineixin i implementin protocols de publicació de dades, en compliment de les directives europees sobre dades obertes i reutilització de la informació del sector públic (RISP),[11] i respectant els principis FAI,[12] amb els grans proveïdors de dades. Cal fer les accions necessàries perquè els grans generadors de dades en català es constitueixen en subministradors de dades per a la IA/TL de manera sostinguda i programàtica. Per tal d'aconseguir-ho, s'estan abordant tots els aspectes necessaris incloent-hi aspectes tècnics i legals. Tècnicament, les accions han de permetre la publicació de dades que idealment permetin l'accés programàtic en formats processables. Per això és necessari el desenvolupament d'API o mecanismes de descàrrega/compartició de dades, definició de formats i metadades, etc. El primer exemple és l'accés a les dades de veu i text del Parlament de Catalunya que permet accedir a les dades de text i veu de les sessions del Parlament.

Part important són també les dades anotades; AINA ha invertit esforços considerables en la generació de dades anotades per a diferents tasques i s'està treballant en la generació d'un corpus d'instruccions que ens haurà de permetre «instruir» diferents models generatius.

Un altre aspecte fonamental en la infraestructura d'AINA són els models del llenguatge. Com ja hem vist, els models són una peça clau en el desenvolupament de noves aplicacions. La generació de models utilitzant tècniques d'aprenentatge profund i dades massives és una tasca extremadament costosa en recursos computacionals i humans, que difícilment poden afrontar les empreses o els grups de recerca catalans. Últimament, la recerca i generació de models massius ha quedat en mans de les grans tecnològiques, que, tot i que cada vegada treballen més amb dades multilingües, solen deixar de banda les llengües no globals. Cal, doncs, garantir que l'ecosistema català disposi de models suficients. Així doncs, AINA vol generar models de la llengua en les arquitectures més novedoses i amb més impacte, incloent-hi models monolingües, multilingües i multimodals, prestant especial atenció a la participació en grans projectes multilingües (nacionals i internacionals) i garantint la presència del català en aquests models massius. En aquest sentit, actualment s'està treballant en la generació de tres models GPT3 trilingües (català, castellà i anglès) de diferents mides que després es podran adaptar a tasques i dominis específics.

A AINA, les tecnologies de la parla tenen especial rellevància. L'objectiu de les tecnologies de la parla és la implementació del mitjà més natural d'interacció com a interfície entre humans i màquines. La parla incorpora aspectes que no solem trobar en les formes escrites de la llengua com les variants dialectals i, fins i tot, emocions. La parla també aporta connotacions de trets personals com ara el gènere i l'edat. Per aquestes raons, la incorporació de la parla als productes tecnològics és molt important, tant des del punt de vista cultural com per fomentar l'ús de la llengua en les tecnologies, especialment per a llengües minoritzades com el català. Comparat amb les tecnologies de text, la incorporació de les tecnologies i funcionalitats de la parla té una barrera d'entrada, per la varietat de maneres de parlar, i també per la dificultat suplementària que introdueix el processament del senyal acústic. Per això, és imprescindible un esforç sostingut de captació de veus de tot l'àmbit geogràfic català. AINA ja va engegar la campanya de Common Voice al Principat, que s'ha estès a les Illes Balears i properament s'estendrà al País Valencià gràcies a la col·laboració del Projecte Vives, amb què el català ha passat a ser la segona llengua en nombre d'hores a la plataforma de referència. La generació de models de síntesi de la parla, el reconeixement de la parla, i la traducció automàtica parla a parla és una prioritat del projecte.

Les tecnologies de la traducció automàtica tenen també un paper clau a AINA. La traducció automàtica permet salvar les barreres idiomàtiques en entorns multilingües, tot garantint l'ús de la llengua pròpia. Això la converteix en una tecnologia clau per a la salvaguarda de les llengües minoritzades i de la diversitat lingüística. Entre altres funcionalitats essencials, destaquem el suport a la traducció documental, la traducció de continguts web, la subtitulació automàtica o la traducció en temps real. Per tal de dotar el català de les eines tecnològiques que permetin assolir aquestes funcionalitats AINA es proposa construir i posar a dis-



Font: bot.aina.bsc.es

posició de la comunitat models de traducció basats en els últims avenços tecnològics entre el català i llengües d'interès, ja sigui per motiu de negoci (llengües europees, xinès, etc.) o proximitat geogràfica i social (aranès, llengües de la península, llengües de la immigració). Donada l'escassetat de dades existents per a la majoria d'aquests parells de llengües, està prevista la recerca i desenvolupament de tècniques d'entrenament no supervisat més adequades per a la traducció automàtica. Aquesta recerca és potencialment beneficiosa per a un gran nombre de llengües del món i contribuirà a posicionar la recerca catalana en aquest àmbit.

Com en altres casos, les tecnologies de traducció desenvolupades per AINA permetran disposar de solucions tecnològiques pròpies, independents de les grans tecnològiques (p.ex. Google Translate), tot garantint la sobirania tecnològica i el control de l'accés a les dades, així com assegurar la presència del català dins de les plataformes més populars de programari lliure. Cal emfatitzar que tots els recursos generats per AINA es publiquen en obert i amb llicències el més permissives possible.

Finalment, un dels objectius d'AINA és posicionar Catalunya com a referent en l'àmbit de la IA/TL. El mercat de la IA/TL està en plena expansió. L'augment de la taxa d'adopció de la IA en els diversos sectors industrials i la demanda creixent de tecnologies com traducció i parla, així com l'augment de les àrees d'aplicació (p.ex. detecció de desinformació) impulsen el creixement del mercat. Catalunya, amb un ecosistema de recerca i teixit industrial propici, té el potencial per liderar la recerca i el desenvolupament en TL de llengües mitjanes i petites. Així doncs, AINA ha d'identificar els avenços de la tecnologia i respondre-hi mitjançant la vigilància tecnològica i identificar noves necessitats de les empreses i de la societat mitjançant la vigilància sectorial i de mercat. L'objectiu últim és propiciar l'adopció de les tecnologies innovadores, fomentar-ne l'adopció, la participació de les empreses del sector, la transferència tecnològica i l'exploració de nous àmbits d'aplicació.

Notes:

1. Mikolov, Tomas; *et al.* (2013). «Efficient Estimation of Word Representations in Vector Space». arXiv:1301.3781
2. Pennington *et al.* (2014). «GloVe: Global Vectors for Word Representation».
3. Bojanowski *et al.* (2017). «Enriching Word Vectors with Subword Information»
4. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11
5. «Language Models are Few-Shot Learners» *https://arxiv.org/abs/2005.14165*
6. Un FLOP es una unitat de mesura de la potència de càlcul d'un ordinador que correspon a una operació de coma flotant per segon
7. *https://ai.googleblog.com/2023/03/palm-e-embodied-multimodal-language.html?m=1*
8. *https://www.aljazeera.com/economy/2023/3/16/chinas-baidu-unveils-chatgpt-rival-ernie*
9. *https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2017)598621*
10. *https://www.greens-efa.eu/en/article/press/victory-for-language-equality-in-the-european-parliament/*
11. *https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Gobierno_Abierto_Inicio/pae_Reutilizacion_de_la_informacion_en_el_sector_publico.html*
12. *https://www.go-fair.org/fair-principles/*