LLENGUA I LITERATURA | Maite Melero | Actualitzat el 07/07/2021 a les 10:15

La tecnologia obre portes a la diversitat lingüística digital

La Intel·ligència Artificial s'alimenta de grans quantitats de dades | Del 2018 ençà, els models preentrenats i l'aprenentatge no supervisat han obert les portes a una supervivència digital que abans només es podia garantir per a les llengües globals

El magnat Elon Musk va fundar el 2015 el grup d'investigació OpenAI, creadora del revolucionari model GPT-3. | Steve Jurvetson/Flickr.

Novè article del dossier «El català al món digital»

Quins són els punts forts i febles del català en l'àmbit digital?
L'any 2016 es va publicar una anàlisi de debilitats i fortaleses del català en l'àmbit digital (Bel i Marimon, 2016). En aquest document es ressaltava com a fortalesa l'interès precoç que les tecnologies de la llengua van despertar a Catalunya, sobretot comparat amb la resta de l'Estat, amb la possible excepció del País Basc. Efectivament, ja des de mitjans dels anys noranta del segle passat es van començar a utilitzar **sistemes de traducció automàtica de forma intensiva**, primer en publicacions bilingües de premsa (*El Periódico*, *Segre* i, més tard, *La Vanguardia*), i després a l'administració catalana. Aquests sistemes de traducció automàtica eren tecnològicament molt més rudimentaris que els que hi ha ara, però la proximitat de les dues llengües entre les que es volia traduir (castellà i català) feia que la qualitat fos prou acceptable per rendibilitzar-ne l'ús.

Aquesta experiència primerenca amb l'ús real de la tecnologia lingüística va potenciar l'aparició d'una sèrie d'empreses petites i mitjanes (reunides en l'associació Clusterlingua en aquell moment), que varen constituir un primer nucli del que avui constitueix un sector en plena expansió. Al mateix temps, aviat van sorgir una **desena de grups de recerca** dedicats a aquestes tecnologies, repartits per totes les universitats catalanes. D'entre els productes desenvolupats durant aquells anys destaca el processador FreeLing (Padró i Stanilovsky, 2012), programa de codi obert d'anàlisi lingüística del català i altres llengües, amb més de 250.000 descàrregues des del 2009. De la mateixa manera que Freeling, la majoria de recursos digitals per al català han estat produïts amb finançament públic pels grups d'investigació de les diferents universitats catalanes, a més de l'Institut d'Estudis Catalans i del TERMCAT, el centre de terminologia per a la llengua catalana.

Cal mencionar també dues iniciatives col·laboratives molt importants nascudes al principi de la digitalització: es tracta de Softcatalà i de Viquipèdia, que demostren la gran capacitat d'autoempoderar-se que té una comunitat lingüística com la catalana. El català, que ocupa la posició 91 en la llista de llengües del món pel seu nombre de parlants, ocupa en canvi la posició 15 en el rànquing multilingüe de Wikipedia, segons el nombre d'articles. Per contraposició, la Wikipedia en castellà ocupa el número 9 d'aquest mateix rànquing, essent la segona llengua en nombre de parlants.

A més de la importància de la Viquipèdia i de la influència de Softcatalà en la catalanització de platafomes digitals i portals web, i la creació d'eines d'idioma (correctors, traductors, etc.), **la presència del català a Internet és certament superior a la que li pertocaria per nombre de parlants i situació sociopolítica**. Tanmateix, la creixent globalització en el consum de continguts (plataformes digitals, xarxes socials com Youtube, Twitter o Instagram) debilita la posició de les llengües d'àmbit més petit, com el català, i reforça les grans.

En aquest context global, els usuaris de les xarxes socials **opten per la llengua que els dóna una major audiència**, això es fa més evident en les generacions més joves, incrementant així la diglòssia digital d'arrel generacional.

Aquest seria, doncs, un seriós punt feble per al català. És a dir, la competència en l'àmbit digital de les llengües globals, com ara el castellà o l'anglès, tant pel que fa a l'accés a continguts digitals -per exemple, en quina llengua fem les cerques a Internet-, com a l'ús mateix de la tecnologia -en quina llengua ens obliguen a parlar els *xatbots* o els assistents virtuals.

Val la pena assenyalar que les tecnologies de la llengua han experimentat un salt tecnològic molt important en els últims anys lligat a la revolució que les xarxes neuronals d'aprenentatge automàtic han portat a tot l'àmbit de la Intel·ligència Artificial. Els sistemes intel·ligents, ja sigui un traductor o un *xatbot*, s'entrenen (aprenen) a partir de grans quantitats de dades lingüístiques, normalment anotades de forma manual per a la tasca. Per exemple, **un traductor automàtic aprèn a partir de traduccions prèvies**. Això vol dir que quantes més dades -i de més qualitat- es disposin, millor serà la qualitat de les aplicacions que podem construir.

La manca de dades lingüístiques disponibles en una llengua amb un mercat petit, com el català, encareix el desenvolupament dels productes que l'integren. Aquest encariment limita l'interès que podrien tenir les grans empreses a l'hora d'incorporar el català i alhora el fa inaccessible a les empreses emergents d'àmbit local.

La dificultat d'accés a les dades és una debilitat generalitzada del sector, que afecta totes les llengües, degut a carències regulatòries i a la falta de cultura de reutilització i posada en valor d'aquest tipus de dades, però que en el cas de llengües amb mercats reduïts constitueix un problema de primer ordre.

En resum, podem apuntar com a punts forts del català en l'àmbit digital:

L'existència d'un sector tecnològic de petites i mitjanes empreses amb una llarga experiència
Un ecosistema universitari i de recerca fort (com s'assenyala a Costa-Jussà i Melero(2020))
Un sector públic i privat amb experiència llarga en l'ús d'aquestes tecnologies, singularment la traducció automàtica
Una comunitat lingüística de perfil tecnològic alt i empoderada
Una presència a Internet alta relativa al nombre de parlants, com il·lustra la fortalesa de la Viquipèdia

I com a punts febles:

La globalització de les xarxes socials i la conseqüent competència de les llengües globals
L'absència del català en aplicacions innovadores d'intel·ligència artificial (assistents virtuals, *xatbots*...)
La dificultat de posar a l'abast dels proveïdors d'aplicacions i serveis de tecnologia lingüística, dades i recursos lingüístics en català, necessaris per al desenvolupament d'aquestes aplicacions i serveis intel·ligents.

[noticiadiariambautor]93/229[/noticiadiariambautor]
Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis digitals?
Ja sabem que la IA s'alimenta de grans quantitats de dades. Mentre que fins ara aquestes dades havien d'estar anotades o etiquetades manualment per cadascuna de les tasques que es pretenia resoldre (traducció automàtica, sistemes de diàleg, etc.) amb un cost pràcticament només a l'abast de llengües grans com l'anglès, actualment, hi ha una línia d'investigació molt prometedora que permet millorar les aplicacions intel·ligents utilitzant dades no etiquetades, és a dir, text lliure. Són les anomenades **tècniques d'aprenentatge no supervisat**.

Dins d'aquesta línia, **l'any 2018 es produeix un fet que revoluciona les tecnologies de la llengua** i les converteix actualment en el sector de la IA amb l'avenç més espectacular. Es tracta de l'aparició dels models massius de llengua. Per construir un model de llengua, per exemple per al català, necessitem quantitats massives de text, però no cal que aquest text sigui anotat, és suficient que sigui text en català. Aquest model general es pot llavors ajustar a la tasca que es vulgui -per exemple, crear un *xatbot*- utilitzant conjunts anotats molt més petits dels que es necessitaven abans.

Un model de llengua és per tant un recurs de primer ordre, que serveix de base per resoldre multitud de tasques lingüístiques. De fet, com més massiu el model pre-entrenat, més petit cal que sigui el conjunt de dades anotat per la tasca. Per exemple, amb només 10 frases s'ha ensenyat al model GPT-3 de l'anglès a escriure un assaig sobre perquè els humans no han de tenir por de la intel·ligència artificial (podeu trobar aquests assaigs i moltes altres mostres de text artificial a la xarxa).

Els models pre-entrenats i l'aprenentatge no supervisat han obert les portes a una supervivència digital que abans només es podia garantir per a les llengües globals. La tecnologia avança rapidíssimament i cada pocs mesos apareixen noves arquitectures, que fan obsoletes les anteriors. El que persisteix però és la importància de les dades, tant més les dades de qualitat, netes i ben classificades.

Què pot fer el català per evitar caure en la irrellevància en plena revolució tecnològica? **Un accés convenient i ben regulat a les dades és imprescindible** per al desenvolupament de nous productes, aplicacions i serveis. Les polítiques de dades obertes són essencials per a la innovació en intel·ligència artificial i tecnologies de la llengua. Els models de negoci de les empreses generadores de dades i unes polítiques reguladores insuficients provoquen el control de les dades a mans d'un conjunt reduït d'agents i limiten greument la investigació i el desenvolupament tecnològic, en particular en llengües minoritàries. Cal promoure polítiques adequades de dades obertes basades en l'ètica, la transparència i l'accessibilitat a les dades, tant del sector privat com del sector públic, tot garantint els drets de la ciutadania a la privacitat i la confidencialitat. I cal també una cooperació estreta entre la indústria i els diferents organismes que generen, posseeixen, necessiten i utilitzen les dades.

Quins actors o recursos caldria activar perquè fos possible?
En relació al tema crític de les dades, la **disponibilitat lliure de dades de qualitat en català** creades amb suport públic, ja siguin publicacions de la pròpia administració, o d'organismes que reben finançament públic, com ara la Corporació Catalana de Mitjans Audiovisuals (CCMA), particularment en el que concerneix la tecnologia de la veu, **resulta fonamental**.

La sensibilització i la implicació dels poders públics és el més important en la tasca de dotar tecnològicament una llengua no global com el català i garantir-ne la supervivència digital. Tanmateix, cal també la **conscienciació i l'empoderament tecnològic de la pròpia comunitat de parlants**. Els mateixos factors que posen en risc una llengua, li ofereixen les eines per fer-la sobreviure. Això inclou, actors privats amb accés a dades (editorials, institucions diverses, empreses de comunicacions, etc..) i també plataformes col·laboratives com la ja mencionada Viquipèdia o Common Voice de Mozilla, de recollida de dades de veu. Totes aquestes iniciatives són susceptibles d'incrementar la quantitat de recursos disponibles en català i elevar el seu perfil

tecnològic.

Obtenir dades 'en cru' o 'primàries' és el primer pas en la cadena de valor. El següent és el processament computacional d'aquestes dades, incloent-hi tasques essencials de neteja i adequació. Un altre pas consisteix a anotar alguns subconjunts d'aquestes dades, ja sigui per adaptar el model preentrenat bàsic per resoldre tasques específiques -per exemple, un sistema conversacional- com per crear conjunts de referència que serviran per a poder avaluar automàticament el rendiment de les aplicacions.

L'accés lliure aquestes dades ja processades o 'secundàries', és a dir als conjunts de dades netes i als subconjunts de dades anotades, també és molt important. Aquestes dades secundàries se solen anomenar 'corpus'.

D'altra banda, la construcció dels models massius de llengua requereix de **grans capacitats computacionals** que no estan a l'abast de tots els actors. Centres com el Barcelona Supercomputing Center hi juguen aquí un paper central.

El processament computacional d'aquestes quantitats massives de dades deixa una **petjada de carboni notable** i s'ha d'evitar per tots els mitjans malbaratar aquests recursos. Per tant, de la mateixa manera que s'ha d'obrir l'accés a les dades per a tothom, s'ha d'afavorir l'accés lliure als models ja pre-entrenats construïts sobre aquestes dades (ja siguin generals de llengua o de àmbits específics, biomèdic, legal, etc) per tal d'evitar repetir complexos processos computacionals inútilment. Aquests models els podríem considerar dades 'terciàries'.

Tant els corpus preparats com els models preentrenats han de ser accessibles com a codi obert a llibreries de referència -Huggingface, Spacy, etc.-, de manera que qualsevol desenvolupador de tecnologia (grans corporacions, petites empreses, investigadors acadèmics) en pugui disposar lliurament i construir amb ells solucions tecnològiques per al català.

Aquests són alguns dels objectius que persegueix el projecte AINA, que vol garantir la supervivència digital del català (Melero i Costa-Jussa, 2021).

[noticiadiariambautor]93/224[/noticiadiariambautor]
Com encaixa la promoció del català en el context d'un mercat global dominat per llengües majoritàries?

Semblaria que la lògica de mercat afavoreix només les llengües amb mercats molt grans però això no deixa de ser una visió simplificadora. Un cop establerts aquests grans mercats, **a la indústria li interessa diversificar-se, omplir nous nínxols**. En aquest sentit el mercat lingüístic del català és un nínxol interessant i que, tot i no ser llengua majoritària, té unes dimensions no negligibles en l'entorn europeu.

Com hem anat dient, **la propera generació de sistemes intel·ligents necessitarà menys dades anotades per aprendre** i això és una bona notícia per al català, per dues raons: perquè la quantitat de dades es pot ajustar a les que genera una llengua de les dimensions del català i perquè s'abarateixen els costos que comporta l'anotació manual.

En resum, els avenços tecnològics actuals, una política adequada d'accés a les dades primàries i una gestió adequada de les recursos computacionals generats (dades anotades, models) permeten albirar un futur optimista per a les eines digitals en català en un mercat global.

[noticiadiariambautor]93/227[/noticiadiariambautor]

Com encaixa internament la promoció digital del català en un context de diglòssia a favor del castellà?
La diglòssia digital es produeix entre els parlants bilingües d'una llengua minoritària i d'una llengua global; aquests parlants, abans que perdre el tren digital, opten per la llengua gran que els dóna

accés a més continguts o els permet accedir a tecnologies més desenvolupades. **La diglòssia digital es un risc real per a tota llengua que hagi de compartir espai amb una llengua hegemònica**. És el cas de moltes llengües al món, sovint com a conseqüència de processos de descolonització. També es el cas d'algunes llengües europees, com ara l'irlandès gaèlic o el gal·lès que han de competir amb l'anglès, el sard que ha de competir amb l'italià i evidentment el cas del català, que ha de competir amb el castellà.

Tots els catalanoparlants són també competents en castellà i per tant pot tenir menys motivació a l'hora d'exigir la presència del català en el món digital. Es podria fins i tot dir que requereix un cert activisme o una certa actitud militant en favor de la llengua per assumir aquesta promoció de forma activa. Ens podem trobar que si l'esforç a favor de la promoció digital del català no es fa a temps o amb prou intensitat, **els usos adquirits facin difícil l'adopció del català com a llengua digital** per part d'aquells que ja s'haguessin habituat a utilitzar el castellà.

Tanmateix, cal ser optimistes. **Per primera vegada la situació tecnològica és propicia a la diversitat lingüística en el món digital**. Cal aprofitar l'avinentesa, creant l'ecosistema adequat per treure'n partit. I confiar que la disponibilitat d'eines tecnològiques de qualitat en català impulsarà el seu ús de forma natural.

[noticiadiariambautor]93/226[/noticiadiariambautor]

Bibliografia rellevant:

Bel, N.; Marimon, M. (2016). "Les indústries de la llengua i la tecnologia per al català". *Llengua i Ús: Revista Tècnica de Política Lingüística*, (58), 17-26.
Costa-jussà, M.R.; Melero-Nogués, M. (2020). "Converses al voltant de la intel·ligència artificial en clau catalana", *Revista de Llengua i Dret, Journal of Language and Law*, 74, 90-99.
Melero, M.; Costa-jussà, M.R. (2020). "AINA, un projecte d'intel·ligència artificial en clau catalana", *Revista de Llengua i Dret, Journal of Language and Law*
Melero, M. (2018). "El futur de les llengües en l'era digital: oportunitats i bretxa lingüística". *Revista de Llengua i Dret, Journal of Language and Law*
Padró, Ll.; Stanilovsky, E. (2012). *FreeLing 3.0: Towards Wider Multilinguality. Proceedings of LREC-2012*.
Rivera, R.; Tarín, C.; Villar, J.P.; Badia, T.; Melero, M. (2017). *Language equality in the digital age - Towards a Human Language Project*. Parlament Europeu. Consutat el 28 de juliol de 2018

Articles del dossier:
[noticiadiariambautor]93/234[/noticiadiariambautor]
[noticiadiariambautor]93/233[/noticiadiariambautor]
[noticiadiariambautor]93/229[/noticiadiariambautor]
[noticiadiariambautor]93/232[/noticiadiariambautor]
[noticiadiariambautor]93/226[/noticiadiariambautor]
[noticiadiariambautor]93/231[/noticiadiariambautor]
[noticiadiariambautor]93/230[/noticiadiariambautor]
[noticiadiariambautor]93/227[/noticiadiariambautor]
[noticiadiariambautor]93/223[/noticiadiariambautor]
[noticiadiariambautor]93/224[/noticiadiariambautor]
[noticiadiariambautor]93/222[/noticiadiariambautor]
[noticiadiariambautor]93/225[/noticiadiariambautor]