Pensem.

LLENGUA I LITERATURA | Marta R. Costa-jussà | Actualitzat el 16/06/2021 a les 15:30

El repte de convertir la tecnologia lingüística en una línia estratègica de país

Catalunya té una gran capacitat formativa en l'àmbit de la lingüística computacional, però reté poc talent | Dissenyar com aprenen les màquines impacta dramàticament en sectors varis més enllà del lingüístic, i el país hi pot tenir molt a dir

Catalunya té una gran capacitat en l'àmbit lingüístic i computacional. | pressfoto-freepik.com

Quart article del dossier «El català al món digital»

Quins són els punts forts i febles del català en l'àmbit digital?
D'una banda, el català és una llengua consolidada amb una riquesa de recursos considerable (Bel & Marimón, 2016). Per exemple, el català té ràdio, televisió i diaris, això fa que la generació de dades sigui constant. No estem parlant d'una llengua majoritària ni d'extremats pocs recursos. Per posar-ho en context, agafem la Wikipedia: el català és la 20a. llengua amb més pàgines de contingut, d'un total de 320 llengües representades. Així mateix, el català té cert suport institucional que vetlla perquè aquest recursos es continuïn generant. Per exemple, el Departament de Cultura col·labora en el doblatge de pel·lícules en català a Netflix. Per últim, en aquest llistat de punts forts, també destacaríem que **Catalunya té una gran capacitat formativa de professionals en l'àmbit lingüístic i computacional**.

D'altra banda, com a punts febles, el fet que el català no estigui formalment present a Europa fa que s'estigui frenant una font de dades multilingüe molt gran i que és de gran rellevància per millorar tecnologies com la traducció automàtica. La capacitat formativa que tenim contrasta amb la poca capacitat de retenció de talent, doncs **molts experts de l'àmbit de la lingüística computacional acaben marxant fora** per manca d'oportunitats locals. Finalment, el català té un suport molt polititzat i no té una estratègia clara que puguin aprofitar els centres o grups de recerca experts en aquest tema que hi ha a totes les universitats.

Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis digitals?
Abans de parlar de les accions a promoure, seria adient qüestionar-nos **si el català i la tecnologia lingüística poden esdevenir una línia estratègica per a Catalunya**, en la qual vulguem destacar, crear teixit empresarial, ocupabilitat i expertesa. Si és així, hauríem de plantejar-nos si tenim líders en aquest sector que ens puguin guiar per tal de fer-ho. Així mateix, hauríem de pensar si tenim gent formada al territori per seguir aquesta línia.

Tenim candidats/es per liderar aquesta iniciativa, ja sigui residint al territori o que estan fent carrera a l'estranger. Només cal veure la gran quantitat de noms catalans que surten en articles de lingüística computacional o ocupant càrrecs de responsabilitat en l'àrea (Costa-jussà & Melero, 2020). Respecte a la gent formada, i coherentment amb els punts forts que anomenàvem, tenim unes **universitats punteres en tecnologia lingüística** tant des del punt de vista lingüístic com computacional. Per tant, tenim elements que apunten a què tindria molt sentit optar per aquesta línia estratègica.

Ara bé, per fer-ho, no és qüestió d'engegar un projecte amb durada limitada. **Històricament s'han donat moltes subvencions** per desenvolupar sistemes de traducció automàtica, correcció ortogràfica i s'han realitzat accions com la creació i el finançament del Centre de Referència d'Enginyeria Lingüística (CREL-Centro de Referencia de Ingeniería Lingüística). Aquestes iniciatives han fomentat el desenvolupament de recursos i eines pel processament del català en l'àmbit digital. Ara bé, tot això crec que no ha estat suficient per retenir talent, per exemple. S'hi hauria d'apostar més.

Una estratègia per retenir talent seria **que la Generalitat creés un centre/institut de tecnologia lingüística**. Això ja ha passat en d'altres línies, en la que s'han creat noves entitats com l'Institut de Ciències Fotòniques (ICFO) o el Centre Tecnològic de Telecomunicacions de Catalunya (CTTC). En l'àrea de la llengua, el més semblant que existeix a Catalunya és el centre TALP, que fusiona dos grups de la mateixa universitat (UPC), un de processament de text i un de processament de la parla. De fet, això és el que tenen al País Basc, el HiTZ: Centro Vasco de Tecnología de la Lengua. La diferencia principal és que el govern basc finança aquesta entitat. També és cert que a Catalunya tenim la sort de comptar amb molts més centres i universitats i ens quedem curts si només fusionem dos grups de recerca.

En qualsevol cas, aquest centre de tecnologia lingüística a més a més de mantenir el talent que ara mateix està emigrant a l'estranger, podria atraure diners del sector privat i ser una font de recerca puntera i de creació de *start-ups* relacionades amb la tecnologia lingüística. La missió d'aquest centre podria ser vetllar pel català en tot moment **desenvolupant tecnologia que es pogués utilitzar pels milers de llengües al món** que són minoritàries. Catalunya seria un desenvolupador d'aquesta línia puntera creant productes que podrien ser interessants per altres governs, països en situacions similars i/o molt més precàries.

[noticiadiariambautor]93/222[/noticiadiariambautor]
Quins actors o recursos caldria activar perquè fos possible?
Com deia, **caldria involucrar finançament públic a llarg termini** amb el compromís econòmic que es requereix, com ha fet Alacant creant la Fundación Ellis en Intel·ligència Artificial o el centre HiTZ que anomenàvem amb anterioritat. Caldria definir l'estructura d'aquest centre que tindria un fort lideratge potent, carismàtic, un consell d'experts relacionats amb Catalunya, residint en territori català o a l'estranger, col·laboracions i representacions de tots els centres locals amb expertesa en llengua.

Com encaixa la promoció del català en el context d'un mercat global dominat per llengües majoritàries?
La tecnologia va en aquesta direcció. **El gran repte de tecnologies punteres d'Intel·ligència**

Artificial com és l'aprenentatge profund és funcionar amb poques dades. Aquí podem ser punters en tecnologia mentre beneficiem la nostra llengua i milers d'altres llengües minoritàries que tenen un número similar o extremadament menor de parlants.

Actualment, els sistemes amb millors resultats aprenen a base d'exemples etiquetats. Es tracta de sistemes supervisats. Què vol dir això? Doncs que per tenir un sistema de traducció automàtica es necessiten exemples de dades que tinguin traduccions. Si vull un traductor entre anglès i català, es necessiten milions de frases traduïdes entre anglès i català. Ara bé, **hi ha un aprenentatge que es diu no-supervisat**, que actualment no funciona tan bé com el supervisat. En el cas de la traducció, puc construir un traductor entre anglès i català, amb dades de l'anglès i del català que no tenen perquè ser traduccions uns dels altres. Per exemple, puc utilitzar dades del diari *Ara* en català i dades del *Times* en anglès.

Les traduccions de llengües minoritàries són de pitjor qualitat que les de llengües majoritàries; això també passa amb el reconeixement facial, que funciona millor o pitjor segons la teva raça o el teu gènere

I encara hi ha línies més innovadores, com els aprenentatges *zero-shot, one-shot* o *few-shot*, que són aprenentatges que no requereixen exemples (només instruccions, com seria 'tradueix'), un exemple (com seria *casa=house*) o alguns exemples (com seria *casa=house*, *formatge=cheese*). El cas de pocs recursos és un repte inherent en les aplicacions d'intel·ligència artificial i no només és aplicable a la llengua, sinó també a aplicacions mèdiques, reconeixement facial o conducció autònoma. A més a més, la recerca en aquesta línia no només està motivada per millorar la qualitat de les aplicacions, sinó que -més important que això- hi ha la motivació d'entrenar els sistemes automàtics amb poques dades per tal de reduir el desorbitat consum energètic d'aquests models entrenats amb moltes dades. Finalment, mantenir una quantitat limitada de dades també ajudaria a auditar les dades per reduir els biaixos actuals. Resulta que **els sistemes actuals no tenen la mateixa qualitat segons el col·lectiu que els utilitzi**. Sabem això per traducció on les llengües minoritàries tenen pitjor qualitat que les majoritàries. Però també passa amb el reconeixement facial, que funciona millor o pitjor segons la teva raça o el teu gènere. Això passa perquè **les dades amb les quals s'entrenen els sistemes no són representatives de la diversitat que tenim a la societat**. Més enllà del problema de la qualitat, també succeeix que els sistemes automàtics amplifiquen els nostres estereotips (Costa-jussà, 2019). Això té un alt impacte social, perquè per exemple hi ha sistemes automàtics de selecció de personal que descarten per gènere simplement perquè aquella posició històricament l'han ocupada més persones d'un gènere concret.

Tot i així, aquests aprenentatges (no supervisat o *zero*, *one*, *few-shot*) encara han de millorar molt. Però són, d'una banda, una necessitat per reduir el consum energètic i una oportunitat perquè molt llengües minoritàries com el català tinguin una representació equitativa a llengües majoritaries en el món digital.

La meva pregunta: **volem que Catalunya sigui representativa en el desenvolupament d'aquesta tecnologia que impactarà dramàticament en sectors varis més enllà del lingüístic?**

[noticiadiariambautor]93/225[/noticiadiariambautor]

Bibliografia rellevant:

Bel, N.; Marimón, M. (2016). Les indústries de la llengua i la tecnologia per al català. *Llengua i Ús: Revista Tècnica de Política Lingüística* [Barcelona: Generalitat de Catalunya. Departament de Cultura. Direcció General de Política Lingüística], núm. 58.

Costa-jussà M.R. (2019). *An Analysis of Gender Bias studies in Natural Language Processing*. Nature Machine Intelligence, Vol 1, Num 11, pages 495-496.

Costa-jussà & Melero (2020). Converses al voltant de la intel·ligència artificial en clau catalana. *Revista de Llengua i Dret*.

Mitchell, M. (2019). *Artificial Intelligence - A guide for thinking humans*. Farrar, Straus and Giroux.

Agraïments:
En aquest escrit exposo plantejaments estratègics que he tingut ocasió de discutir amb professionals de la comunitat científica catalana. En cap cas no em refereixo a què les meves paraules expressin l'opinió d'aquesta comunitat, ja que no tinc potestat per fer-ho. Tanmateix, podreu veure plantejaments que són fruit d'enriquidores converses amb una gran varietat de col·legues. Aprofito per agrair-los aquestes estones d'interacció, que ens fan créixer tant en l'àmbit professional com personal.

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]