LLENGUA I LITERATURA | Noé Casas | Actualitzat el 25/06/2021 a les 09:45

I si la IA permetés doblar pel·lícules automàticament al català?

La traducció automàtica d'alta qualitat de textos és una realitat gràcies a les tècniques d'Intel·ligència Artificial | Grans empreses com Amazon i Google treballen per poder obtenir doblatge instantani

El doblatge de pel·lícules i sèries podria patir un canvi substancial amb la irrupció de la Intel·ligència Artificial. | Freepik

Cinquè article del dossier «El català al món digital»

Quins són els punts forts i febles del català en l'àmbit digital?
Potser el principal punt feble del català en l'àmbit digital és el **reduït nombre de persones** que poden -o decideixen- utilitzar un servei digital o consumir contingut digital només si està disponible en català. Aquest punt feble ve derivat del nombre de catalanoparlants, que no és gaire elevat en comparació amb d'altres llengües majoritàries, afegit a la situació de diglòssia en la qual es troben el català i el castellà. Així, pels creadors de continguts digitals o proveïdors de serveis digitals, el potencial retorn de la inversió necessària per a oferir contingut o servei en català, pot no ser gaire alt. Hi ha alguns factors que poden, malgrat tot, portar els creadors de contingut o proveïdors de serveis a oferir-los en català, com la possibilitat d'obtenir un avantatge competitiu en el mercat català, la construcció d'una imatge de marca sensible als drets lingüístics dels catalanoparlants o inclinacions personals dels propis creadors, però les dades suggereixen que no és així en la majoria dels casos.

El principal punt fort -molt fort- del català a l'àmbit digital és el compromís d'un gran nombre de persones i entitats amb la promoció de la llengua. Fet que s'aprecia en la quantitat de contribucions a projectes col·laboratius en llengua catalana, com la Viquipèdia, amb un nombre de contribucions més elevat que les Wikipedies de llengües amb molts més parlants. Fet que, afegit al ric ecosistema de recerca i emprenedoria a Catalunya, ha conduït a què en l'últim any sorgeixin molts conjunts de dades públiques en català, amb les que crear sistemes d'Intel·ligència Artificial (IA) sobre els que construir serveis digitals. Què significa això? I per què és un avantatge? Molts serveis digitals es construeixen sobre sistemes d'IA que es basen principalment en tècniques d'aprenentatge profund. Aquests sistemes aprenen a desenvolupar la seva funció (és a dir, "s'entrenen") a base d'imitar exemples, normalment en grans quantitats. Un cas de sistema d'intel·ligència artificial que necessita dades per a ser entrenat són els assistents de veu com **Siri o Alexa que, per a cada llengua suportada, necessiten ser entrenats amb moltes hores de gravacions de veu** de persones diferents. Un altre exemple és el sistema de traducció de Google, que necessita milions de parelles de frase i la seva traducció en les llengües entre les que es vulgui traduir. Si no hi ha dades d'aquest tipus disponibles públicament, o si les dades no permeten el seu ús per a fins comercials, llavors l'empresa haurà d'invertir per a generar per si mateixa les dades amb les que entrenar els seus sistemes, augmentant així la inversió necessària per desenvolupar el seu suport de català.

En l'últim any, projectes com Mozilla Common Voice o el projecte AINA de la Generalitat de Catalunya han lliurat nombrosos conjunts de dades en català, **fent que es passi d'una manca quasi total de dades públiques a una gran disponibilitat**. Aquesta disponibilitat de dades en català ha derivat en l'aparició de sistemes d'IA lliures entrenats amb aquestes dades, com sistemes per obtenir transcripcions de gravacions de veu, o sistemes de traducció. Aquests sistemes poden ser utilitzats directament per petites i mitjanes empreses que, d'aquesta manera, no necessiten

entrenar els seus propis sistemes, reduint així la inversió necessària per desenvolupar serveis en català amb aquest tipus de tecnologies.

[noticiadiariambautor]93/222[/noticiadiariambautor]
Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis digitals?
D'acord amb l'*InformeCAT 2020*, mentre que totes les aplicacions de missatgeria instantània (Whatsapp, Telegram, etc) rellevants a Espanya suporten el català, només el 30,3% de les grans marques líders a Catalunya tenen la seva pàgina web en català. Dels serveis digitals basats en veu, Google Maps no suporta la veu en català, però Waze sí que ho fa. Entre els assistents de veu, Google Assistant suporta el català, Siri planeja suportar-lo i Alexa no el suporta.

Per tal que empreses privades considerin rendible la inversió de proporcionar els seus serveis o continguts en català, una estratègia directa és que la inversió necessària es redueixi. Però... com aconseguir-ho? Una forma és oferir-los part de la feina ja feta. Per això, és clau **un dels avantatges del català enfront d'altres llengües no majoritàries: la gran proximitat lingüística amb una llengua majoritària, el castellà**. Aquest fet permet que els sistemes de traducció automàtica generin traduccions d'alta qualitat entre el castellà i el català, de manera que la quantitat de revisió humana necessària sobre les traduccions automàtiques sigui petita. Així, hi ha diaris com *La Vanguardia*, *El Periódico* o *Segre* que porten ja anys publicant diàriament edicions en castellà i català gràcies a un sistema de traducció automàtica clàssica corregit per editors humans ('post-editors').

Una estratègia per reduir la inversió necessària perquè les empreses tinguin les seves pàgines web en català seria **posar a la seva disposició traductors automàtics castellà-català de qualitat**, que fossin fàcilment integrables en els seus fluxos de treball. Ja que les empreses que no tenen la seva web en català probablement no disposin de personal que verifiqui les traduccions, seria possible proveir-los també de la traducció inversa, és a dir, que quan tradueixin els textos de castellà a català, aquest nou text en català es tradueixi de nou al castellà, de forma que així puguin verificar si la traducció manté el significat original, tal com suggereixen recerques recents. Aquest hipotètic servei es podria reforçar amb la integració de correccions humanes per part del Servei d'Assessorament Lingüístic (SAL) del Consorci per a la Normalització Lingüística (CPNL).

Una idea similar a la qual proposo de donar accés a sistemes de traducció automàtica, és el servei eTranslation posat en marxa per la Comissió Europea a dins del programa Connecting Europe Facility, tot i que aquest és més ambiciós (tradueix entre 24 llengües) i crec que menys proper a les necessitats reals de les empreses i proveïdors de serveis.

«En un futur proper, les tècniques modernes d'IA permetran no només traduir diaris, sinó generar doblatges de pel·lícules i series automàticament»

Gràcies a la proximitat del català i el castellà, en un futur proper, les tècniques modernes d'IA permetran no només traduir diaris, sinó **generar doblatges de pel·lícules i series automàticament**. Una aproximació que ja està sent desenvolupada és la de fer traducció automàtica de la transcripció dels diàlegs d'un vídeo a la llengua destí i generar la veu corresponent, mantenint la sincronia amb els temps de la veu original. Amazon ha fet progressos considerables aplicant aquest tipus de tècniques de l'anglès a l'italià. D'aquesta manera, cal pensar que la seva aplicació a la generació de doblatge en català des del castellà pot obtenir encara millors resultats. Google va un pas més enllà i proposa complementar l'anterior mètode afegint-hi tècniques de clonació de la veu per obtenir la mateixa veu de l'actor original, i aplicant tècniques de *lip sync* per corregir els llavis

a la imatge perquè encaixin amb la veu traduïda.

Cal recordar que totes les tècniques d'intel·ligència artificial necessàries per fer-ho possible ja existeixen, tot i que la qualitat final encara s'ha de millorar, i que és qüestió de temps que alguna empresa -ja sigui Amazon, Google o alguna altra local- combini i refini aquestes tècniques per **agafar els subtítols en castellà d'una pel·lícula, traduir-los automàticament al català i generar la mateixa veu de l'actor original parlant en català**. Una vegada la qualitat del resultat final sigui acceptable, la inversió necessària per a proporcionar una versió doblada al català d'una pel·lícula o sèrie que ja tingui doblatge al castellà seria una fracció del cost actual, el qual sens dubte contribuiria a augmentar el nombre de continguts amb doblatge al català, permetent que millorés, per exemple, l'actual 0,5% del contingut de Netflix (PDF) amb subtítols o doblatge en català.

Una altra estratègia que ja s'està seguint és la de posar a disposició pública dades en català, com articles de diaris o subtítols de pel·lícules, que permetin a les empreses entrenar sistemes d'intel·ligència artificial. Tot i així, ara s'actua *a posteriori*, recollint dades que ja existeixen i que, de vegades, no poden utilitzar-se per a finalitats comercials. Una alternativa seria canviar a un model proactiu, pel qual els continguts audiovisuals o escrits finançats amb diners públics estiguessin subjectes a la **condició de poder ser utilitzats per entrenar sistemes d'intel·ligència artificial** i es recollissin activament com a part de la producció.

Actualment, quan es posen a disposició pública conjunts de dades, és convenient definir explícitament en quins escenaris i en quins termes es poden utilitzar aquestes dades. Això s'acostuma a fer associant una llicència d'ús a aquestes dades. Existeixen llicències estàndard, que recullen casos comuns de tipus d'ús. Una família de llicències d'aquest tipus són les Creative Commons, que especifiquen diferents casuístiques. Per exemple, permís il·limitat d'ús, permís únicament per a ús no comercial, etc. Tot i així, aquestes llicències no expressen de manera explícita si es permet l'ús de les dades per entrenar sistemes d'intel·ligència artificial. Respecte a això, potser caldria disposar de llicències més clares respecte a l'ús per a aquest fi, de manera que es proporcionés un marc legal segur per a empreses que utilitzessin aquestes dades. Per això, potser convindria **impulsar una nova llicència de Creative Commons** que fos explícita respecte a aquest tipus d'ús.

Quins actors o recursos caldria activar perquè fos possible?

Primer de tot, he de dir que el meu coneixement sobre les competències d'alguns dels òrgans i entitats que enumero a continuació es basa en la seva informació pública a la web, pel que potser no estic encertat amb quina entitat hauria de ser responsable de quina acció. Els prego comprensió al respecte.

Les entitats que persegueixen línies de recerca d'intel·ligència artificial aplicada al processament del llenguatge natural, com les universitats públiques, centres de recerca com el Barcelona Supercomputing Center (BSC) o associacions com Softcatalà o Col·lectivaT ja treballen per crear conjunts de dades i publicar models pre-entrenats, que després puguin ser utilitzats per altres empreses per construir productes i serveis sobre ells. Tot i així, **aquests esforços no són coordinats**, i no reben una promoció conjunta que potenciï la imatge de disponibilitat de recursos. Seria idoni que l'Institut d'Estudis Catalans (IEC), amb el suport del Center for Innovation in Data Tech and Artificial Intelligence (CIDAI), monitoritzés i coordinés la creació i manteniment d'aquestes dades i sistemes pre-entrenats, i els donés visibilitat com a conjunt.

Respecte a les entitats de recerca esmentades, caldria reforçar la recerca i el lliurament de tecnologia en algunes àrees que ja hagin sigut desenvolupades per altres llengües però no per al català, i en àrees en les quals la tecnologia d'intel·ligència artificial encara no hagi arribat al punt de maduresa per ser explotades per part d'empreses locals. El doblatge automàtic de castellà a català seria un exemple.

Sobre la proposta que els continguts audiovisuals finançats amb diners públics estiguessin subjectes a la condició de **poder ser utilitzats per entrenar sistemes d'intel·ligència artificials**, entenc que l'entitat clau és la Corporació Catalana de Mitjans Audiovisuals (CCMA). Tant a la producció de contingut propi (TV3, Catalunya Ràdio) com en les pel·lícules i documentals finançats per la CCMA s'hauria de negociar la inclusió de clàusules per permetre utilitzar el material resultant (diàlegs, àudios) per a aquest fi. Per la seva part, el Consell de l'Audiovisual de Catalunya (CAC) hauria de monitoritzar les produccions que la CCMA aconsegueixi incloure en aquest model, així com monitoritzar el compliment d'aquestes condicions.

Així mateix, respecte a la utilització de material escrit per entrenar sistemes d'IA, el Departament de la Presidència de la Generalitat de Catalunya, a través de l'Entitat Autònoma del Diari Oficial i de Publicacions (EADOP), hauria de compilar i **alliberar els textos publicats per l'administració** per a aquest ús. Aquesta tasca hauria de coordinar-se amb l'IEC, i amb els altres organismes i departaments de la Generalitat que gestionen la publicació de material.

Per donar una millor cobertura legal a l'ús de materials per entrenar sistemes d'intel·ligència artificial s'hauria d'**impulsar la creació de llicències** d'ús estàndard que fossin explícites respecte a aquest tipus d'ús. Una opció directa seria que el Departament d'Empresa i Coneixement de la Generalitat treballés amb l'associació Creative Commons, que gestiona aquest tipus de llicències, per impulsar i promoure la creació de noves llicències CC que permetin explícitament l'ús de dades amb aquesta finalitat. També caldria que el Congrés de Diputats impulsés mesures per ampliar el marc legal de l'ús de dades per mineria de textos de la transposició estatal de la regulació europea (PDF), per permetre altres escenaris a part de l'àmbit de la recerca científica.

Per últim, caldria recollir totes aquestes mesures en una revisió de l'Estratègia d'Intel·ligència Artificial de Catalunya.

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

El llenguatge és un vehicle de cultura i defineix una gran part de la nostra identitat. Tot i així, la seva funció principal i raó de ser és permetre la comunicació. Si un dels objectius d'un comunicador és que el seu missatge arribi al major nombre de persones, **és natural que triï la llengua que més persones puguin entendre**. Corroborant aquesta intuïció, dels 10 milions de llocs web més

populars globalment, un 61,4% estan disponibles en anglès, mentre que només un 3,7% ho estan en castellà, i al voltant del 0,04% en català. Si comparem aquestes xifres amb el nombre d'usuaris a Internet segons la seva llengua, veurem que les proporcions no es conserven, sinó que es concentren en l'anglès.

El català no es l'única llengua en aquesta situació. Per exemple, **a la Xina** la llengua oficial a nivell nacional és el mandarí estàndard, que va ser creat a mitjans del segle XX a partir del vocabulari dels dialectes xinesos del Nord, la pronunciació de Beijing, i la gramàtica vernàcula del moment. El mandarí es va imposar a tot el territori pel Partit Comunista Xinès, i és la llengua que es fa servir a l'escola, a tots el mitjans públics i a l'administració. Per contra, a la Xina s'hi parlen 300 altres llengües a banda del mandarí, una de les quals és **el cantonès**, el qual és el parlat majoritàriament a la província de Guangdong, amb 113 milions d'habitants. Tot i això, el nombre de programes de televisió en cantonès a Guangdong és només un grapat, motiu pel qual la població ha de recórrer a veure les notícies de la televisió de la ciutat semi-autònoma de Hong Kong, on el cantonès sí que és llengua oficial i gaudeix de molt més suport mediàtic. Com el cantonès i el català, **moltes altres llengües a tot el món es veuen en una situació que no afavoreix que la gent local pugui consumir contingut en la seva llengua habitual**.

Tot i així, si ens fixem en el marc local a Catalunya, podem veure una situació potser més favorable. Per exemple, un 30% de les marques amb més notorietat a Catalunya ofereixen la seva web en català, havent-hi sectors específics en què arriba al 67,4%. Per tant observem que, tot i que **hi ha una tendència global a què el món digital es concentri al voltant de les llengües majoritàries**, també veiem una tendència al nostre àmbit local a oferir els serveis digitals en català, tot i que encara hi ha marge per a la millora.

Per tant, entenc que la promoció del català ha de centrar-se en els productes i serveis oferts en i, sobretot, des de Catalunya, i que les estratègies per aconseguir-ho són reduir els impediments i incrementar els incentius per fer-ho. Respecte a reduir els impediments, una manera és **reduint la inversió que les empreses necessiten fer per oferir aquests productes i serveis en català**, amb les mesures que proposo a les anteriors respostes. Respecte a incrementar els incentius, crec que la millor manera és reforçar l'economia local, de manera que capturar el mercat català suposi cada vegada un objectiu més important per a les empreses que ofereixen productes i serveis digitals.

[noticiadiariambautor]93/225[/noticiadiariambautor]

Com encaixa internament la promoció digital del català en un context de diglòssia a favor del castellà? Entenc que la promoció del català a Catalunya es juga en dues grans àrees complementaries: augmentar el nombre de persones catalanoparlants i augmentar l'oferta de productes i serveis digitals en català.

Respecte a augmentar el nombre de catalanoparlants, hem de tenir en compte que el 35,7% de la població catalana ha nascut a fora de Catalunya i, d'ells aproximadament el 80% utilitzen poc o mai el català en el seu dia a dia, segons l'*Informe de Política Lingüística 2019* (PDF). D'altra banda, el 45,9% dels habitants de Catalunya nascuts a fora d'aquesta manifesten interès en aprendre o millorar els seus coneixements de català. Per tant, cal pensar que, tot i que hi ha una porció important dels residents a Catalunya que voldrien incorporar-se al grup de catalanoparlants, l'oferta educativa actual del català no satisfà les condicions per a que això ocorri, potser per falta de temps o motivació per assistir a classes presencials. Per això, tot i la quantitat de material *on-line* d'aprenentatge de català, crec que convindria que la Generalitat impulsés la creació d'una **aplicació mòbil gratuïta d'aprenentatge de català** que permetés que el grup de no-catalanoparlants augmentés sense fricció la seva competència del català.

Respecte a augmentar l'oferta de productes i serveis digitals en català, tal i com he descrit en les meves respostes anteriors, opino que la millor estratègia és **aprofitar les tecnologies basades en intel·ligència artificial per reduir enormement la inversió necessària** de les empreses per

suportar-lo en els seus productes.

Tot i així, crec que aquestes mesures no són efectives en el contingut audiovisual de petits creadors en plataformes com Youtube o Tiktok, on només hi ha un idioma principal (a diferència de las pàgines web o els doblatges, a on hi ha diferents idiomes en igualtat de condicions) i els incentius simplement es regeixen pel nombre de visualitzacions. Si es vol potenciar el contingut en català en aquest tipus de plataformes, crec que seria necessari impulsar un **sistema d'incentius paral·lel al de les pròpies plataformes**, oferint primes als vídeos en català amb major nombre de visualitzacions. Això estimularia la creació de contingut en català fins, potser, poder obtenir massa crítica suficient perquè els alineés amb els incentius de les pròpies plataformes.

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]