BERRIA
Ostirala, 2022ko apirilaren 1a

BERRIA

Euskarazko kalitate handiko corpusa sortu du EHUko Informatika fakultateko Ixa ikerketa taldeak: EusCrawl. Nahi duenaren esku jarri du, gainera. Euskarazko hainbat komunikabideren edukietan oinarritu du. Corpusak funtsezkoak dira hizkuntzan oinarritutako aplikazioak sortzeko.

Euskara ona makinentzat

Jakes Goikoetxea *Donostia*

Gaur egungo herritarrak hizkuntzari lotutako aplikazio ugari ditu inguruan: sakelakoan edo ordenagailuan idazten ari dela hurrengo hitza izan daitekeena erakusten dutenak, hitzez galdetu eta erantzuten duten laguntzaile adimendunak (Alexa, Siri, Google Assistant...), bezeroen arretarako txatbotak, itzultzaile automatikoak...

Gertuko, ohiko eta berehalako bihurtu diren aplikazio horien muinean, baina, itzaleko lana eta entrenamendua dago. Zehazki, adimen artifiziala, *machine learning*-a eta *deep learning*-a. *Machine learning*-a, ikasketa automatikoa, adimen artifizialaren adar bat da: algoritmoei datu pila bat ematen dizkiete, eta gai dira datu horietan patroiak, ereduak, identifikatzeko eta iragarpenak egiteko. *Deep learning*-a, ikasketa sakona, ikasketa automatiko mota bat da, sare neuronaletan oinarritua.

Funtsean, hizkuntzaren prozesamendua da. Ordenagailuek gizakien hizkuntza ulertzeko ahalegina, adimen artifiziala erabiliz. Teknologia horrek datu pila bat behar ditu, sistemak ikasteko eta entrenatzeko. Hizkuntzan oinarritutako teknologia denez, testuak eman behar zaizkio. Baita euskaraz ere. Ixa taldeak kalitate handiko euskarazko corpus bat osatu du, eta eskura jarri du, nahi duenak erabil dezan: EusCrawl (*ixa.ehu.eus/euscrawl*). EHU Euskal Herriko Unibertsitateko ikerketa talde bat da Ixa. Hizkuntzaren tratamendu automatikoan lan egiten du.

Euskararen kasuan, hizkuntza gutxitua denez, zaila du testu corpus erraldoiak biltzea. Euskarazko corpusak badaude. Batzuk ez daude erabili nahi dituenaren eskura. Beste batzuk bai: «Facebookek, Googlek eta horiek sortuak», azaldu du Aitor Soroak, Ixa taldeko eta Hitz ikerketa zentroko ikertzaile eta EHUko Informatika fakultateko irakasleak. «Baina corpus haiek erabat automatikoki sortuta zeuden: web osoa hartzen dute; webguneak zer hizkuntzatan dauden bereizteko programa automatikoak aplikatzen dituzte; eta edukiak hizkuntzaren arabera automatikoki sailkatzen eta biltzen dituzte. Hizkuntza guztien corpusak dituzte».

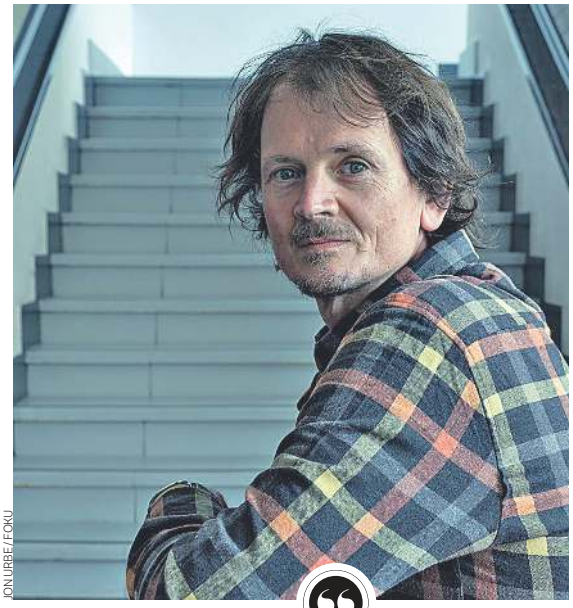
Googlerenak eta Meta AI-renak –lehen Facebook– dira euskarazko testu masa handienak: Googlerena, mC4, mila milioi hitzekoa; Meta AIrena, CC100, 416 milioi hitzekoa. Haien kalitatea, baina, zalantzan jarri da, euskarazko edukiak bereizteko programa automatikoek hainbat akats egiten dituztelako.

Komunikabideak

Kalitate oneko euskarazko corpusa sortu du Ixak. Ez du, Interneteko erraldoien gisan, begi estuko sarea hartu eta Internet osoan arrantza egin, horrela euskarazkoak bai, baina zaborra eta beste hizkuntza batzuetan dauden edukiak ere harrapatzen direlako. «Guk kontrakoa egin dugu», argitu du Soroak. «Aurrena iturri on batzuk identifikatu genituen; gero programa batzuk sortu genituen, haietatik informazioa xurgatzeko; eta horrela sortu da corpusa».

Edukiak Creative Commons lizentzia librearekin banatzen dituzten Interneteko zenbait webgune aukeratu zituzten, komunikabideak batez ere: Tokikom (tokiko 76 komunikabide biltzen dituen elkartea), Berria, eskualdeetako *Hitza* egunkariak, euskarazko Wikipedia, *Argia* eta Bilbo Hiria irratia. Haiek sortutako edukiak xurgatu egin zituzten –*crawl*, ingelesez–. Emaitza: EusCrawl. 12,5 milioi dokumentu eta 423 milioi hitz.

Ikasketa sakonaren barruan hizkuntza eredua deitutako teknika edo teknologia dago, sare



JON URBE / FOKU

neuronaletan oinarritua. Hizkuntza eredu horiek testuarekin entrenatzen dituzte: «Testua irakurtzen dute, hizkuntzaren patroiak ikusten dituzte, eta, milioika eta milioika hitz irakurriz, ikasi egiten dute. Horrela, hizkuntzari buruzko eredu matematiko erraldoi bat sortzen duzu. Hori bai, ondo idatzitako testuak eman behar zaizkio». Ez dute ikasi bakarrik egiten. Testu berriak sortzeko gai ere badira. Hizkuntza ereduek hizkuntzaren egitura egitura matematiko bihurtzen dute, nolabait esateko. «Hizkuntzari buruzko aplikazioak egiteko gaur egungo tresna onenak dira», nabarmendu du Soroak.

Ikasi egiten dute, «baina ez dakigu oso ondo zer ikasten duten», onartu du Soroak. Kutxa beltzekin parekatu ditu. «Ikerkuntza arazo bat dago: jakitea zer demontre dagoen hor barruan eta zer egin hori hobeto kontrolatzeko. Ataza bat ematen diozu, eta ikasi egiten du. Input bat ematen diozu, eta output bat ematen dizu».

Ez dira kontrolatzeko errazak, sistema erraldoiak baitira. Hainbat parametro dituzte. Bi adibide ezagun: GPT-3 hizkuntza ereduak, esaterako, 175.000 milioi parametro ditu; Bert-large-k, berriz, 350 milioi.

Hizkuntza ereduak

Ixa taldeak, EusCrawl osatu eta gero, euskararen bi hizkuntza eredu sortzeko eta entrenatzeko erabili zuen. Hizkuntza eredu horietako bat gaur egun euskararako dagoen eredu handiena da, 355 milioi parametrokoa. Hizkuntza eredu berri horiek euskarazko beste corpus batzuekin ere entrenatu zituzten: Googleren mC4rekin eta Meta AIren CC100ekin, berez, EusCrawl baino kalitate txarragokoak.

> **Euskarazko ahalik eta corpus handiena biltzen saiatu behar dugu, euskarazko aplikazio hobeak nahi baditugu»**
>
> **Aitor Soroa**
> EHUko Informatika fakultateko irakasle eta Ixako eta Hitz-eko ikertzailea

Probetan ikusi zuten baietz, EusCrawlen testuen kalitatea besteena baino hobea zela, baina ezusteko ondorio bat ere atera zuten: hizkuntza eredu guztiei hizkuntzaren prozesamendurako zenbait eginkizun jarri zizkieten, eta emaitzak berdintsuak izan ziren. Alegia, ez zegoen hainbesteko alderik EusCrawlekin entrenatutako eta Googleren eta Meta AIren corpusekin entrenatutako hizkuntza ereduen artean.

«Horrek erakusten digu munstro hauek entrenatzeko garrantzitsuagoa dela testuen kantitatea, kalitatea baino», ondorioztatu du Ixak. «Beraz, euskarazko ahalik eta corpus handiena biltzen saiatu behar dugu, euskarazko tresna eta aplikazio hobeak nahi baditugu».

Euskarazko komunikabide askoren testuak xurgatu dituzte. Handien artean, EITBren edukiak falta dira. Argitaletxeenak ere bai. Eta beste hainbat: sare sozialak... Sare sozialei dagokienez, hizkuntza ereduak hizkuntza horren zenbat eta erregistro gehiago izan, orduan eta eredu aberatsagoa da.

«Hala ere», ohartarazi du Soroak, «euskararen corpus ezagun guztiak bilduta ere, hizkuntza nagusien tamainatik oso urrun geratuko ginateke, eta horrek euskarazko hizkuntza ereduei goi borne bat ezartzen die». Arriskua: euskararentzat sor daitezkeen tresnen kalitatea ez izatea ingelesarentzat sortzen direnen parekoa, esaterako.

Egoera horri aurre egiteko, Ixak bi helburu estrategiko ezarri ditu: alde batetik, corpus handiagoak biltzea, euskarazko testu ekoizle guztien edo gehienen testuak erabili ahal izatea; bestetik, testu gutxiagorekin ikasteko gai izango diren hizkuntza ereduen ikerketa bultzatzea, hizkuntza gutxitua izatearen mugei aurre egin ahal izateko. Beste herrialde batzuetan ikertzaileei lizentzia librerik gabeko testuak erabiltzen uzten diete, corpusak osatzeko, teknologia horien garapena lehentasuna baita.

Estrategiaren premia

Ixak sortu du EusCrawl, Ixak erabili du hizkuntza ereduak sortzeko, eta Ixak planteatu ditu helburu estrategikoak. EHUko ikerketa talde bat da Ixa. Ez al dute hizkuntzaren prozesamendurako teknologiek herri estrategia bat behar? Ez al da erakundeen lana halako estrategiak bultzatzea eta garatzea? «Apustu bat behar da, baina ez Ixarena bakarrik, erakundeen apustua behar da», Soroaren iritzi.

Ikertzaileak Espainia aipatu du. Espainiako Gobernuak badu Hizkuntzaren Teknologiak Bultzatzeko Plana. Eusko Jaurlaritza ere antzeko zerbait sortzeaz hitz egiten ari dela aipatu du. «Euskararentzat behar-beharrezkoa da halako estrategia bat, bai ikerkuntzaren aldetik, bai aplikazioen aldetik. Ez badugu euskaraz egiten, jendeak beste hizkuntza batean egingo du». Euskal Herrian goi mailako ikerketa taldeak daude hizkuntzaren prozesamenduan.

EusCrawl, euskarazko kalitate handiko corpus librea, ez da Euskal Herrira begira soilik egindako ekarpena. Munduko edozein ikertzailek erabili ahal izango du. Erabiltzen ari dira jada, BigScience proiektuan: hizkuntza eredu eleaniztun eta librea sortu nahi dute –normalean enpresa handiek baino ez dituzte egiten, garestiak baitira–. Proiektu irekia eta kolektiboa da. Euskara ere txertatu dute, EusCrawlen bidez. Beraz, sortutako hizkuntza ereduak euskaraz ere jakingo du.

EusCrawlek badu beste erabilgarritasun bat: hizkuntzalaritza, hizkuntzaren azterketa. Corpusa balia dezakete erabilera ikusteko, esaterako.

ARGI ALDIAN
Ana Galarraga Aiestaran

Elhuyar Zientzia

Arren itxuraz, gizonez jantzita

Erdialdeko eta Hego Amerikan, bada *Florisuga mellivora* izeneko kolibri espezie bat. Heldutan, hegazti eta beste animalia askotan gertatzen den bezala, arrek eta emeek itxura desberdina dute. Hegazti gehienetan, gazteek eme helduen tankera dute, eta heltzean, arrek beste lumaje bat hartzen dute, normalean, koloretsuagoa eta deigarriagoa. *Florisuga mellivora* kolibrietan, baina, alderantzizkoa gertatzen da: gaztetan ar helduen antza dute lumajean, eta, heltzean, emeek hartzen dute bestelako lumajea. Ez guztiak, ordea: emeen % 20k arren itxurari eusten diote.

Ikertzaileek ez dakite horren atzean arrazoi genetikoak ala ingurunekoak ote dauden, baina ondorio bat, behintzat, ikusi dute: ar-itxura duten emeek jazarpen txikiagoa jasaten dute, besteekin alderatuta. Hain zuzen ere, arrak dira jazarleak, eta, bereziki, eme arrunten aurka erabiltzen dute indarkeria, jana eskuratzeko. Ar-itxura duten emeek, berriz, ez dute erasorik jasaten, eta ez dute eragozpenik arrek adina jateko.

Ugalketaren ikuspegitik, ar-itxura izatea ez da abantaila bat. Aitzitik, ohiko lumajea eta arrarena duten bi emeen artean, lehena aukeratzen dute arrek gortea egiteko. Baina ar-itxurakoak ere gorteatzen dituzte; beraz, itxura aldatzeak dakarren irabazia galera baino handiagoa da.

> ***Eta XXI. mendean? Bada, emakumeek sinatutako artikulu zientifikoek aukera gutxiago dituzte aldizkari ospetsuetan argitaratuak izateko, eta aipamen gutxiago jasoko dituzte***

Kolibri horiek burura ekartzen dute gizonen aukerak eskuratzeagatik gizon-itxura hartu duten emakumeen kasua. Historian, asko izan dira horrela jokatu dutenak, tartean, zientzialariak. Ezagunenetako bat Sophie Germain da. 1776an jaio zen, eta txikitatik zen zientziazalea. 18 urte zituela, Eskola Politeknikoa ireki zuten Parisen, baina gizonak baino ez zituzten onartzen. Alabaina, Germainek aurkitu zuen han ikasteko modua: postaz jasotzen zituen lezioak eta bidaltzen lanak, Monsieur Antoine-August Le Blanc izena hartuta.

Joseph Louis Lagrange irakaslea Le Blancen lan bikainaz jabetu zen, eta ezagutu nahi izan zuen. Germainek ezin izan zuen emakumea zela ezkutatu, baina Lagrangek ez zuen baztertu; alderantziz, Germaini matematikaren munduan aurrera egiten lagundu zion. Azkenean, matematikari handia izatera iritsi zen, eta Parisko Zientzien Akademian onartua izatea lortu zuen.

Garai bertsuan, Jeanne Baret botanikaria gizonez jantzita ontziratu zen munduari bira emango zion lehen frantziar espedizioan. Izan ere, o ezustekoa!, emakumeek debekatua zuten espedizioetan parte hartzea. Munduari itzulia eman zion lehen emakumea bilakatu zen, jada emakumez jantzita, eta botanikaren arloan sekulako ekarpena eginda.

Eta XXI. mendean? Bada, emakumeek sinatutako artikulu zientifikoek aukera gutxiago dituzte aldizkari ospetsuetan argitaratuak izateko, eta aipamen gutxiago jasoko dituzte. Eta jantziek ere elementu garrantzitsua izaten jarraitzen dute, onartuak ala baztertuak izateko, adibidez, kongresuetan. Hautespen naturala deituko diote gero.