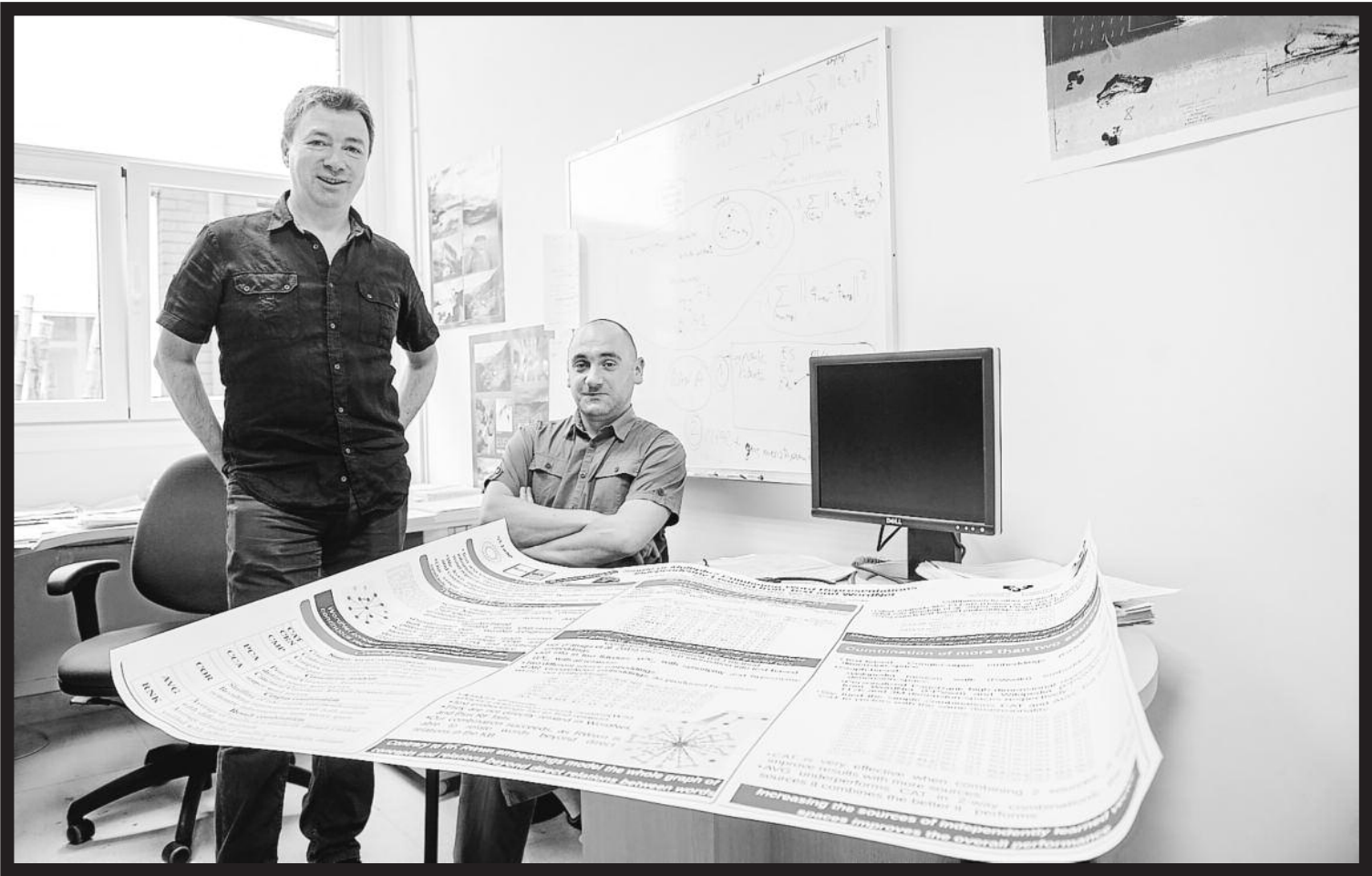




Eneko Agirre (zutik) eta Josu Goikoetxea, Agirreren bulegoan, ikerketaren posterra erakusten. Juan Carlos RUIZ | ARGAZKI PRESS

infraganti

ENEKO AGIRRE

JOSU GOIKOETXEA

EHUko Informatika fakultatean, Donostian, hartu gaitu Eneko Agirre Google Research-ek saritutako irakasleak. Aitor Soroa eta Oier Lopez de la Callerekin batera ikerketan lagun duen Josu Goikoetxea ere batu da elkarriketara. IXA taldekoak dira, eta zenbakiak eta letrak uztartuz altxorren mapa osatu dute makinek erraten dieguna uler dezaten.

Hitzak irakurtzean guk letrak ikusten ditugun lekuan zenbakiak eta koordinatuak irudikatzen dituzte Eneko Agirrek eta Josu Goikoetxeak. Hitz bakoitzari zenbaki sorta bat edo bektore bat esleitzen diote. «Zenbaki bakoitza esanahiaren tasun edo ezaugarri semantiko bat da. Orduan,

zenbakien konbinazio desberdinekin esanahi bat edo beste bat lortzen da», agertu du Goikoetxea doktorego ikasleak. Esanahia puntu bat da espazio batean. Zenbat eta hurbilago egon puntuak, semantikoki ere hurbilago daude. Bertze modu batera kontatu du Agirre Lengoaia eta Sistema Informatikoak saileko irakasleak: «Ordenagailuek hitzak ikusten dituztenean ez dakite ezer

gure munduari buruz, ez dakite hitzak zeri lotuta dauden eta zer esan nahi duten. Bai itzulpen automatikoan eta bai bilaketak egitean, Google bertan edo beste edozein bilatzailetan, gertatzen dena da hitz bera ez badu topatzen berdin idatzita ez dakiela zer egin. Orduan, makinek munduari buruzko ezagutza behar dute, hau da, hitzen esanahiei buruzko informazioa, gauzak elkarrekin lotzeko». Adi-

bidez, jakiteko “eserleku” eta “aulki” sinonimoak direla. Zenbakiak erranen dute “katu” hitza koordinatu batean dagoela eta bertze zenbaki batzuek bertze puntu batean kokatuko dute “kutxa”. “Banku” hitza “kutxa”-tik hurbil dago. Hala jakin daiteke gertu daudenek esanahi antzekoa dutela. “Kutxa” aipatzen den esaldi bat itzultzen ari bazara, makinak ez badu lehenago ikusi hitz hori, badaki “banku”-ren antzekoa dela eta honen itzulpenarekin itzuliko du. Kutxen finantza informazioa bilatzean ere, lotutako dokumentuak topatuko lituzke nahiz eta “kutxa” hitza ez aipatu. Hau da, munduari buruzko informazioa gehitzen diote hitz bakoitzari mapako lokalizazio bezala. Baina hainbertze hitz daude, eta berriak sortzen dira, gainera... Nola egin lan hori guztiekin? Goikoetxeak esplikatu du alde batetik corpusak erabiltzen dituztela, Wikipedian dagoen testu osoa, adibidez. Bestetik, grafoak baliatzen dituzte. Wikipediako artikuluko bateko esteka batek bertze artikuluko batera bidaltzen zaitu. Kontzeptuen arteko egitura erraldoi bat dago loturekin eta horri deitzen zaio grafoa. Testuak izaera bateko informazio semantikoa du eta grafoak bertze batekoa. «Osagarriak direla ikusi dugu eta bi informazio mota horiek bateratu nahi izan ditugu». IXA taldearen metodoa guztiz automatikoa da. Ezagutzen dituzten hizkuntzekin probatzen dute, euskara, gaztelania eta ingelesarekin. Baina edozein hizkuntzarentzat balio du. «Gertatzen dena da ez dakigula ondo ala gaizki egiten duen, baina berdin egiten du», diote irriz. **ALT XORRAREN MAPA** Wikipediako informazio guztia mapa batera ekartzen dute. Hala, hitz bakoitza bere lekuan dago. Hitzen zerrenda luze bat da, bakoitza bere koordinatuekin. Alt xorren mapa bezalakoa da. «Nahi dituzun hitzak irudikatzen ahal dituzu, edo distantziak kalkulatu», dio doktorego ikasleak gure harridura aurpe-

giarekin barrez. «Orduan, itzul- tzerazoaazenean itzultzaile auto- matiko batek mapa hau hartzen du eta laguntzen dio bere lana hobeto egiten», gaineratu du lankideak. Mapa horrek hainbat erabilera izan ditzake. Adibidez, hagitz ohikoa da argazki artxibo batean argazki bat ezin aurkitu ibiltzea irudia sartu duenak ez dituelako zehazki zuk idatzitako hitzak idatzi. Mapok hagitz erraz aplika daitezke halakoe- tan, hitzen atzean dagoen esa- nahia delako garrantzitsua eta ez hitza bera.

Google Researchek saria eman zien proposamena baliagarria izan daitekeelako hamaika kon- tutarako. «Mapa berean irudika- tzen ditugu hizkuntza guztiak. Ez bakarrik hitzak, esanahiak ere bai. Adiera bakoitza. Adibi- dez, banku baten argazkia bila- tzean zein banku nahi duzun bereizi ahal izango duzu, eser- tzeakoak ala banketxeak».

Hori lortu nahian dabilta. Oi- narrizko ikerkuntza da eurena. «Badakizu zure mapa hori oso ona bada, erabilerak berehalako- ak izango direla. Baina inoiz ez dakizu helduko zaren ala ez. Ur- tebete barru agian esango dugu emaitzak hobetu ditugula eta %75ean gaudela. Hori nahikoa da zure bilatzailea edo itzulpen zerbitzua hobetzeko? Beharbada bertan integratzen duzu eta ez du lana guztiz ondo egiten. Hori da ikerkuntzan inoiz ez dakizu- na, zenbat beharko duzun gero erabiltzaileek onura nabaritze- ko», argitu du Agirrek.

Google Researchen saria ho- netan ikertzeko da. Proposame- na gustatu zaiolako eman die aukera, baina ez du erran nahi proposamenak derrigorrez ongi funtzionatzea ekarri behar due- nik. Hemendik urtebetera Goo- gle etxeak ikerketa nola joan den ikusiko du. «Hura bisitatzen egon ginen. Gure arloko ikerla- riak dira, gu bezalakoak. Antze- ko gaiak ikertzen dabilta».

EUSKARAREN ALDE BALIATU

Doktorego ikasleak euskaran ja- rri du fokua. «Ez ditugunez ho- rrenbeste baliabide, sektoreak ez dira kalitate lar onekoak eta metodo hauekin ingeleseko edo



Agirre eta Goikoetxea, EHUko Informatika fakultatean. Juan Carlos RUIZ | ARP

gazelaniazko informazio se- mantikoa sartu nahi dugu eus- karazko sektoreetan, bilaketen eta itzulpenen kalitatea hobe- tzeke». Testuen tamaina txikia- goa da euskaraz, grafoena ere bai. Informazio gutxiago dago. Hizkuntza nagusi batzuk daude sarean: ingelesa, gaztelania, txi- nera... «Gutxi batzuk dira, ba- tzueta frantsesarentzat eta ale- manarentzat ere ez daudelako hainbeste testu jasota».

Helburuetako bat da, ingele- seko mapa hagitz ongi egiten dutenez, bertatik ikasitakoa eus- karazkoa hobetzeko baliatzea. «Euskara hizkuntza gutxitua izateko ez dago hain gaizki beste batzuekin alderatuta, inondik inora. Hala ere, hobetu nahi du- gu eta eleaniztasun hau ustiatu nahi dugu euskararen alde ere», adierazi du Goikoetxeak. Hiz- kuntzak batze hori interesgarria

da beren artean antzekotasunak daudela ikusten delako.

Burua batez ere erabilera kon- putazionalan jarrita dute, bai- na hizkuntzalariek ere erabil de- zakete, hitzen erabilerak aztertzeke balio duelako.

MILIOIKA HITZEN IRUDIA

Hiru milioi hitzeko edo gehia- goko mapa osatu dute ingelesez. Hitzak ez dira bakarrik hiztegie- tan topatzen ditugunak; markak ere badira, jendeak erabiltzen dituen espresioak, Whatsapp, izen propioak, bereziak, leku izenak... Edozer gauza. Horrek balio du makinak jakiteko hi- tzen atzean zer dagoen. Hitz bat entzutean, «kutxa» edo «katu», gauza aunitz etortzen zaizkigu, kutxei eta katuei buruz gauza pila bat dakizkigu. «Egiten du- guna da bai Wikipediatik eta bai testu dokumentuetatik katuari

Hizkuntzarekin lan egiten duten programa gero eta gehiago daude, tartean makinekin hitz egitekoak. Baina erraten dieguna ulertzeko gure munduaren irudi edo ispilu bat behar dute

IXA ikertaldea ekipo bat bezalakoa da eta ahal den onena izatea komeni zaie. Beti daude prest jendea hartzeko. Fisikariak, Irakasle ikasketak egindakoak eta filosofoak ere badira taldean

ku artean: itzultzaile automati- koa, zuzentzaile ortografikoa... 60 bat ikertzaile dira. Heren bat hizkuntzalariek dira eta bi he- ren informatikariak.

Sare neuronalak erabiltzen di- tuzte. Horiek egiten dutena da hitz baten esanahia ikasteko in- guruko hitzen esanahia kon- tuan hartu. «Hori da gure oina- rri linguistikoa; gainerakoa matematika da».

LIGAKO PUNTUAK BEZALA

Mapak nola osatzen dituzten eta beren eguneroko lana nola- koa den galdetuta, irriz eta aho batez erantzun digute: «Ordena- gailuaren aurrean. Alde batetik sare neuronalak dituzu, eta, bes- tetik, baliabideak: testu edo cor- pus handiak ala grafoak. Progra- man sartzen dituzu eta ordu, egun edo asteetan egon ahal da makina prozesatzen. Amaitzen duenean fitxategi erraldoi bat daukazu. Hitz bakoitzak sektore bat du esleituta. Probak egin, ebaluatu eta akatsak konpon- tzen ditugu. Jende gehiago dago gauza hauek ikertzen eta beren lana irakurri, aztertu, ideia be- rriak pentsatu eta martxan jar- tzen ditugu. Zenbaki batzuk sor- tzen ditu sistemak eta Ligako puntuak bezala dira. Zenbat eta puntu gehiago, talde hobea».

Beraiena ekipo bat bezalakoa da, eta, ahal den onena izatea komeni zaienez, beti daude prest jendea hartzeko. Proiektu honek eta IXA taldeak lantzen dituen gaiak master batean (Language Analysis and Proces- sing) ematen dituzte. Etortzen denak ez du zertan informatika- ria edo hizkuntzalaria izan. Ma- tematikariak, fisikariak, itzul- tzaileak, Irakasle ikasketak egindakoak eta filosofoak ere badira taldean. Hagitz progra- matzaile onak, gainera, gauzak ikasi egiten baitira. •

Maidier lantzi Goienetxe

buruz agertzen den informazio guztia prozesatu eta formula matematikoen bidez mapan ko- katzeko erabili, bestela makinak 'katu' irakurtzean lau hizki di- tuela soilik dakielako».

Hizkuntzarekin lan egiten du- ten programa gero eta gehiago daude: itzulpen automatikoa, galdera-erantzun sistema, bila- tzaileak eta bertze aunitz, adibi- dez kotxean irratia erratea «ja- rri Info7» edo Osakidetza deitzean makina batekin min- tzatzea txanda hartzeko. Baina ongi funtzionatzeko hitz bakoi- tzaren gibelean zer dagoen jakin behar dute.

IXA ikerkuntza taldeko kideak dira proiektuan dabilzanak. Talde honek 26 urte darama la- nean. Hizkuntzaren prozesamendua aztertzen du, gero eta pisu handiagoa hartzen ari den arloa. Egitasmo ugari dituzte es-