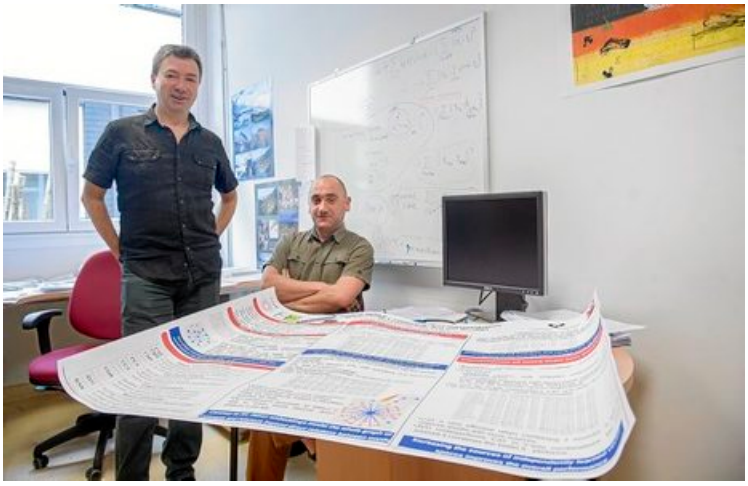
2016/05/14

ERRITARRAK | 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 elkarrizketara. IXA taldekoak dira, eta zenbakiak eta letrak uztartuz altxorraren mapa osatu dute makinek erraten dieguna uler dezaten.

MAIDER IANTZI GOIENETXE



H itzak irakurtzean guk letrak ikusten ditugun lekuan zenbakiak eta koordenatuak 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,

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

hau da, hitzen esanahiei buruzko informazioa, gauzak elkarrekin lotzeko». Adibidez, jakiteko “eserleku” eta “aulki” sinonimoak direla.

Zenbakiek erranen dute “katu” hitza koordenatu 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 artikulu bateko esteka batek bertze artikulu 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.

Altxorraren mapa

Wikipediako informazio guztia mapa batera ekartzen dute. Hala, hitz bakoitza bere lekuan dago. Hitzen zerrenda luze bat da, bakoitza bere koordenatuekin. Altxorraren mapa bezalakoa da. «Nahi dituzun hitzak irudikatzen ahal dituzu, edo distantziak kalkulatu», dio doktorego ikasleak gure harridura aurpegiarekin barrez. «Orduan, itzultzera zoazenean itzultzaile automatiko 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 halakoetan, hitzen atzean dagoen esanahia delako garrantzitsua eta ez hitza bera.

Google Researchek saria eman zien proposamena baliagarria izan daitekeelako hamaika kontutarako. «Mapa berean irudikatzen ditugu hizkuntza guztiak. Ez bakarrik hitzak, esanahiak ere bai. Adiera bakoitza. Adibidez, banku baten argazkia bilatzean zein banku nahi duzun bereizi ahal izango duzu, esertzekoak ala banketxeak».

Hori lortu nahian dabiltza. Oinarrizko ikerkuntza da eurena. «Badakizu zure mapa hori oso ona bada, erabilerak berehalakoak izango direla. Baina inoiz ez dakizu helduko zaren ala ez. Urtebete 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 dakizuna, zenbat beharko duzun gero erabiltzaileek onura nabaritzeko», argitu du Agirrek.

Google Researchen saria honetan ikertzeko da. Proposamena gustatu

zaiolako eman die aukera, baina ez du erran nahi proposamenak derrigorrez ongi funtzionatzea ekarri behar duenik. Hemendik urtebetera Google etxeak ikerketa nola joan den ikusiko du. «Hura bisitatzen egon ginen. Gure arloko ikerlariak dira, gu bezalakoak. Antzeko gaiak ikertzen dabiltza».

Euskararen alde baliatu

Doktorego ikasleak euskaran jarri du fokua. «Ez ditugunez horrenbeste baliabide, bektoreak ez dira kalitate lar onekoak eta metodo hauekin ingeleseko edo gaztelaniazko informazio semantikoa sartu nahi dugu euskarazko bektoreetan, bilaketen eta itzulpenen kalitatea hobetzeko». Testuen tamaina txikiagoa da euskaraz, grafoena ere bai. Informazio gutxiago dago. Hizkuntza nagusi batzuk daude sarean: ingelesa, gaztelania, txinera... «Gutxi batzuk dira, batzuetan frantsesarentzat eta alemanarentzat ere ez daudelako hainbeste testu jasota».

Helburuetako bat da, ingelesezko mapa hagitz ongi egiten dutenez, bertatik ikasitakoa euskarazkoa hobetzeko baliatzea. «Euskara hizkuntza gutxitua izateko ez dago hain gaizki beste batzuekin alderatuta, inondik inora. Hala ere, hobetu nahi dugu eta eleaniztasun hau ustiatu nahi dugu euskararen alde ere», adierazi du Goikoetxeak. Hizkuntzak batze hori interesgarria da beren artean antzekotasunak daudela ikusten delako.

Burua batez ere erabilera konputazionaletan jarrita dute, baina hizkuntzalariek ere erabil dezakete, hitzen erabilerak aztertzeko balio duelako.

Milioika hitzen irudia

Hiru milioi hitzeko edo gehiagoko mapa osatu dute ingelesez. Hitzak ez dira bakarrik hiztegietan topatzen ditugunak; markak ere badira, jendeak erabiltzen dituen espresioak, Whatsapp, izen propioak, bereziak, leku izenak... Edozer gauza. Horrek balio du makinak jakiteko hitzen atzean zer dagoen. Hitz bat entzutean, “kutxa” edo “katu”, gauza aunitz etortzen zaizkigu, kutxei eta katuei buruz gauza pila bat dakizkigu. «Egiten duguna da bai Wikipediatik eta bai testu dokumentuetatik katuari buruz agertzen den informazio guztia prozesatu eta formula matematikoen bidez mapan kokatzeko erabili, bestela makinak ‘katu’ irakurtzean lau hizki dituela soilik dakielako».

Hizkuntzarekin lan egiten duten programa gero eta gehiago daude: itzulpen automatikoa, galdera-erantzun sistema, bilatzaileak eta bertze aunitz, adibidez kotxean irratiari erratea «jarri Info7» edo Osakidetzara deitzean makina batekin mintzatzea txanda hartzeko. Baina ongi funtzionatzeko hitz bakoitzaren gibelean zer dagoen jakin behar dute.

IXA ikerkuntza taldeko kideak dira proiektuan dabiltzanak. Talde honek 26 urte darama lanean. Hizkuntzaren prozesamendua aztertzen du, gero eta pisu handiagoa hartzen ari den arloa. Egitasmo ugari dituzte esku artean: itzultzaile automatikoa, zuzentzaile ortografikoa... 60 bat ikertzaile dira. Heren bat hizkuntzalariak dira eta bi heren informatikariak.

Sare neuronalak erabiltzen dituzte. Horiek egiten dutena da hitz baten esanahia ikasteko inguruko hitzen esanahia kontuan hartu. «Hori da gure oinarri linguistikoa; gainerakoa matematika da».

Ligako puntuak bezala

Mapak nola osatzen dituzten eta beren eguneroko lana nolakoa den galdetuta, irriz eta aho batez erantzun digute: «Ordenagailuaren aurrean. Alde batetik sare neuronalak dituzu, eta, bestetik, baliabideak: testu edo corpus handiak ala grafoak. Programan sartzen dituzu eta ordu, egun edo asteetan egon ahal da makina prozesatzen. Amaitzen duenean fitxategi erraldoi bat daukazu. Hitz bakoitzak bektore bat du esleituta. Probak egin, ebaluatu eta akatsak konpontzen ditugu. Jende gehiago dago gauza hauek ikertzen eta beren lana irakurri, aztertu, ideia berriak pentsatu eta martxan jartzen ditugu. Zenbaki batzuk sortzen 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 Processing) ematen dituzte. Etortzen denak ez du zertan informatikaria edo hizkuntzalaria izan. Matematikariak, fisikariak, itzultzaileak, Irakasle ikasketak egindakoak eta filosofoak ere badira taldean. Hagitz programatzaile onak, gainera, gauzak ikasi egiten baitira. •