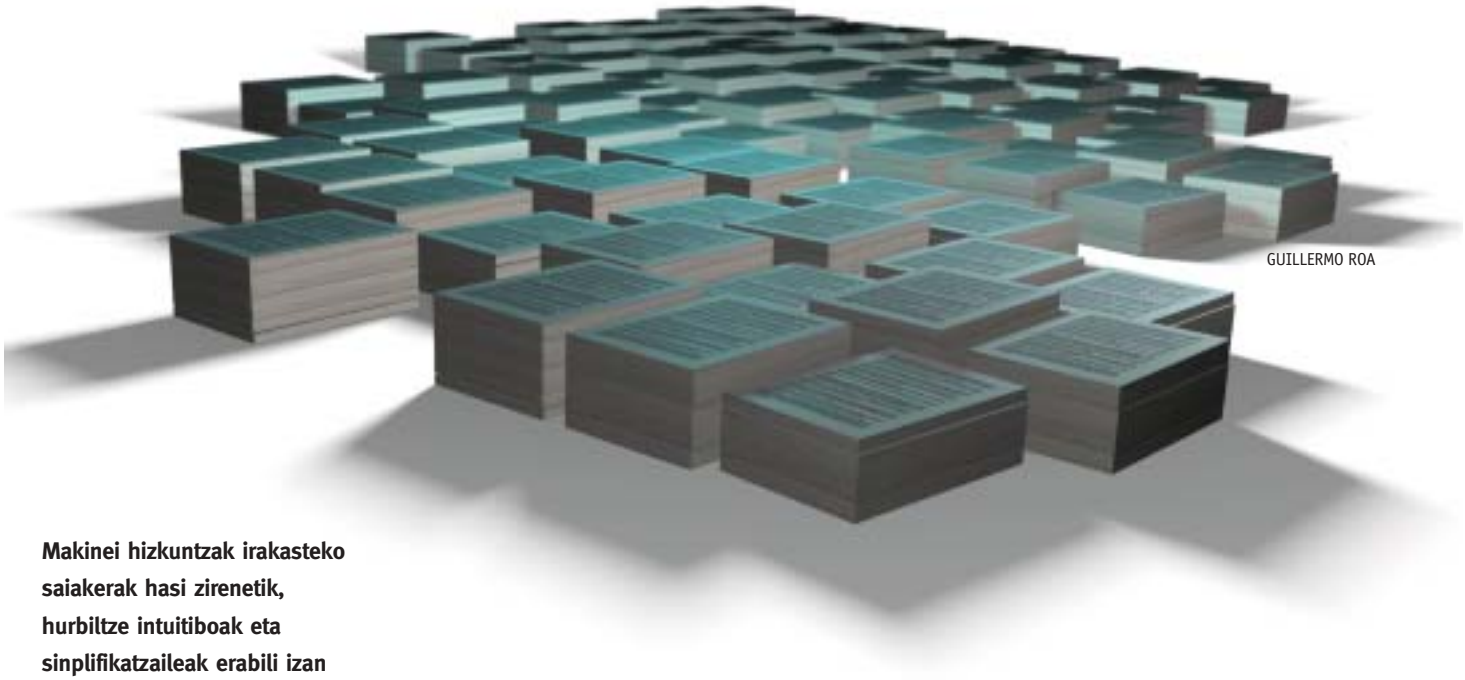
Igor Leturia Azkarate
Informatikaria eta ikertzailea

TAMAINAK AXOLA DU

testu-bilduma erraldoiak, hizkuntzaren prozesamenduan beharrezkoak



GUILLERMO ROA

Makinei hizkuntzak irakasteko saiakerak hasi zirenetik, hurbiltze intuitiboak eta sinplifikatzaileak erabili izan dira. Hizkuntzalarien ezagutza linguistikoak makinek ulertzeko moduko egituretara pasatzen ziren informatikarien laguntzarekin, eta egitura horien bidez tratatzen zen hizkuntza. Azken urteetan, baina, gero eta gehiago erabiltzen dira corpus handietan eta metodo estatistiko hutsetan oinarritutako teknikak.

Hizkuntzaren prozesamendua ia-ia ordenagailuen sorreratik existitzen da. Joan den mendeko 40ko hamarkadan sortutako lehen makina elektroniko programagarriak, Bigarren Mundu Gerra medio, batez ere mezuak deszifratu eta kodeak apurtzeko erabili ziren, baina, gerra amaitu ondoren, hizkuntzaren prozesamendua asko lantzen hasi zen, batez ere itzulpen automatikoaren arloan.

Hasiera haietan, batez ere matematikariak aritzen ziren horretan, eta oso teknika sinpleak erabiltzen zituzten, kriptografiaren ohiturek eraginda: funtsean, hiztegien eta hitz-ordenaren aldaketen bidez lortu nahi zuten itzulpen automatikoa. Baina segituan konturatu ziren hizkuntzak hori baino gehiago zirela, eta eredu linguistiko konplexuagoak erabili beharra zegoela. Hala, taldeetan hizkuntzalariak sartzen joan ziren, eta Saussure eta Chomskyren teoriak aplikatzen. Geroztik, eta hamarkada askotan zehar, hizkuntzaren prozesamenduko alor guztietan (morfologian, ortografia-zuzenketan, sintaxian, adieren desanbiguazioan...) hurbiltze bat izan da nagusi: hizkuntzalarien intuizioan oinarritutako ezagutza ordenagailuek tratatu ahal izateko moduko egitura sinpleetara egokitzea (erregelak, zuhaitzak, grafoak, programazio-lengoaiak...).

Adam Kilgarriff

“Nahiko erraz bil daitezke testu hutsezko datu-base erraldoiak”

Hizkuntzaren prozesamenduan, corpusen erabilerak iraultza ekarri du azken urteotan, eta, zalantzarik gabe, Adam Kilgarriff ingelesa horren lekuko izan da. Urteak eman ditu ingelesezko corpusekin lanean, eta, gaur egun, Internet bera corpus gisa erabiltzen dutenen artean erreferentzia da. Ildo horretan lan egiteko balio duen Sketch Engine **(www.sketchengine.co.uk)** *tresnaren sortzailetako bat da. EHUko IXA taldeak antolatutako hizkuntzaren prozesamenduari buruzko SEPLN 2009 kongresuan izan zen, Donostian.*

Zein dira zailtasun nagusiak makina batek hitz egin ahal izan dezan?

Asko daude. Gizakiak gauza asko dakizki. Jaio ginenetik ari gara gauzak ikasten, ikusten, sumatzen... jakinduria handia dugu buruan, eta badakigu zein den ideien arteko erlazioa. 50 urteko ikerketa ez da nahikoa adimen artifizialak gauza bera egin ahal izateko. Guk datu guztiak buruan bilduta ditugu. Makinek hitz egiteko duten zailtasun handiena hortik dator: oraindik ez dugu asmatu modurik material pilo bat ordenagailuari emateko harentzat erabilgarria suerta dadin.

Bestalde, hizkuntzarekin lotutako arazo asko ditugu. Edozer gauza esateko modu asko dago, eta ordenagailuentzat oso zaila da ulertzea bi esaldi ezberdinek ideia bera adierazten dutela. Ez du ulertuko “Leku hau zoragarria da” eta “Hondartza eder bat dago hemen” esaldiek oinarrian ideia bera adierazten dutela. Kontrakoa ere gertatzen da; esaldi bakar batek esanahi bat baino gehiago izan dezake. “Sagu bat ikusi dut” esaldiak esanahi ezberdina du Miramar Jauregian edo biologiako laborategi batean.

Horiek dira arazo orokor nagusiak (baina badira beste arazo txiki asko).

Behar-beharrezkoa da adimen artifiziala erabiltzea hizkuntzaren prozesamenduan?

Hizkuntzaren prozesamenduan gero eta gauza gehiagorako ari dira erabiltzen ikasketa automatikoa. Baina adimen artifiziala ez da gauza bakar bat; estrategia ezberdin asko garatu dira esparru ezberdin askotan. Hizkuntzaren tratamendurako interesatzen zaidan hurbilketa da patroiak aurkitzea datu askotatik abiatuta. Ume batek hori egiten du; patroiak bilatzen ditu soinuetan, esanahietan, gramatikan eta abarretan, eta, gerora, horrek sortzen du umearen lexikoa. Gure lana ere horixe da. Adibidez, hitz batetik abiatu, eta hitz horrekin batera testuinguru berean azaltzen diren hitzak bilatzen ditugu.

Ikasketa automatikoari esker, adibidez, patroiak bila daitezke, eta ezagutza eraiki ordenagailu bidez. Beraz, hizkuntzaren prozesamenduaren arazo nagusietako bati aurre egiteko modu bat da, alegia, hitz bakar batek esanahi bat baino gehiago dituen kasua ebazteko bide bat. Hori egin dezakegu corpus handiak erabiltzen baditugu.

Corpusik onena Internet al da?

Helburuaren araberakoa da. Nik egiten ditudan lan askotan, zenbat eta datu gehiago erabili, orduan eta hobeto funtzionatzen du. Baina sareak zailtasun batzuk ere ekartzen ditu. Spam asko dago. Beraz, datu horiek kudeatzeko estrategia onena da Googlek eta Yahook erabiltzen dutena: bildu webgune asko eta asko, eta testua bakarrik bilatu, informazio gutxiagorekin lan egiteko (gigabyte batean bideo gutxi sartzen dira, baina testu-kantitate izugarri handia dago). Horrela, nahiko erraz bil daitezke testu hutsezko datu-base erraldoiak. Gaur egun, ingelesezko handienak 5.500 milioi hitz ditu. Eta horrelakoetatik patroi asko aurki daitezke.

Arazo bat dago, ordea: solasean arituko den makina batek hitz egingo duen hizkuntzak ez du izan behar, adibidez, zientzialariek artikuluetan idazten duten estilo bera. Gu solasean ari gareneko hizkuntza izan beharko luke. Beraz, horretarako, ez du balio artikuluetan edo egunkarietan idatzitako testuez osatutako corpus handi batek. Solasean oinarritutako corpus handi bat behar da, txatetan oinarritutakoa. Baina testu horiek biltzea zaila da, eta konfidentzialtasunak are zailagoa egiten du. Gure ikerketarako, blogetako testuak jasotzen ditugu, haietan idazkerak formaltasun gutxiago izaten baitu.

GUILLERMO ROA

Baina metodo horiek ere beren mugak dituzte. Alde batetik, hizkuntzalaririk onenek ere ezin dute kontuan izan hizkuntza batek eskaintzen duen kasuistika guztia; bestetik, hizkuntzek konplexutasun eta aberastasun handiegia dute egitura sinpleen bidez adierazteko. Muga horiek, gainera, are handiagoak dira solaserako hizkuntzan. Hala ere, beste biderik ez zegoen; garaiko makinen ahalmena kontuan izanda hori zen hizkuntzarekin aritzeko modu bakarra. Eta, teknika horien bidez, aurrerapena mantso samarra izan da urte askotan.

©ISTOCKPHOTO.COM/CHIEFERU

CORPUSEN ETA ESTATISTIKAREN ETORRERA

Azken bi hamarkadetan, baina, hurbiltze enpirikoago bat ari da nagusitzen hizkuntzaren prozesamenduan, testu-bilduma handien ustiaketan eta metodo estatistikoetan oinarritua. Ezagutza intuitiboan oinarritu beharrean, hizkuntza-lagin erreal handiak, hau da, corpusak, erabiltzen dira hizkuntzaren ahalik eta kasu gehien kontuan hartzeko. Eta horien gainean estatistika edo ikasketa automatikoaren gisako metodoak erabiltzen dira, teknika linguistiko gutxi erabiliz. Hizkuntza egitura konputagarrien bidez modelizatzen saiatzen diren kasuetan ere, ereduak corpusetatik erauzten dituzte automatikoki. Horregatik, metodo estatistikoekin lan eginda, makina batek hitz egiteko ahalmena izan dezan, testu-bilduma erraldoi bat eta bilduma horrekin lan egiteko baliabideak izan behar ditu eskuragarri.

Bi faktorek eragin dute nagusiki metodologia-aldaketa hau. Batetik, gaur egungo ordenagailuek, lehengoek ez bezala, datu-kopuru ikaragarriak maneiatzeko gaitasuna dute. Bestetik, inoiz baino testu gehiago dago eskura formatu elektronikoan, batez ere Internet sortuz geroztik.

Hala, corpusak eta teknika estatistikoak erabiltzen dira ortografia-zuzentzaileetan (hitz okerraren antzeko testuinguruak corpusetan bilatuta), itzulpen automatikoan (itzulpen-memoriak edo webgune eleanitzetako testuak erabiliz, hitz, sintagma edo esaldi ahalik eta handienen itzulpenak estatistikoki lortzeko), adieren desanbiguazioan, terminologia-erauzketa automatikoan... Eta orokorrean esan daiteke zenbat eta corpus handiagoak izan orduan eta emaitza hobeak lortzen dituztela sistemek. Adibidez, Googleko Franz Joseph Och-ek bere itzulpen automatiko estatistikoaren sistema aurkeztu zuen 2005eko ACLren (Association for Computational Linguistics) kongresuan, 200.000 milioi hitzeko corpus baten gainean entrenatutakoa. Eta, geroztik, haien sistema da itzulpen automatikoan erreferentzia nagusia eta lehiaketa guztiak irabazten dituena. Eta antzera gertatzen da beste alorretan ere.

ETORKIZUNA, HIBRIDAZIOA

Alabaina, metodologia honek ere mugak ditu. Hizkuntza eta ataza batzuetan, corpus benetan erraldoiak erabiltzen dira jadanik, eta esan daiteke honezkero goi-muga jo dutela, oso zaila baitute lortutako emaitzak askoz gehiago hobetzen jarraitzea. Beste hizkuntza eta alor batzuetan ez dago hain corpus handirik, eta metodo estatistiko hutsekin ezin dira hain emaitza onak lortu.

Horregatik, metodo estatistikoak hobetzeko azken aldiko joera da teknika linguistikoekin konbinatzea, eta metodo hibridoak sortzea. Eta etorkizunean ere hori izango da bidea hizkuntzaren prozesamenduan aurrera egiteko. Makinek laster hizkuntza ulertu eta egoki trata dezaten nahi badugu, eta makinek hitz egitea nahi badugu, beharrezko izango da matematikariak, informatikariak eta hizkuntzalariak eskutik joatea.●

Hizkuntza eta ataza batzuetan, corpus benetan erraldoiak erabiltzen dira jadanik, eta esan daiteke honezkero goi-muga jo dutela, oso zaila baitute lortutako emaitzak askoz gehiago hobetzen jarraitzea.