- Artikuluak
- Argazkiak
- Sarean
- Agenda
- Hileko zerua
- Dosierrak
- Multimedia
- Elhuyar aldizkaria
- Elhuyar 1974-1996
- Zientzia-hiztegiak
- Euskara teknikoa
- Dibulgazio-liburuak
- CAF-Elhuyar sariak
- Teknoskopioa

[asteko laburpena]
[gaiak]

XML RSS jarioak
podcasta
bideocasta

Erabiltzaileak

ARTIKULUAK

Makinen hizkuntzari buruz hizketan. Adituen solasa

2009/11/01 | Roa Zubia, Guillermo |

Hizkuntzaren prozesamenduan dauden joerez eta euskarak beste hizkuntzekin alderatuta dituen berezitasunez hitz egiteko elkartu ditugu aditu batzuk. EHUko IXA taldeko Kepa Sarasola, Iñaki Alegria eta Eneko Agirre informatikariekin izan gara. Hain zuzen ere, IXA taldeak aurtengo hizkuntzaren prozesamenduari buruzko SEPLN kongresua antolatu du Donostian, eta gai horretako aditu asko bildu ditu.

Zein dira, gaur egun, hizkuntzaren prozesamenduaren erronka nagusiak?


(Argazkia: Iñigo Ibáñez)

Eneko Agirre: Nik uste dut ulermenarekin lotutako kontuak direla. Azken urte hauetan egindako ikerketa salto kualitatibo handia izan da, baina horrek ez du esan nahi makinak orain "ulertzen" gaituenik. Nik uste dut pauso txikiak eman direla, eta makinek gauzatxoak ulertzen dituzte gero eta eremu gehiagotan. Zer den leku bat, adibidez. Abizenekin beti dago arazoa; Azpeitia zer da: pertsona bat edo leku bat? Edo enpresa bat? Gauza horiek ulertzen hastea aurrerapauso bat da. Eta nahiz eta oso sinplea iruditu pertsonei, testuingururik gabe zailak dira. Hortaz, erronka da makinari horrelako ezagumendu-zatitxoak irakastea.

Izan ere, corpusetan oinarritutako metodo matematiko eta estatistikoak nolabait goia jotzen ari dira; egin zezaketena egiten ari dira, eta zailtasunak dituzte hortik aurrera egiteko. Erregeletan oinarritutakoek ere eman zuten berea, eta pixka bat trabatuta gelditu ziren. Beraz, nik uste dut orain erronka dela erregelak testuetatik ikastea, eta corpusetatik saiatzea erregela haiek nola edo hala ikasi eta kontrastatzen, eta jakitea zer ikasi duen ondo eta zer gaizki.

Kepa Sarasola: Gaur egun erronkak zein diren ikusteko, bi maila egon daitezke: bata, aplikazioak eta, bestea, hizkuntzaren barruko tripak, oinarrizko tresnak, gero aplikazioetan erabili behar direnak. Esan daiteke lexikoan ditugun beharrak gaur egun ia % 100 beteta daudela. Orain dela 20 urte ez zegoen hiztegi konputazionalik, denak paperezkoak ziren. Orain, aldiz, Interneten dituzu hitz guztien esanahiak, nola esaten diren beste hizkuntzetan eta abar. Morfologiaren aldetik, hizkuntza zailetarako (euskara bezalakoetarako), % 95-98 beteta dago. Sintaxian % 90 ondo egiten du ingeleserako.

Orduan, zeri begira gaude? Bada, semantikari eta pragmatikari. Eta horretarako, hemen, aldaketa ikaragarri bat dago. Orain dela 20 urte, edozein gairi buruz hitz egiteko, ez genuen zeri heldu. Gaur, adibidez, Wikipedia dugu, edo Wordnet, Internet bera eta abar. Testuen esanahia ulertu ahal izateko baliabide berriak ditugu orain. Eta horrek ate bat ireki digu, baina oraindik ez da askorik landu.


Kepa Sarasola. Arg.: Iñigo Ibáñez.

SEPLN 2009 kongresuan horri eman al zaio indarra?

Iñaki Alegria: Kongresura hausnarketa egiten duten hizlari gonbidatu batzuk etorri ziren. Uppsala Unibertsitateko Joakim Nivre sintaxian adituak, adibidez, aditzera eman zuen corpusen erabileraren bitartez sintaxiaren arazoa ez dela % 100 ebazten, baina oso landuta dagoela. Semantikaren ildotik, nolabait Enekok aipatu duen egoera aurkeztu zuen. KYOTO proiektua ere aurkeztu zuten: wiki plataforma baten bidez hitzen eta terminoen esanahiak definitu ahal izateko sistema baten proiektua. Bestalde, datuetatik ezagutza eraucteari buruz ere hitz egin zuten. Eta Kataluniako Unibertsitate Politeknikoko Horacio Rodriguezek eman zuen hitzaldian aipatu zuen adimen artifizial klasikoaren erronka batzuei saiatu behar dugula berriro heltzen, baina datu gehiagorekin, eta bide berriago batzuetatik. Eta ni ere, apur bat, iritzi horretakoa naiz.

Bide horretan Googlek oso emaitza onak lortu ditu adimen artifizialaren oinarrizko metodo batzuk erabilita. Baina ezagutza sakonagorik erabiltzen ez badute, epe laburrean berrikuntza gutxi aterako da hortik.

Antzeko edukiak:

2009/11/01
Tamainak axola du: testu-bilduma erraldoiak, hizkuntzaren prozesamenduan beharrezkoak

2002/11/22
Euskara eta Ingeniaritza linguistikoa

2007/07/01
Andy Way: "Itzulpen automatikoaren erronka handiena kalitatea da"

2003/09/01
Hizking21 hizkuntz ingeniaritza XXI. mendearen atean

dokusare

Artikulu berrienak:

- Lur arraroen bila itsaspean 2011/07/09
- Edzard Ernst: "Tratamendu alternatiboek ez dute benetako eraginik pazienteengan" 2011/07/05
- Malaspina 2010, klima-aldaketaren uberan 2011/06/18
- Elhuyar Fundazioak eta DIPCk bideo amateurren ON ZIENTZIA lehiaketa jarri dute abian, bigarrenez 2011/06/15
- Energia arriskutsua ontzi barruan 2011/06/11


Elhuyar Z. eta T. 258. zk.

Bilatu:
Hitza:
Bilatu
Bilaketa aurreratua

Google aipatu duzu, zenbateraino ari dira enpresa handi horiek hizkuntza prozesamenduan ikertzen?

I. A.: Nik uste dut Google asmatzen ari dela eginda dagoenari etekina ateratzen. Inbertsio handiak egiten ditu, etekin ona ateratzen die; ospea lortu du, eta marka bat egin du. Ezagutza hori edo tresna horiek jende guztiarentzako aplikazioetan eta industria-mailan integra litezke. Baina ez dute ematen nahiko informazio, eta aplikazioen eskaera espero baino txikiagoa da.

E. A.: Ikerkuntzan ez dakizu nor etorriko den ideia onarekin. Nahiz eta ikerkuntza-talde handia izan, beharbada ideia onak ez dira hortik aterako; ezin da hori iragarri. Horregatik, enpresa handiek, Googlek adibidez, haien proiektuak garatzeaz gain, ikertzaile arrakastatsuak fitxatu egiten dituzte.


Iñaki Alegria. Arg.: Iñigo Ibáñez.

Jende asko joan da Googlera. Estatu Batuetan aipatu izan dute ikertzaile onenak Googlera joan direlako. Gazteen artean jende asko hartu dute, eta unibertsitateetan hori nabaritu dute. Jendea hara joan da; gero esan dute Googlen dena ez dela hain polita, baina oso gutxik egin dute ospa handik.

I. A.: Arlo honetan, zehaztuta dago zein diren dirua ematen duten aplikazioak. *Killer applications* esaten zaie. Historikoki, hiru aplikazio-mota sartu dira talde horretan: itzulpen automatikoa, *proofing tools* (alegia, testu-editoreetarako tresnak, zuzentzaileak batez ere) eta bilaketa. Hain zuzen ere, Googleren hasiera bilaketaren mundua izan zen. Orain, itzulpen automatikoa tratatzen ari da, eta, azkenaldian, telefonoetako sistema eragileen arloan eta *proofing tools* etan ere ari dira sartzen. Nolabait, arriskua izan daiteke Googlek ikerketa horiek guztiak monopolizatzea.

Zuen lanean eragina izango du arrisku horrek, ezta?

K. S.: Gu, alde batetik, pozik gaude argi eta garbi ikusten delako lantzen ditugun teknikak baliagarriak direla. Behin eta berriz frogatzen da. Baina, bestetik, kezka dugu Googlek zenbat datu dituen ikusita, haiek bakarrik dituztelako. Haiek dakite jendeak zer eskatzen duen, zer bilatu nahi duen. Eta bilaketaren emaitzetan jendeak zer aukeratzen duen. Haientzat oso garrantzitsua da hori, sistema hobetzeko. Hitz bat eskatuta jende gehienak laugarren aukera klikatzen duela; eta, handik gutxira, laugarren hori lehengoa izango dela. Erabileraren datu horiek oso garrantzitsuak dira; baina Googlerenak dira.

E. A.: Googlek badaki berrikuntza dela aurrera egiteko bidea. Energia guztiak berrikuntzara zuzentzen dituzte.


Eneko Agirre. Arg.: Iñigo Ibáñez.

I. A.: Eta diruari ematen diote lehentasuna. Dirua non, haiek han. Eta horrek ondorio batzuk ditu. Adibidez, Googlek euskaraz oso gaizki bilatzen du. Eta esan zaie. Baina ez zaie interesatzen. Une batean erabaki zuten gehienez berrogei hizkuntzarekin lan egitea. Gainerakoetan, hitzez hitzeko bilaketa egiten dute. Hori arazo bat da, baina markak indar handia du. Gainera, aplikazio askotan integratzen da eta abar. Baina, gaur egun, Elebila aplikazioak askoz hobeto bilatzen du euskarazkoa.

Zer egoeratan dago euskara hizkuntzaren tratamendurako beste hizkuntzekin konparatuta?

I. A.: Ingelesa da erreferentzia. Esate baterako, kongresura Etiopiako ikertzaile bat etorri zen. Han, amhareraz hitz egiten dute. Hizkuntza semitikoa da, beste teklatu mota bat erabili behar dute, baina, telefono mugikorretan horrelako teklaturik ez dagoenez, mezuak ingelesez bakarrik bidaltzen dizkiote elkarri.

Argi dago euskara txikia dela. Ikuspuntu ekonomizista batetik, eskaera txikia du, eta, beraz, arazoak daude. Ikerketa mailan, berriz, gu pozik gaude. Arlo batzuetan, behintzat, erreferentzia bat gara beste hizkuntza txiki batzuetarako. Corpusetan oinarritutako aplikazioek inbertsioak egitea eskatzen dute, corpusak berak lortzeko.

E. A.: Hizkuntza gisa, euskarak badu berezko tipologia bat, baina konputazio-aldetik ez da bereziki zailagoa beste hizkuntzekin konparatzen badugu. Morfologia tratatzea zailagoa bada ere, beste alor batzuetan, fonetikan esate baterako, oso erraza da. Hizkuntza bakoitzak bere alde zailak eta errazak ditu, baina oro har, hizkuntzaren ezaugarri guztiak kontuan hartuta, hizkuntza guztien zailtasuna antzekoa da.


(Argazkia: Iñigo Ibáñez)

Eta, beste hizkuntzekin konparatzeko, ikusi behar da hizkuntza bakoitza hiztun-kopuruaren arabera. Nik uste dut euskara nahiko gertu dagoela gehien hitz egiten diren hizkuntzetatik. Alde handiena erabiltzen diren corpusen tamaina txikia da; nik uste dut hori dela gabezia nabarmenena euskaraz. Ingelesez, adibidez, milaka milioi hitzeko corpusak daude. Eta makinek corpus handietatik ikasten dute. Baina, baliabideen arabera, zerrendaren goialdean gaude.

K. S.: Hiztunen kopuruaren araberako zerrendan 256.a ikusi nuen euskara, eta ikerketan lehen 50en artean gaude. Hori zergatik? Bada, laguntza ofizialak egon direlako, eta, nire ustez, honetan gabiltzanok gauzak ordenatuta egiten ditugulako. Modu ordenatu eta planifikatu batean egin ditugu gauzak. Une batean sortzen dituzun tresnak eta baliabideak gerora ere baliagarriak dira. Modu inkrementalean egiten dugu lan.

IXA taldekoek euskararen prozesamenduan egiten dute lan. Ez dira bakarrak. Baina robot batek euskaraz hitz egiteko ahaleginetan erreferentziazko ikertzaile dira. Enpresa handiek, adibidez, euskarazko aplikazioak garatu nahi izango balituzte, seguru asko, haiengana jo beharko lukete. Besteak beste, ANHITZ proiektuaren garapenean parte hartu dute, zientziako galderei erantzuten dien pertsonaia birtual bat sortuz. Hitz egiten duen robot bat, azken batean. Hizkuntzaren prozesamenduaren eredu ona da; kanpotik ikusita, ANHITZ ez dirudi aplikazio iraultzailea denik, ez baitu fikziozko robot batek bezain azkar eta erraz erantzuten. Proiektuaren atzean dagoen lana ezagutzen duenak, aldiz, oso balorazio ona egiten du. Asko dago egiteko hizkuntzaren prozesamenduaren arloan; ez dago zalantzarik. Baina eginda dagoena lan ikaragarria da, horren zalantzarik ere ez dago.

Imma Hernaez: "Gaur egungo sintesi-sistemen ahotsak arras ulergarriak dira"

Imma Hernaezek EHUko Aholab laborategian egiten du lan. Makinek ahotsa ezagutzeko eta sintetizatzeko sistemetan aditua da. Besteak beste, ANHITZ proiektuan hartu du parte, zientzia-galderei erantzuten dien pertsonaia birtual bat egiten. Proiektu horretan, pertsonaiak ahotsa ezagutzeko eta hitz egiteko sistemak garatu zituzten Hernaezek eta Aholab laborategiko langileek.

Zein dira ahotsa ezagutzeko eta sintetizatzeko gainditu beharreko zailtasun nagusiak?

Zailtasunak ez dira berdinak ezagutzan edo sintesian. Ahotsa ezagutzeko, hizkuntzaren barietateak izateak berak zailtzen du lana, dialektoak, azentuak, erregistroak eta abar baitaude. Gainera, ahotsa oso aldakorra da, hainbat faktoreren arabera. Pertsonaren aldarteak, osasunak, eguneko orduak eta beste faktore batzuek aldatu egiten dute hizketa. Eta, horretaz gain, inguruko arazoak ere izaten dira: zarata, audio-sistemen kalitatea eta abar.

Ahots sintetikoari naturaltasuna, bat-batekotasuna eta gizatiartasuna ematea da zailena; alegia, ahotsari nahi dugun 'nortasuna' ematea.

Zure ustez, zer dago gaindituta eta zer ez?


(Argazkia: Imma Hernaez)

Ahotsaren ezagutzan, ezagutu beharreko hiztegia murritza denean eta sistemari ahotsa hitzez hitz ematean, oso emaitza onak lortzen dira, inguruko baldintzak txarrak izanda ere. Baldintza horietatik aldentzean hasten dira arazoak: bat-bateko hizketarako (hau da, murriztu gabeko hiztegia duen eta era jarraituan ebakitzen den hizketarako), ez dira lortzen oso emaitza onak, oraindik. 'Pilotu' moduko mikrofonoa erabili behar da nahitaez, eta sistema egokitu egiten da hizlariaren ahotsera; alegia, trebatu egiten da sistema, hizlariaren zenbait ahots-lagin erabiliz.

Gaur egungo sintesi-sistemen ahotsak arras ulergarriak dira. Ahotsaren naturaltasuna ere lortzen da, esaldiak edo paragrafoak laburrak baldin badira eta testuak irakurtzean estilo neutroa erabiltzen bada. Emozioa edo adierazkortasuna adieraztean, halere, sintesi-sistemek porrot egiten dute oraingoz; naturaltasunetik hurbil dauden gaur egungo sistemak corpusetan oinarrituta daude, hau da, datu-base erraldoiak baliatzen dituzte, eta amaierako kalitatea datu-base horien tamainaren araberakoa da: zenbat eta handiagoa datu-basea, orduan eta hobea kalitatea.

Gainera, pertsona bakar baten ahotsa da beti, eta, ahotsa aldatu nahi bada, datu-base berriak egin behar dira. Hortaz, metodorik onena litzateke datu-base txikiagoak erabiltzea, baina, ahots desberdinak sortzeko, zenbait parametro aldatzea ahotsa sortzeko erabilitako ereduari; oraingoz,

hala ere, ez dakigu doi-doi zein izan behar duten parametro horiek, seinalearen kalitatean galera esanguratsurik izan ez dadin.

Nola dago euskara beste hizkuntzekin konparatuta? (Ez dakit hizkuntza berezia den berez ahoskeraren ikuspuntutik).

Ikerkuntzaren ikuspuntutik, euskara ez dago beste hizkuntzetatik oso urrun, batez ere metodoei eta teknikei begiratzen badiegu. Bestelakoa da egoera ikuspuntu komertzialetik (ezagutzaren arloan, batez ere): sistema komertzialak eraikiko badira, enpresa garatzaileek datu-base estandarrak behar dituzte sistemak trebatzeko eta testatzeko, beste hizkuntzekin erabiltzen duten softwarea erabili ahal izateko. Eta horrelakoak oso gutxi ditugu. Bestalde, orain arteko garapenak euskara baturako baino ez dira egin, oro har, eta ahozko euskararen errealitatea ez da gure ondoko hizkuntzena bezalakoa (Europako hizkuntza nagusiena, esaterako). Batuaren eta euskalkien arteko distantzia oso handia izan daiteke, eta ezagutza-sistemak ez badira euskalkietara egokitzen, litekeena da gizartearen atal mugatu batek baino ez erabiltzea halako sistemak.