Hizkuntza-teknologia «Datu Handi»en garaian: bilatzaileak, itzultzaileak…

Itziar Aduriz (Bartzelonako Unibertsitatea)
Iñaki Alegria, Olatz Arregi
Arantza Diaz de Ilarraza, Kepa Sarasola
(UPV/EHU)

Sarrera

Datu Handien garaian bizi gara, Datu andanaren garaian, letra larriz. Makrodatu ere esan izan zaio ingelesezko *Big Data* kontzeptuari. Zer dago, ordea, kontzeptu horren atzean? Adibidez, egunero, milioika pertsona aritzen da argazkiak, bideoak eta testuak bidaltzen eta partekatzen. IBMren arabera, munduan «2012. urtean, egunero 2,5 exabyte ($2,5*10^{18}$ byte) sortu ziren. Eta, hortik, % 75, testua, hizketa edo bideoa da».[1]

Zenbat da exabyte bat? Gmail-ek 2012an exabyte bat memoria behar zuen bere erabiltzaileen mezu guztiak biltegiratzeko (260 milioi erabiltzaile eta 10.240 megabyte erabiltzaile bakoitzeko). Erraldoia zen zenbaki hori 2012an; erraldoia zen egunero mugitzen zen informazio kantitatea, eta urtez urte ikaragarri igotzen ari da.[2]

Gauzak (kopuruak) horrela, eta interakzio digitalaz ari garenez, tresna informatikoak behar dira informazio hori guztia modu eraginkorrean baliatzeko. Testuinguru globalizatu eta eleaniztun honetan, hizkuntza-teknologia funtsezkoa da, pertsona, enpresa eta erakundeen arteko komunikazioa bideragarria izango bada.

Eta erronka ikaragarria da. Izan ere, testu andana eskuragarri hori fintzen den heinean, paraleloan, hizkuntza-teknologia ere handitzen eta hobetzen ari da. Orain dela 25 urte sortzen ziren tresnak, oro har, jakintza linguistikoan oinarritzen ziren. Mende berriak teknika estatis-

1. http://www.bbc.com/news/business-26383058
2. http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/

tikoen erabileraren nagusitasuna ekarri zuen. Eta, azkenaldian, ikasketa sakona (*deep learning*) eta neurona-sareak (*neural networks*) erabiltzen dira datu multzo erraldoi horiei ahalik eta zuku handiena ateratzeko.

Artikulu honetan, Ixa Taldearen oraingo jarduera orokorra azaldu nahi dugu, bi ildo nagusi bereizita. Lehenengo puntu orokorrean, azalduko dugu zelan ari garen lantzen aipatutako teknologia berri horiek, eta euskararekin ere erabiltzeko egiten ari garen saioak deskribatuko. Bigarrenean, berriz, hizkuntza-teknologietan baliabide urriko hizkuntzen ikuspuntutik bideratzen ari garen zenbait proiektu eta doktorego-tesi aipatuko ditugu. Ekarpen handitzat jotzen dugu puntu hori; izan ere, errotik aldatzen du hizkuntza «handi» eta hizkuntza «txiki»en arteko erlazioa eta bien arteko interakzioen norabidea. Ikuspegi horrek lan interesgarriak bultzatu ditu euskararen kasuan; berritzaileak, hizkuntza nagusien unibertsoan topatzen ez direnak, hain zuzen ere.

Senez aldizkariko «Besterena nuen neuregana» atal honetan, irakurtzen ari zareten honetaz gain, beste hiru artikulu idatzi ditugu euskararako aplikazioen berri emanez. Bata, itzulpengintza automatikoaz, gai hori bereziki lotuta dagoelako aldizkari honekin; bigarrena, testu-corpus handiak bildu eta aztertzen dituen *Lexikoaren behatokia* proiektuaz; eta azkena, Garaterm eta TZOS Internet-aplikazioez, zeinak lagungarriak baitira terminologia lantzeko.

Oraingo ikerketa eta etorkizunekoa Ixa Taldean

Ixa Taldean ia 30 urtez aritu gara hizkuntza (euskara gehienbat) lantzen konputagailuen eskakizunei erantzunez, eta ordenagailuak (programak) hizkuntzaren beharretara egokitzen. Une honetan, 57 kide gara: 16 hizkuntzalari, 38 informatikari, eta hiru teknikari; haietatik 45 doktoreak gara.

Sarreran aztertu dugunari jarraituz, testu eskuragarriaren tamaina egunez egun zelan handitzen den ikusita eta teknologia nola aldatzen den kontuan izanda, Ixa Taldeak gaurko garai digitalean lantzen duen estrategia aztertuko dugu, hau da, zer ikuspuntutatik erantzuten dien erronka berri horiei. Horretarako, eskuartean darabiltzagun lan garrantzizkoenak aipatuko ditugu: tresnak eta aplikazio berriak.

Horrela, aurten hasita, hurrengo bost urteetan ikertu nahi ditugun lerro nagusiak hauek dira: 1) itzulpen automatikoa, 2) eduki-erauzketa eta -kudeaketa, 3) irakaskuntzarako laguntzak, 4) komunikabideak eta sare sozialak lantzekoak, eta 5) hodeiko konputazioa. Ikus ditzagun banan-banan iker-lerro horiek.

Itzulpen automatikoa

Azken hamarkadan, Europako beste ikertalde batzuekin aritu gara elkarlanean zenbait proiektutan (OPENTRAD, OpenMT, OpenMT-2, TACARDI, TADEEP, ENEUS eta QTLeap). Une honetan, oinarri sendoa dugu estatistikan eta erregeletan oinarritzen diren itzulpen-

tekniketan. Sintaxia eta semantika era koordinatuan konbinatzen hasi gara. Eta gehienbat albisteen eta informatikaren arloetan esperimentatu dugu.

Hurrengo urteetan, teknika aurreratuagoarekin ere aritu nahi dugu, oraingo itzultzaileak hobetze aldera. Batetik, ezagutza linguistiko osoa erabili nahi dugu (informazio pragmatikoa barne, adibidez); eta, osoaz gain, landua (unitate lexikal konplexuak ezagutzen dituena, adibidez). Bestetik, ikasketa automatikorako teknika sakon berriak eta neurona-sareen inguruko teknikak aplikatu nahi ditugu itzulpen automatikoaren prozesuan. Haietako batzuk, hain zuzen ere, gaur egunean martxan ditugun tesi eta proiektuetako gaiak dira. Espainiera-euskara hizkuntza bikoteaz gain, ingelesa-euskara ere lantzen jarraitu nahi dugu. Zer domeinutan? Testu itzuli ugari eskaintzen dizkigutenetan; osasunaren eta kontsumoaren domeinuetan, esate baterako. Monografiko honetan, artikulu oso bat ondu dugu itzulpen automatikoaren inguruan, egin nahi duguna xehetasunez deskribatzen duena.

Eduki-erauzketa eta -kudeaketa

Batzuetan ez dugu asmatzen Interneten aurkitu nahi dugun informaziora heltzen edo aspaldian posta elektroniko bidez bidali ziguten mezu berezi hura aurkitzen. Izan ere, sarreran aipatu dugun bezala, eskuartean ditugun testu kopuru handi eta gaietan zabal horiek modu egokian kudeatzeko, alor berriak sortu dira azkenaldian: informazioaren berreskurapena (*information retrieval*)[3] eta informazio- edo datu-erauzketa (*information extraction*), esate baterako. Badira beste aplikazio mota berezituago batzuk ere; besteak beste: laburpen automatikoa (*summarization*), dokumentu-sailkatzaileak (*classification*), dokumentuak bideratzea (*routing*), dokumentuak multzokatzea (*clustering*), dokumentuak iragaztea (*filtering*).

Alor horietan egin ditugun lanak osatzen jarraitu nahi dugu, teknika eta aplikazio berrien bidez. Hona hemen, labur, haietako batzuen funtzioak:

- Testuetan agertzen diren entitate izendunak hautematea (pertsonak, enpresak, erakundeak…), baita toki-izenak, denborazko edota zenbakizko adierazpenak ere, eta entitate horien arteko erreferentziakidetasuna ezagutzea. Halako tekniken bitartez jakingo dugu, adibidez, «Macron», «Emmanuel Macron» eta «Frantziako presidentea» pertsonei buruzko entitateak direla eta, gainera, pertsona berari egiten diotela erreferentzia artikulu batean.
- Terminologia-erauzketaren tratamendua. Adibidez, testu batean azaltzen diren termino berriak edo esanguratsuenak zein diren jakiteko.
- Testuetatik kontzeptuen eta entitateen arteko erlazioak erauztea teknika konbinatu baten bitartez, honelako galderei erantzuteko: «Zenbat ordaindu du halako enpresak beste hura erosteko?».

3. Parentesi artean, ingelesez erabiltzen den terminoa azalduko da.

- Testu batean deskribatzen diren gertaerak edo gertaera-sekuentziak erauztea, informazioa berreskuratzeko beste bide bat den heinean.
- Kontzeptu-erauzketarekin lotuta, testu batetik mapa kontzeptual bat sortzea testuan azaltzen diren kontzeptuekin eta beren arteko erlazioekin.
- Kausa/efektu erlazioak identifikatzea eta horren araberako diskurtsoaren egitura automatikoki erauztea.
- Sentimendu- eta iritzi-analisiaren tratamendua. Adibidez, testu bat modu positiboan ala negatiboan idatzita dagoen jakiteko.
- Testuen arteko antzekotasun semantikoa aztertzea. Asko aurreratu da arlo horretan, eta, 2012az geroztik, SemEval lehiaketan parte hartzen ari gara, antzekotasun semantikoa neurtzeko atazan, eta, 2015az geroztik, antzekotasun semantikoaren interpretazioaren atazan.
- Domeinuei dagokienez, hizkuntza aztertzeko gure tresnak egokitu ditugu, medikuntzako, zuzenbideko eta turismoko testuetatik informazioa ahalik eta modu egokienean erauzteko.
- Oro har, edukiak aztertzeko baliagarri izango diren 5 tesi bukatzen ari dira 2017 honetan Ixa Taldean.

Irakaskuntzarako laguntzak

Hauek dira gure erronkak irakaskuntzaren arloan:

- Idazlanen ebaluazioan lagungarri izan daitezkeen neurri kualitatibo eta kuantitatiboak definitzea: hiztegi-aberastasuna, espresioen erabilera, errore ortografiko eta sintaktikoak, etab.
- Testu errealak erabilita irakasleei laguntza ematea material didaktikoa eta ariketak sortzeko. Domeinuaren araberako bilaketa aurreratuak egin ditzaketen sistemez baliatzea, tutoreek dauzkaten material didaktikoak adibideekin osatzeko. Puntu jakin honetan, aztertu nahi dugu ea lagungarria ote litzatekeen antzekotasun semantikoa lantzen duten sistemak ere erabiltzea.
- Euskara ikasten ari direnen idazlanak biltzen jarraitu nahi dugu, egiten diren errore tipikoak identifikatzeko eta ikasleei tresna lagungarriak eskaintzeko, horrelako errorerik egin ez dezaten.

Komunikabideak eta sare sozialak lantzeko tresnak

Eskuko telefonoek eta sare sozialek pertsonen arteko komunikazio moduak aldatu dituzte. Ez dago zalantzarik. Komunikabideek eta sare sozialek funtsezko papera jokatzen dute gaurko

gizartean. Alor horretan kokatzen dira, esate baterako: Wikipedia, blogak, foroak eta sare sozialak (Twitter, Facebook, LinkedIn, Pinterest, Tuenti, etab.). Horiek dira, gaur egun, pertsonen arteko komunikabide erabilienak.

Testuinguru horretan sortzen den informazio kantitatea, gorago ikusi dugun bezala, handia da, edukietan zabala eta, gainera, eleaniztuna. Izan ere, kasu gehien-gehienetan, hizkuntzaz ari gara, hizkuntza sortzen baita komunikatzeko. Beraz, informazio andana hori denbora errealean prozesatu nahi badugu, hizkuntza-teknologiaz baliatu behar dugu. Ez dago beste aukerarik.

2012an hasi ginen halako arloak jorratzen Ixa Taldearen ikerketa-lanetan. Txioen tratamenduan, batez ere itzulpen automatikoan, eta txioetako testuen normalizazioaren inguruan aritu gara lanean orduz geroztik. Wikipedia ere erabili izan dugu itzulpen automatikoan. Bestalde, Wikinews-eko testuekin, hiru hizkuntzako corpusa osatu dugu (euskara, ingelesa eta gaztelania), eta anotazio-prozesuak landu ditugu. Anotatutako elementuen artean, edukien kudeaketan[4] giltzarri direnak jo ditugu begiz lehenik; esate baterako: gertaerak, gertaera-sekuentziak, gertaeren arteko erlazioak, denborazko erlazioak, tokiak adierazteko adierazpenak eta partehartzaileak. Elementu horiek guztiak dokumentu barruan (*intra*) zein dokumentuen artean (*cross*) aztertu ditugu.

Code-Switching fenomenoa ere aztertzen ari gara azkenaldian. Foro sozialetan mintzakideak elkarrizketa berean hizkuntza batetik beste batera maiz pasatzen direnean gertatzen da fenomeno hori. Horrelakoetan, prozesamendu automatikoaren bitartez jakin beharko genuke detektatzen noiz gertatzen diren halako hizkuntza-aldaketak, eta, ondorioz, egokitzen ere jakin beharko litzateke, dagokion tresna linguistikoa aplikatzeko.

Iker-lerro nagusi hauek jorratu nahi ditugu:

- Oro har, sareko edukien analisia eta prozesamendua (Wikipedia, komunikabide sozialak, blogak, foroak, etab.).
- Kultura-ondareko espazio digitaletako informazio-erauzketa eta -errepresentazioa, ondorengo urrats batean espazio horietarako atzipen pertsonalizatuak sortu ahal izateko.
- Eduki masibo, heterogeneo eta ez-egituratuaren tratamendua, errepresentazio baliokide egituratu eta prozesagarriak lortzeko.
- Komunikabide sozialetako eta sare sozialetako testu-datuen fusioa eta integrazioa, bereziki interbentzioak noiz gertatzen diren kontuan hartuta.
- Albisteak kontsumitzeko interfaze aurreratuak.

Hodeiko konputazioa

Hodeian konputatzeak aukera ematen du konputazio-programak zerbitzu moduan dabiltzan ereduak garatzeko (*Software as a Service, SaaS*). Halako zerbitzuak Internet bidez eskaintzen

4. Ikus, gorago, edukien kudeaketari dagokion atala.

dira, eta erabiltzaileak baliatu ditzake konputagailu ahaltsurik eduki beharrik gabe (zerbitzaria, memoria handia, etab.), aplikaziorik instalatu gabe, eta ezagutza tekniko handirik gabe.

Hizkuntzaren prozesamenduan ohikoak izan diren *pipeline* arkitektura sekuentzialak (hau da, hainbat programa bata bestearen atzetik exekutatzea) makina birtualekin eta edukitzaileekin inplementatutako arkitektura banatu bihurtu dira hodeiko konputazioan. Horrela, testuen tratamendu paraleloa eta banatua erraz bidera daiteke. Horri esker, lana errazago eta arinago egin ahal izan da azken urteotan.

Konputazio-eredu berri hori landu duten lau proiektutan lan egin du Ixa Taldeak orain arte: 1) PATHS (*Personalised Access To cultural Heritage Spaces*, 2011-2013), 2) LoCloud (*Local content in a Europeana cloud*, 2013-2015), 3) OPENER (*Open Polarity Enhanced Named Entity Recognition*), eta 4) NEWSREADER (*Building structured event indexes of large volumes of financial and economic data for decision making*).

Euskararako eta baliabide urriak dituzten hizkuntzetarako hausnarketak. Zenbait proiektu berezi

Azkenaldian Ixa Taldean garatzen ari garen zenbait proiektu eta doktorego-tesik ez dute pareko erreferentziarik hizkuntza nagusien ikerketaren unibertsoan; izan ere, baliabide urriko hizkuntza batek duen problematika bereziari erantzun bereziak eman nahian sortuak dira. Haietako batzuen azalpen motz bat ekarri dugu atal honetara, batez ere DSS2016 proiektuaren barruan *Hirikia* izeneko *kultura-kaian* garatu ditugun proiektuak hizpidera ekarrita.[5] Ekimen horretan, beste eragile eta enpresa askoren arteko lankidetzari esker, kultura zabaltzeko hainbat ekarpen egin ditugu hizkuntza-teknologiatik.[6] Hona hemen garatu ditugun proiektu nagusien azalpen labur bat:

5. Olatz Arbelaitz. «Informatikatik ekarpena kulturara (DSS2016-Hirikia)»: http://www.naiz.eus/eu/hemeroteca/gaur8/editions/gaur8_2017-02-25-06-00/hemeroteca_articles/informatikatik-ekarpena-kulturara-dss2016-hirikia

6. https://www.ehu.eus/ehusfera/ifbloga/2016/06/14/informatika-kulturaren-alde-dss2016-la-informatica-herramienta-de-cultura/

- Ohar Eleanitzak[7] eta Gida Eleaniztunak[8] proiektuetan, QR kodeen bidez,[9] bisitariei azalpenak eta audioak ematen zaizkie, hainbat hizkuntzatan. QRak irakurtzen dituen aplikazio (app) bat erabiliz, mugikorrean ezarrita daukan hizkuntzan hartzen ditu mezuok bisitariak, irakurtzeko edo entzuteko moduan.
- Interprest aldibereko itzulpena antolatzeko azpiegitura merke bat da. Kabina beharrik gabe, Internet eta sakelako telefonoak erabiliz funtzionatzen du.
- Hiztegi-makina edo Hitz Machine[10] hainbat hizkuntzatan 100 hitzeko hiztegitxoak inprimatzen dituen makina sortu zuten Hirikilabs laborategian, eta Donostiako autobusgeltokietan egon zen makina hori.
- Behagunea proiektuak,[11] sare sozialetan (Twitter-en batez ere) DSS2016ri buruz ematen diren iritziak aztertuta, iritzi positibo, negatibo eta neutroen arteko proportzioa zenbatekoa den erakusten du ingurune bisual batean.
- Ondarebideak plataforma digitalarekin,[12] Donostialdean kultura aldetik ikusgarri diren obra eta elementu digital andanari bizia emateko erakusketa digitalak antola daitezke.

Donostiapediarekin, Wikipedia aberastu, eta Donostiari buruzko liburu bat idatzi da, 100 egileren artean.[13] Ixa Taldeak sortutako Matxin itzultzailea erabili ahal izango da laster Wikipediako artikuluak itzultzeko, Content Translation sistemarekin.[14]

Bukatzeko, 1. taulan, zenbait baliabide, tresna edo aplikazio interesgarri aipatuko ditugu, azken bost urteetan Ixa Taldean landuak, oso interesgarri izan baitaitezke baliabide urriak dituzten hizkuntzen ikuspuntutik; euskararenetik, esate baterako.

Oinarrizko tresnak	Corpusa
Elkarola: baliabideak eta demoak. **Dialekta**: aldaera dialektalak. **Eus-SRL**: rol semantikoak. **Diskurtso-markatzaileak**: hiztegia. **IxaKat**: euskarazko analizatzailea. **Ixa-pipes:** analizatzaileak (8 hizkuntza). **UKB**: hitzen esanahien desanbiguatzailea.	**Lexikoaren Behatokia**: komunikabideetako corpus monitorea. **Ancora-net**: baliabide semantikoen integrazio eleaniztuna.

7. http://www.ehu.eus/ehusfera/ifbloga/2015/05/08/kartelak-20-hizkuntzatan-albaola-museoan-posiblea-da/
8. http://gidaeleaniztunak.elhuyar.eus/
9. Ingelesezko *Quick Response Code* terminoaren laburtzapena.
10. http://euskarriak.eus/Euskarriak/hitzmachine/
11. http://behagunea.dss2016.eu/
12. https://ondarebideak.dss2016.eu/
13. https://eu.wikipedia.org/wiki/Wikiproiektu:Donostiapedia_Liburua
14. https://www.mediawiki.org/wiki/Content_translation

Medikuntza	Testu-sorkuntza
Deteami: sendagaien erreakzioen detekzioa txosten medikoetan. **Extrecm**: kontzeptu medikoen arteko erlazioen erauzketa. **Osaku**: osasun-txostenen kudeaketarako laguntza-sistema.	**Poetauto**: Bertso sortzaile automatikoa
Kitxua	**Kuba**
Kitxuaren lexikoa eta morfologia	**Kubako** eskola-hiztegia

1. taula: Azken bost urteetan Ixa Taldeak sortu dituen zenbait produktu

Ikertzaile eta profesionalen beharra: formazioa

Euskararen garapenerako alor estrategikoa den neurrian, profesional eta ikertzaileen behar handia dago, eta, gaur egun, hizkuntzaren prozesamenduan trebatutako pertsonen falta nabarmena dago, bai maila lokalean, Euskal Herrian, baita maila globalean ere, Europam edo mundu-mailan. Hori dela eta, formazio zehatza eskaintzea izan da beti Ixa Taldearen helburuetako bat.

Gurekin ikasi duten pertsona gehienak hizkuntza-teknologiaren arloan dabiltza lanean, ikerketan edo enpresan, baina, esan bezala, aditu gehiago behar dira. Teknologia berriek bultzatuta, azken hamarkadan, enpresa eta erakunde asko sortu dira Euskal Herrian hizkuntzaren industriako erronka berriei erantzuteko. Horren ondorioz, garapen teknologikoek hizkuntzaren industrien lan-merkatua handiagotu eta produktu berriak sortu dituzte. Horiei guztiei aurre egiteko, ezinbestekoa da jendea etengabe prestatzea hizkuntza-teknologietan.

Horregatik, Ixa Taldeak badu konpromisoa formazio espezifiko aurreratua eskaintzeko. Une honetan, 24 doktorego-tesi abian dira taldearen barruan, eta 25 lagun ari dira gure masterrak egiten. Teknologien arloan aditu gehiago prestatzearren, lau programa eskaintzen ditugu graduatuentzat:

- Itzulpengintza eta Teknologia graduondokoa, euskaraz.[15] Itzulpengintzako lan-merkatuaren erronka berriei erantzuteko gai izango diren itzultzaile profesionalak prestatzen ditugu. EHUko sail bik antolatzen dute: Ingeles eta Alemaniar Filologia eta Itzulpengintza eta Interpretazio Saila, eta Lengoaia eta Sistema Informatikoak Saila. Langune klusterrak eta UEUk laguntzen dute antolaketan. Jarduera gehienak online egiten dira.
- Hizkuntzaren Azterketa eta Prozesamendua masterra, euskaraz eta ingelesez.[16] 2001ean hasi ginen ematen, eta, ordudanik, 130 ikasle baino gehiago izan ditugu. Ikasle horiek dira, hain zuzen ere, gaur egun Hizkuntza Teknologia lan-eremua dinamizatzen ari diren informatikari eta hizkuntzalariak. Masterrean, hizkuntzaren prozesamenduan ezinbestekoak diren bi arloak uztartzen dira: teknologia, batetik, eta hizkuntzalaritza, bestetik. Hori dela eta, arlo askotako ikasleak biltzen dira.

15. http://www.ueu.eus/ikasi/gradu-edo-graduondoko-ikastaroa/494/Itzulpengintza+eta+teknologia
16. http://ixa.si.ehu.es/master

- Language and Communication Technology Erasmus Mundus Masterra, ingelesez.[17] Nazioarteko kontsortzio zabal batean integratuta, ekarpen berezia egiten dugu ospe handiko nazioarteko master horretan: kontsortzioan orokorrean ematen den ingelesaren ikuspuntu globalaz gain, baliabide gutxiko hizkuntza baten ikuspegia ematen dugu. Ixa Taldeak 29 urteko esperientzia du horretan, eta nazioarteko erreferentziatzat jotzen da.
- Hizkuntzaren Azterketa eta Prozesamendua doktorego-programan,[18] erronka teknologiko berriei erantzuteko gai diren doktoreak prestatzen ditugu. Azken 5 urteetan, 20 tesi aurkeztu dira, eta, 2017an, beste 24 ikasle doktoregai daude.

Bukatzeko

Artikulu honetan, Ixa Taldearen azken urteotako ibilbidea azaldu nahi izan dugu. Izan ere, azaldu nahi izan dugu Datu Handien garaian murgilduta gauden honetan nola egokitu diren hizkuntza-teknologiak datu pila hori modu eraginkor batean erabiltzeko eta ustiatzeko, eta zer bide berri hartu duten.

Hizkuntza handiek markatzen dituzte egokitzapen eta berrikuntza horien eredu eta joerak. Joera horiei jarraituz, baina, hainbat aplikazio interesgarri sortu dira euskaraz ere azken urteotan.

Zenbait kasutan, ordea, teknika berriak aplikatzea ezinezkoa suertatzen da, testu-masa handiak eta, oro har, baliabide ugari eskatzen dituztelako. Horixe gertatzen da maiz euskararen kasuan, bai eta baliabide urriko beste hizkuntzetan ere. Horren aurrean, zer egin? Zein da irtenbidea? Artikuluaren bigarren zatian dago hausnarketaren erantzuna. Izan ere, euskarak hartu dituen bide *independenteak*-edo azaldu ditugu, bai kultura zabaltzeko zereginei begira, baita baliabide berriak sortzeko ahaleginari begira ere. Ixa Taldearen inguruan eginiko hausnarketa hau baliagarria izan daiteke baliabide urriko hizkuntzentzat, jarrera-aldaketa interesgarria dakarren neurrian.

Hizkuntza guztientzat da estrategikoa alor honetako teknologian indarra jartzea, txikiak nahiz handiak izan. Hizkuntzaren prozesamenduan inbertitzea eta hizkuntzaren garapena eskutik doaz, lotuta doaz neurri handi batean. Horregatik da hain garrantzizkoa hizkuntzen teknologian formazio egokia duten ikertzaileak egotea, etorkizun hurbilean planteatuko zaizkigun erronkei modu eraginkorrean aurre egiteko.

17. http://lct-master.org/

18. http://www.ehu.eus/eu/web/doctoradohaplap/aurkezpena

Resumen

La cantidad de información textual disponible de forma electrónica está creciendo sustancialmente, por lo que las nuevas tecnologías se han convertido en instrumentos necesarios para aprovechar al máximo dicha información. En este artículo describimos tres nuevas técnicas que abren nuevos horizontes, como son la computación en la nube, el aprendizaje profundo y las redes neuronales. Estos grandes recursos han traído consigo la posibilidad de crear nuevas aplicaciones en las lenguas mayoritarias del mundo. ¿Pero son estos recursos igualmente prácticos en el caso del euskera? Teniendo en cuenta que la cantidad de textos disponibles en lengua vasca es inmensamente menor que los que encontramos en inglés, es necesario experimentar con estas nuevas técnicas para valorar hasta qué punto nos resultan prácticas. Por otro lado, ¿ha de actuarse siguiendo el mismo procedimiento en el caso de las lenguas con pocos recursos? Además de conocer las tendencias que se observan en las lenguas mayoritarias, el objetivo es conocer qué recursos, instrumentos y aplicaciones resultan más productivos. Existen en el mundo 190 lenguas que, si bien cuentan con una mínima presencia en Internet, no han desarrollado aún tecnologías del lenguaje, y para las que una estrategia diferente puede ser la clave del éxito. El Grupo IXA ha dedicado algunos proyectos y tesis doctorales al estudio de este ámbito. Así mismo, se han planteado e incorporado nuevas propuestas de tratamiento de la cultura vasca dentro del proyecto de DSS2016.

Résumé

La quantité de textes informatifs dont nous disposons au niveau mondial croît vertigineusement, et nous avons besoin de nouvelles technologies pour pouvoir exploiter au maximum toute cette information. Dans cet article nous décrivons trois nouvelles techniques qui ouvrent de nouveaux horizons : l'informatique en nuage, l'apprentissage en profondeur et les réseaux neuronaux. Ces techniques ont permis de créer de nouvelles applications dans les langues hégémoniques du monde. Mais, ces ressources sont-elles valables pour des langues comme l'euskara ? Compte tenu du fait que la quantité de textes disponibles en notre langue est infiniment moindre que celle des textes rédigés en anglais, il nous faut expérimenter ces nouvelles techniques pour mesurer jusqu'où elles sont valables dans notre cas. Par ailleurs, les langues ayant peu de ressources sont-elles tenues de suivre forcément les pas des langues majoritaires ? En plus de connaître les tendances que l'on observe dans les langues majoritaires, nous devons également savoir quels sont les ressources, les applications et les instruments les plus productifs pour nous. Il existe dans le monde 190 langues qui, bien qu'elles aient une présence minime sur Internet, n'ont pas encore développé des technologies du langage. Une stratégie différente pourrait donc leur servir de référence. Au sein du Groupe IXA, nous avons consacré un certain nombre de projets et de thèses doctorales à l'étude de cette question, de même que nous avons présenté de nouvelles propositions sur le traitement de la culture basque, dans le cadre de DSS2016 (Donostia/San Sebastián capitale européenne de la culture 2016).

Abstract

The amount of textual information available in electronic form has grown dramatically in recent years. Accordingly, new technologies are required in order for us to take full advantage of this information. In this paper, we describe three techniques that have opened new pathways: cloud computing, deep learning, and neural networks. For the majority languages of the world, big textual data have led to the creation of a new generation of applications. But are these resources equally useful for Basque? Although the amount of electronic text in Basque is about three orders of magnitude lower than that in English, we must experiment with these new techniques to analyze their usefulness. On the other hand, is it really necessary to behave in the same way in the case of minority languages? We would like to determine not only the tendencies that stem from majority languages, but also the most beneficial and productive resources, tools and applications. There are 190 languages that have a minimum presence on the Internet, but do not yet use these language technologies. We argue that, for them, a strategy of going beyond the path of the "big" languages may be a key factor for success. The IXA Group has produced several projects and PhD theses specifically on the development of language technologies for Basque. We have also introduced new ideas to promote the Basque culture in the "Donostia/San Sebastián 2016 European Capital of Culture" (DSS2016) event. We explain these initiatives in this paper.