ADIMEN ARTIFIZIALA ETA EUSKARAREN ERABILERA: AUKERAK ETA ERAGIN-EREMU BERRIAK

IGOR LETURIA AZKARATE

ELHUYARREKO HIZKETA-TEKNOLOGIEN ARDURADUNA

JOSU AZTIRIA URTARAN

ELHUYARREKO ADIMEN ARTIFIZIALA HIZKUNTZAN UNITATEKO KOORDINATZAILEA

TESTUINGURUA ERABAT ALDATU DA

Euskarazko hizkuntza-teknologien ibilbidea ez zen atzo goizean hasi. 2001ean hasi ginen hainbat erakunde ikertzen eta baliabideak sortzen euskararentzat, eta orain egindako ibilbide oparo horren ondorioz euskarazko hizkuntza-teknologiak beste hainbat eremu aurreratuetara eramateko gai gara.

Euskararentzat ikertzetik, euskaratik ikertzera pasatu gara, alegia; hau da, euskararentzat oinarrizko baliabideak eta tresnak sortzetik, euskaratik sektore ekonomiko aurreratuenetara (Fabrikazio aurreratua eta biozientziak, kasu) eramateko gai gara . Hizking 21, Anhitz, Berbatek, Be2Tek, ElkarOla, BerbaOla, Modela, Quales, Modena, Tando, DLNP4 eta DeepText bezalako ikerketa-proiektu estrategikoen bidez, Eusko Jaurlaritzak bideratutako Etortek eta Elkartek deialdien bidez finantzatu direnak. Bide honetan guztian elkarlanean aritu gara Vicomtech, Tecnalia, EHU/UPVko Ixa eta Aholab ikerketa-taldeak eta Elhuyar Fundazioa, eta nabarmendu eta azpimarratu behar den elkarlana dela uste dugu.

Azken urteotan, esan bezala, euskarazko hizkuntza-teknologietan ibilbide oparoa egin da, eta beste hizkuntzetatik (ingelesetik eta gaztelaniatik, kasu) oraindik urrun bagaude ere, beste hainbat hizkuntza gutxituk eta estatu-hizkuntza askok ez duten oinarri teknologikoa dugula esan daiteke beldurrik gabe.

Euskararen garapena murritzagoa da gaztelaniarekin eta ingelesarekin alderatuz, batez ere, baliabide gutxiago dituelako. Hizkuntza-baliabide gutxiago, estatu-babesa murritzagoa delako, hiztun gutxiago dituelako, finantza-baliabideak eskasagoak direlako eta merkatu-aukerak ere murritzagoak; hau da, eskala-ekonomiak martxan jartzeko zailtasun handiagoak dituelako.

Oro har, esan daiteke estatuaren babesik ez duten gainerako hizkuntza gehienak antzeko egoeran daudela, hau da, hiztun gehien dituzten estatuko hizkuntzen azpitik. Gaztelania eta frantsesa, aldi berean, ingelesa baino garapen-egoera txikiagoan daude hizkuntzaren eta hizketaren teknologiei

dagokienez. Berdin esan daiteke, hala nola, alemana edo portugesa bezalako hizkuntzetan. Hori, berriz ere, hiztunen kopuruarekin dago lotuta, baliabideekin eta merkatu-aukerekin.

Hala ere, ezin da txartzat jo hiztunen kopurua eta estatukoa ez den izaera kontuan hartzen baditugu. Esan dezakegu hiztun gehiago dituzten estatuko hizkuntza askoren ia maila berean dagoela, eta horietako batzuk baino hobeto ere badagoela. Hori da, hein handi batean, arestian aipatutako eragileek azken hamarkadetan egin duten ahaleginari esker, eta ikerketa-proiektuak finantzatzeko ateak ere ireki dituztelako erakunde publikoek.

Azken urteetan adimen artifizialak hizkuntzan izaten ari den garapen disruptiboak eraldaketa handiak ekarri eta ekarriko ditu gainera. Berrikuntza teknologiko sakon eta eraldatzaile hauen norabidea ongi zehaztea da euskal gizarteak duen erronka handienetako bat, izan ere, aukera handiak ekar ditzake gizarte-berrikuntzaren alorrean, eta zehazki, euskararen garapen-prozesuan. Esaterako, adimen artifizialak eta hizkuntza-teknologiek orain arte pentsaezinak ziren aukerak ekarri dituzte enpresa eta erakundeen hizkuntzen kudeaketan nahiz euskararen erabilera areagotzeko estrategietan. Hizkuntza-politika ez da berdina izango hemendik aurrera (ez luke berdina izan behar) eta euskararen garapena neurtzeko adierazleetan ere aldaketak eta egokitzapenak egin beharko dira.

Adibide txiki bat: ate joka ditugun bozgorailu adimendunak gure etxe eta lantokietan nagusitzen badira, zein izango da etxeko edo laneko hizkuntza? Orain arte bezala ulertu behar al dugu zer den etxeko hizkuntza? Nola neurtuko dugu hori? Aski ikertua dago, gainera, ingurune digitalean nagusitzen diren hizkuntza-praktikek eragin zuzena dutela kaleko erabilera eta aurrez aurreko komunikazio-hizkuntzan.

Covid-19aren krisiak, gainera indarrez zetozen hainbat joera indartu eta azeleratu ditu, hala nola digitalizazioa eta adimen artifizialaren aurrerapen teknologikoen bultzada. Testuingurua erabat aldatzen ari da, baita euskara eta euskal hiztunontzat ere.

TEKNOLOGIAREN EBOLUZIO IKARAGARRIA

Euskaraz bezala gainontzeko hizkuntzetan ere, hizkuntza- eta hizketa-teknologiak eboluzionatzen joan dira urteetan zehar, neurri handi batean, teknologiaren eta gailuen ahalmenaren unean uneko egoerak baldintzatuta.

Ordenagailuen sorreratik berriki arte, hizkuntza teknologiek erregeletan oinarritutako metodoak baliatu dituzte. Hizkuntza ulertu, tratatu, itzuli zein sortzeko, informatikari eta hizkuntzalariz osatutako lantaldeek lengoaia informatikoetan eta ordenagailuetako datu-egituretan idazten zituzten hizkuntza naturalaren arauak, maila guztietan: lexikoa, morfologia, sintaxia, semantika, pragmatika… Baina hizkuntza baten makinaria oso konplexua izan ohi da, arauak bezainbeste dira salbuespenak eta maila ugaritako anbiguotasunak. Hori dela eta, lortzen ziren emaitzak ez ziren oso onak izaten, ez bada kasu gutxi batzuetan (hizkuntza oso antzekoen arteko itzulpen automatikoa, adibidez).

Duela 20-30 urte metodo estatistikoak hasi ziren gailentzen. Hizkuntza tratatzeko *machine learning* edo ikasketa automatikoa deritzon teknika baliatzen dute metodo estatistikoek: testu baten analisi sintaktikoa egiteko, itzulpen bat egiteko edo hizketa sortzeko ikasteko adibide mordoa hartzen dira (testuak eta euren analisiak, edo dagozkien itzulpenak, edo horietatik sortutako audioak) eta ikasketa automatikoa egiten duen programa bati ematen zaizkio; horrek bere barneko kalkulu estatistikoen pisuak egokitzen ditu, eta sarrera berriak ematen zaizkionean irteera berriak kalkulatzeko gai da. Sistema horiek aurrekoek baino emaitza hobeak lortzen zituzten kasu gehienetan.

Azken aldaketa azken bost urteetan gertatu da, modu generikoan *deep neural networks* edo sare neuronal sakonak deitzen diren tekniken eskutik. Sare neuronalak, izatez, ikasketa automatikoaren azpimultzo bat dira, antolaketei eta datu-egiturei dagokienez giza burmuineko neurona-sareak imitatzen dituztenak. Sare neuronalak duela 30 urte asmatu ziren eta orduan interes handia

erakarri eta itxaropen handia sortu zuten, baina espektatibak ez ziren bete eta bazterrean gelditu ziren. Azkenaldi honetan, ordea, sare neuronalak erabiltzen hasi dira berriz arlo askotan, sistemen kalitatean jauzi oso handia ekarri baitute. Zergatik gertatu da hau, zergatik lehen ez zuten funtzionatu eta orain bai? Hiru arrazoi nagusi aipa genitzake: lehenengoa, gaur egun askoz informazio digital gehiago dagoela sistema horiek entrenatzeko (testuak, audioak, itzulpenak…); bigarrena, makina askoz ahaltsuagoak daudela gaur egun, sareak datu kantitate handiekin entrenatzea ahalbidetzen dutenak; eta, azkenik, sare neuronalen antolaketa edo paradigma berri eta konplexuagoak asmatu direla, ataza gehiagotan arrakasta lortu dutenak (sare antolaketa konplexuago horiek aurrekoetatik ezberdintzeko erabiltzen da “sakona” adjektiboa).

Neurona-sare sakonei esker lortu den kalitate-jauzia hain da handia, gaur egun hizkuntza- eta hizketa-teknologien alor guztietan erabiltzen hasi baitira, hizkuntza guztietan bezala euskaran ere. Teknologia horiek apur bat ezagutzen edo baliatzen dituen edonork argi ikusi du nola hobetu diren aurrez zeuden sistemak eta nola agertu diren aplikazio berriak. Euskarari dagokionez, ikusi dugu nola itzulpen automatikoko sistemen belaunaldi berriak emaitza txundigarriak lortzen dituzten (hemen aipa daitezke Elhuyarren Itzultzailea.eus, sei hizkuntzen artean itzultzeko gai dena eta dokumentu osoak ere itzultzen dituena, edo Eusko Jaurlaritzaren Itzultzaile neuronala, euskara eta gaztelania artean itzultzen duena), edo nola hizketaren ezagutzako sistemak agertu diren (Elhuyarren Aditu.eus), edo hizketa sintesiko sistemek naturaltasun hobea lortzen duten. Baina, esan bezala, sare neuronalak hizkuntza-teknologien arlo guztietan aplikatzen ari dira (zuzentzaile automatikoak, iritzien erauzketa, hizkuntzaren sorkuntza, laburpen automatikoa, elkarrizketa-sistemak…), eta honek ekarriko du luze gabe tresna berri harrigarriak ikustea, euskaraz ere bai.

ADIMEN ARTIFIZIALAREN GARAPENAK EUSKARAREN ERABILERAN IZAN DITZAKEEN ALDAKETAK ETA AZTERKETARAKO GAKOAK

Etorkizunari begira, adimen artifizialak ireki ditzakeen bideak ikusirik, hainbat agertoki aurreikusten saiatuko gara. Alde batetik, uste dugu Internet euskararen erabileraren behatoki gisa indartuz joango dela, hau da, euskara nola eta zenbat erabiltzen den xeheago jakiteko baliabide erraldoia bihurtuko da; beraz, Internet corpus gisa ustiatzeko tresnak findu eta hobetu beharko ditugu. Bestetik, adimen artifizialak sare neuronal eta ikasketa automatikoaren bidez aukera teknologiko berriak ekarriko ditu, eta, orain arte bezala, prest egon beharko dugu euskararentzat ekar ditzakeen aukerak baliatzeko. Aukera batzuk zehazten ere saiatu gara

- Hizketaren ezagutzaren bidez ikus-entzunezko edukiak transkribatzea eta ahozko erabilera aztertzea.
- Euskararen erabileraren neurketa fintzeko testu-masa handien tratamendu automatikoa eta adimenduna.
- Web-corpusen sailkatze automatikoa.
- Itzulpen automatiko neuronalarekin euskara hedadura gutxiago edo urrunago dauden beste hizkuntza batzuekin lotzea.
- Sare neuronalak eta ikasketa automatikoa baliatzea zuzentzaileetan txertatutako estilo- eta gramatika-akatsak zuzentzeko.
- Euskararen erabilera inklusiboa bultzatzea hizkuntza-baliabideak eta -tresnak egokituz.

Eta azkenik, hizkuntza-teknologiek datu errealetan oinarrituta ikasten dutenez, pentsatzekoa da erabiltzaileen nahiz teknologien bidez hedatzen den euskara estandarragoa egiten joango dela, hau da, tresnak elikatzeko erabiltzen ditugun datuak erabileran oinarrituko dira, eta erabileran ere eragingo dute era berean. Hori guztia neurtzea oso garrantzitsua izango da.

AUKERA EGINGARRIAK ETA ERAGIN-EREMU BERRIAK DITUGU BEGI-BISTAN

EUSKARA SORTZE-HIZKUNTZA BIHURTZEA

Sare neuronaletan oinarritutako itzultzaile automatikoen emaitza bikainek aukera berriak sortzen dituzte eta euskarazko lan-zirkuituak areagotzeko bideak errazten ditu. Hala nola, euskarazko dokumentuen ekoizpena handitzeko, euskarazko testuen ulermena areagotzeko eta ofimatika-tresnetan integratzeko aukera teknikoak zabaltzen direnez, interbentzio-eremu interesgarria da erakunde publikoetan eta enpresa pribatuetan normalizazio-planak garatzen ari diren profesionalentzat.

Era berean, transkripzio-teknologien emaitzak ere asko hobetu dira (hizketatik testura bihurtzen duten teknologiak) eta erakunde publikoetako batzarrak eta bileren edukiak irisgarriagoa egiteko eta edukiak azkarrago sortzeko bideak irekitzen ditu. Gainera, transkripzio-teknologien eta itzulpen automatikoaren konbinazioarekin bilerak eta batzarrak euskaraz egin arren, gaztelaniaz zuzenean edo diferituan emateko erraztasunak handiak dira. Herritar ororen hizkuntza-eskubideak errespetatuaz euskarazko jarduna lehenesteko aukera ematen du horrek.

IRISGARRITASUNA HOBETZEA

Hizkuntza- eta hizketa-teknologien aplikazioak ez dira soilik eleaniztasunarekin erlazionatutakoak. Erabilera argi bat irisgarritasunean dute teknologia hauek.

Hizketaren sintesia edo sorkuntza erabil daiteke pertsona itsuen edo ikusmen arazoak dituztenen lagungarri. Adibide garbi bat da dislexia edo bestelako ikusmen edo irakurketa arazoak dituzten pertsonak gogoan Eusko Jaurlaritzako Hezkuntza Saileko Berritzegune Nagusiarentzat Elhuyarrek garatutako Irakurle Digitala, edozein webgune, dokumentu edo PDF ozen irakurtzen duena. Edo Elhuyarrek garatutako Bidaide soluzioa, telefonorako audiogidak modu errazean sortzea ahalbidetzen duen sistema; ohiko audiogidez harago, GPS edo Bluetooth bezalako teknologiak baliatuz pertsona itsuak edo bestelako ezgaitasunak dituztenak ere gida ditzake museo edo ibilbide batean zehar.

Hizketaren ezagutzak, aldiz, bestelako ezgaitasun eta arazoetan lagun dezake. Bideoetan pertsona gorrentzat hain beharrezkoak diren azpitituluak sortzea errazten dute. Eta zuzeneko emanaldiak edo irratiko programak aldi berean azpititulatuta ere eskain daitezke, kalitate erabatekoarekin ez bada ere, gutxienez esaten dena ulertu eta jarraitzeko moduan. Ezgaitasun motorrak dituzten pertsonei ere lagun diezaieke hizketaren ezagutzak, idatzi behar dutena ahoz diktatuta makinak transkribatzen duen bitartean. Hauek guztiak egin ditzake Elhuyarren Aditu.eus plataformak, zeina, gainera, tresnetan integratu daitekeen.

ELEANIZTASUNA IKUS-ENTZUNEZKOETAN ETA KOMUNIKAZIOETAN HEDATZEA

Aipatu diren teknologietako zenbaiten konbinazioz, aplikazio aurreratuagoak lor daitezke. Zehazki, hizketaren ezagutza, itzulpen automatikoa eta hizketaren sintesiaren konbinazioak ekarpen handia egin diezaieke ahozko komunikazio eta ikus-entzunezko eleaniztunei.

Izan ere, hizketaren ezagutzak hizketa testu bihurtzen badu, itzulpen automatikoak testua hizkuntza batetik bestera pasatu badezake eta hizketaren sintesiak beste hizkuntza batean dagoen testu hori berriz ere audio bihurtzen badu, finean, hizkuntza batean esaten den hizketa bat beste hizkuntza bateko hizketa bihur daiteke. Hau da, bikoizketa prozesua tresnen bidez automatikoki egina.

Ikus-entzunezkoen kasuan, teknologion egoera kontuan izanik momentuz zaila izan daiteke film eta antzekoetan aplikatzea, hainbat arrazoirengatik: elkarrizketa kolokialak, argota, bolumen aldaketak, atzealdeko zarata… duten audioak transkribatzeko zailtasuna, elkarrizketa informalak itzultzeko arazoak, adierazkortasunaren galera hizketaren

sintesian… Baina albistegi eta dokumentaletan, adibidez, gaur egun jada erabil daiteke. Prozesuaren lehen bi urratsek eskuzko zuzenketa eska dezakete, baina bikoizketa prozesua asko arin dezakete. Eta bikoizketa osoa egin gabe ere, lehen bi urratsen bidez azpititulu itzuliak izango genituzke. Elhuyarren Aditu.eus plataformak lehen bi urratsak (azpititulazio eta itzulpen automatikoa) integratzen ditu jada. Aipatutako teknologien kateaketaren bidez, beste hizkuntzetako produkzioak euskaraz entzuteko edo azpititulatuta ikusteko aukera dago, eta baita euskarazko produkzioa nazioartekotzeko.

Ahozko komunikazio eleaniztunetan ere prozesu bera baliatu daiteke: azpititulu itzuliak izateko hitzaldi, mahai-inguru, parlamentuko saio, udalbatza eta antzekoetan, edo besteek esandakoa norberaren hizkuntzan itzultzeko eta alderantziz nazioarteko bideo-dei batean. Kasu hauetan prozesu hori guztia zuzenean egin behar denez eta ezin direnez urrats bakoitzaren emaitzak zuzendu, baliagarriak izan daitezkeen emaitzak lortzea zaila da momentuz, baina etorkizunean horrelako gauzak egitea ere posible izan daiteke.

Aipatutako aplikazioetan aurrerapenak egiteko, parte hartzen duten teknologia horietako bakoitza hobetzen eta funtzionalitate berriez hornitzen joan beharra dago. Euskalkiak eta elkarrizketa informalak ere ongi transkribatu eta itzultzea, hizketaren sintesiak jatorrizko ahotsak eta adierazkortasuna imitatzea… Horietan guztietan ikerketa lerro irekiak daude, eta gutxinaka aurrerapenak ikusten joango gara.

EUSKARAREN ERABILERA ERREALA NEURTU ETA MONITORIZATZEA

Testuen meatzaritza eta Big Data tekniken bidez testu-masa handiak bildu, sailkatu eta ezagutza erauzteko bideak ireki dira, eta horrek aukera handiak eskaintzen ditu erakunde bateko hizkuntza-jarduna modu sistematiko eta monitorizatuan neurtzeko. Hau da, laginketetan oinarritu beharrean, datu errealetan eta objektiboetan neurtzeko aukera ematen du.

Laburbilduz, testuen meatzaritzak, itzulpen automatikoak eta hizketaren teknologiek euskarazko jarduna areagotzeko, euskararen normalizazioan sakontzeko eta interbentzio-proiektu berritzaileak garatzeko aukera paregabeak eskaintzen dizkigu. Helburuak ongi zehaztu, prozesuak mimoz diseinatu eta zorroztasunez ebaluatuz gero, praktika berriak martxan jartzeko abagune paregabea dugu. Aukera egingarriak ditugu eta interbentzio--eremu berriak aurrez aurre.

Teknologia bidelagun, saiatze horretan ez gara gelditukо, ezagutzaz gure ingurua eraldatzea baita gure asmoa: euskara ingurune digitalean osasuntsu bizitzea, alegia.