02
IKERKETA LERROAK
>>>

ITZULPEN AUTOMATIKOA: AUKERAK, ARAZOAK ETA ERRONKAK

Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Kepa Sarasola
IXA taldea (EHU)
Helbide elektronikoa: i.alegria@ehu.es

SARRERA

Itzulpen automatikoa informatikaren sorreratik datorkigun erronka da. Hasierako konputagailuetarako bilatu ziren lehenengo erabilpenen artean bi erronka zail zeuden: xakea eta itzulpen automatikoa. Mende erdi geroago, lehenengoan adituen maila berdintzera edo gainditzera iritsi den bitartean, bigarrenean asko falta da oraindik hori lortzeko. Egun itzulpen-sistema automatikoek lortzen duten kalitatea oso urrun dago pertsona itzultzaileek lortzen duten kalitatetik. Hala ere, lan errepikakorrenetan teknologiak asko laguntzen die itzultzaileei eta itzulpen-enpresetan derrigorrezko bihurtu dira laguntza-sistemak. Bestalde, antzekotasun handiko hizkuntzen artean itzulpenak egiteko edo bestela itzuliko ez liratekeen dokumentuen itzulpen-zirriborroak edukitzeko, gero eta gehiago erabiltzen da itzulpen automatikoa.

Informatikaren hasiera haietatik pasa diren berrogeita hamar urte baino gehiagotan denetarik egon da komunitate zientifikoaren barruan eta hainbat *euforia-depresio-zalantza* ziklo bereizi izan dira. Ikuspuntu komertzialetik *Systran* izan da enpresen artean gailendu dena, sistema sendoak eta garestiak eskainiz 35 bat hizkuntza-bikotetarako. Horrekin

batera gaur egun web bidezko itzulpen-zerbitzu automatikoak eskaintzen dira doan hainbat hizkuntzatarako, baina gehienen kalitatea ez da ona.

Azken urteetan gure inguruan *euforia* gehiago sumatzen da *depresio* edo *zalantza* baino. Ikerketa-proiektu asko bideratzen ari dira Europan, eta hainbat gobernuk eta erakundek mota honetako programak edo zerbitzuak erosten edo kontratatzen ari dira. Horren adibide dira Kataluniako Generalitat, Galiziako Xunta eta Cervantes Institutua.

Euskararen inguruan ere bada gai honen inguruko berririk. Bi ekimen azpimarratu nahi ditugu:

- *Opentrad* proiektuan[1] garatu den espainieratik euskarara itzultzeko *Matxin* izeneko prototipoa[2].
- Eusko Jaurlaritzak lehiaketa publikora atera duen kontratua[3], espainieratik euskarara itzultzeko tresna bat erosteko.

Artikulu honen helburua hor kokatzen da: itzulpen automatikoaren kontzeptuak, aukerak, arazoak eta erronkak deskribatzea eta eztabaidatzea. Edozein kasutan, hemen azaltzen diren kontzeptuak hainbat iturri interesgarrirekin osa daitezke ([1] [2] [3] [4]), horietatik hartuta baitaude artikulu honetako hainbat ideia.

Azpimarratu behar da teknologia hauen inguruan gero eta diru gehiago mugitzen ari dela, eta horren adibide dira honako bi egitasmoak:

- *Quebec-eko gobernuak 4 milioi dolar inbertitu ditu Quebec-eko Unibertsitateak Outaouais-en duen campusean hizkuntza-teknologia garatuko duen zentro bat sortzeko. (...) Quebec-eko garapen ekonomikorako ministro Raymond Bachand-ek esan du helburua Outaouais eskualdea 2020rako mundumailan itzulpen-teknologian liderra izatea dela.* (Andoni Sagarnaren blogetik[4] hartua).
- Irlandan, 16,8 miloi euro bideratu du Science Foundation Ireland (SFI) erakundeak gai honen inguruan N*ext generation of high tech automatic language translation* (Itzulpen automatikorako teknologiaren belaunaldi berria) egitasmoan[5].

Hasierako konputagailuetarako bilatu ziren lehenengo erabilpenen artean bi erronka zail zeuden: xakea eta itzulpen automatikoa. Mende erdi geroago, lehenengoan adituen maila berdintzera edo gainditzera iritsi den bitartean, bigarrenean asko falta da oraindik hori lortzeko.

OINARRIZKO KONTZEPTUAK

Lehenik eta behin funtsezkoa da bereiztea **konputagailuz lagundutako itzulpena** eta **itzulpen automatikoa**. Lehenean pertsona da prozesuaren gidaria, itzultze-prozesuan hainbat tresna lagungarri dituen arren; itzulpen automatikoan, berriz, makina da itzultze-prozesuaren ardatza nahiz eta giza laguntza egon daitekeen itzuli aurretik testua prestatzeko (aurreedizioa) edo itzuli ondoren txukuntzeko (postedizioa). Lehen arloan aurrerapen handiak egin dira azken urteetan eta egun itzulpen-memoriak erabat hedatuta daude itzultzaile profesionalen artean [5]. Artikulu

Lehenik eta behin funtsezkoa da bereiztea konputagailuz lagundutako itzulpena eta itzulpen automatikoa. Lehenean pertsona da prozesuaren gidaria, itzultze-prozesuan hainbat tresna lagungarri dituen arren; itzulpen automatikoan, berriz, makina da itzultze-prozesuaren ardatza nahiz eta giza laguntza egon daitekeen itzuli aurretik testua prestatzeko (aurreedizioa) edo itzuli ondoren txukuntzeko (postedizioa).

honetan, hala ere, itzulpen automatikoa dugu hizpide, gure ustez arlo soziolinguistikoan duen eragina askoz ere handiagoa izan daitekeelakoan.

Zailtasuna. Itzulpen automatikoaz gaur egun espero daitezkeen emaitzak bi faktorek baldintzatzen dituzte: batetik, hizkuntzen arteko antzekotasuna, eta bestetik, hizkuntza-bikoterako eskuragarri dauden itzulpenen bolumena.

Antzekotasun handiko hizkuntzen artean askoz errazagoa da itzultzea (eskuz zein automatikoki) eta hitzez hitzeko itzulpena eginez emaitza onargarriak lor daitezke. Oso desberdin diren hizkuntzen artean itzultzea, berriz, askoz konplexuagoa da. Horretan berebiziko garrantzia duten bi faktoreak hitz-hurrenkera eta morfologia dira. Hurrenkera libreko hizkuntzek eta flexio aberatseko hizkuntzek zailtasun gehigarria dakarte hurrenkera finko eta flexio sinplea duen hizkuntza batekin itzulpenak egin nahi direnean.

Bestalde, aldez aurretik egindako itzulpen eskuragarrien bolumen handiak asko laguntzen du itzulpen automatikorako hainbat teknika erabiltzen direnean (ikus teknologiak atala), teknika horiek egindako itzulpenen informazioan oinarritzen diren neurrian. Beraz, berebiziko garrantzia dute itzulpen automatikoko egitasmoak garatzeko orduan itzulpen-memoria edo corpus paralelo deritzen baliabideak. Ildo horretan ulertu behar da Europako Batzordeak bideratutako ekimena[6]: Europako hizkuntza ofizialen corpus paraleloak askatzea.

Bistan da bi faktore horiek oso kontuan hartzekoak direla euskaratik edota euskara automatikoki itzultzeko sistemak eraikitzean edo diseinatzean, bi eragozpen handi baitaude: batetik, eremu urriko hizkuntza izateak dakarren corpus paraleloen eskasia, eta bestetik, inguruko hizkuntzekiko duen antzekotasun falta.

Ebaluazioa. Itzulpen automatikoaren kalitatea neurtzeko kontuan hartu behar da funtsezko ezaugarri hau: esaldi bat ondo itzultzeko aukera zuzen bat baino gehiago dago. Beraz, nola jakin daiteke modu automatikoan egindako itzulpen bat ze puntutaraino zuzena den ala ez, ez badugu balizko itzulpen guztien zerrenda?

Gehienetan itzulpen zuzen bakar bat (itzulpen-memorietatik hartuta adibidez) edukitzen dugu eskura emaitzarekin automatikoki konparatzeko, baina hurbilpen hau pobrea da, emaitzak ez baitira fidagarriak, batez ere hurrenkera libreko hizkuntzetan. Lan handiagoa hartuz gero, bigarren itzulpen zuzen bat sor daiteke eskuz, eta automatikoki lortu dena bi itzulpenekin konparatu. Hauxe da sistemak konparatzeko jarraitu ohi den metodoa baina ez da oso fidagarria.

Ebaluazio fidagarriagoa lor daiteke automatikoki lortutako itzulpena itzultzaile bati emanez ahalik eta aldaketa gutxien eginda zuzen dezan [3]. Aldatutako hitz kopuruari dagokion portzentajea oso neurri fidagarria da, baina eragozpen bat du: ez da automatikoa. Automatikotasun eza hau dela eta, sistemaren doitasuna neurtu nahi dugun bakoitzean lana

errepikatu beharko da. Portzentaje horri edizio-distantzia deritzo eta %10 baino txikiagoa izan behar du itzultzaile automatikoak itzulpen-enpresa baten errendimendua igo dezan.

Helburu hori edozein testu motatarako lortzea oso zaila da, eta, egun, antzekotasun handiko hizkuntzen artean edo baliabide eta diru asko inbertituz lortzen da.

ERABILERAK

Aurrekoa irakurrita pentsa liteke diru-xahuketa dela gaur egun euskaratik edo euskara automatikoki itzultzen duen sistemaren bat eraikitzea. Horri erantzun baino lehen, bada kontuan hartzeko beste faktore bat: erabilera. Aurreko pasartean kalitateko itzulpen profesionalaz aritu gara: ordaintzen eta zabaltzen diren itzulpenak. Baina horrelakoak al dira egin daitezkeen itzulpen guztiak?

Itzulpen automatikoaren merkatua aztertzen denean, bi itzulpen mota bereizten dira: batetik, aipatu dugun itzulpen profesionala, **zabalkundeko itzulpena** (*dissemination* ingelesez) deritzona; eta bestetik, **asimilazioa** deritzon eta norberarentzat den itzulpena. Bigarren hori *itzulpen automatikorik gabe egingo ez litzatekeen itzulpen* gisa ikus liteke.

Ulermenari bideratuta dagoen bigarren erabilera horren interesa asko handitu da Interneten eta globalizazioaren eraginez. Hona horren adibideak:

- Gai zehatz baten inguruko artikulu bat bilatu nahi dugu sarean gero ondo itzultzeko, baina ez dakigu zein den interesatzen zaiguna. Artikulua aukeratzeko asimilazio-itzulpena oso lagungarria izan daiteke.
- *Chat* moduko aplikazioetan ondo ezagutzen ez dugun hizkuntza batean hitz egin ohi den lagun bati mezuak bidaltzeko edo bere mezuak jasotzerakoan gurera itzultzeko.
- Enpresa batek delegazioa ireki du atzerrian. Dokumentazio formalaren itzulpena ohiko enpresak erabiliz moldatuko du, baina behar berriak sortzen dira: teknikarien eta langileen arteko komunikazioa, talde moduan sendotzeko ekintzak (langileen intraneta, kasu), lan sindikala, eta abar. Aktibitate hauetan ohiko itzulpena ez da bideragarria, eta batzuetan komunikazio-beharra ingelesaren bitartez ebatz badaiteke ere, behar asko ase gabe geratuko dira, eta horietako batzuetarako asimilazio-itzulpena irtenbide izan daiteke.
- Ikastola batetik guraso erdaldunei abisuak SMS formatuan bidaltzeko erabilgarri izan daiteke. Era berean, guraso horiek ikastolara bidaltzen dituzten mezuak itzultzeko.

Ondorioz, honako hau azpimarratu behar da: teknologia honen hel-

Itzulpen automatikoaren merkatua aztertzen denean, bi itzulpen mota bereizten dira: batetik, aipatu dugun itzulpen profesionala, zabalkundeko itzulpena (dissemination ingelesez) deritzona; eta bestetik, asimilazioa deritzon eta norberarentzat den itzulpena.

Itzulpen automatikoa egiteko programak bi multzo handitan banatzen dira: erregeletan oinarritutako sistemak (RBMT) eta aurreko itzulpenetan oinarritutako sistemak (analogiaz lan egiten dutela esan ohi da).

burua ez da itzultzaileei lana kentzea, baizik eta itzulpen-bolumena handitzea. Sistema hauen kalitatea ez da erabatekoa izango, baina prezioa edo presa direla-eta, itzulpen profesionala erabiltzen ez duenak erabiltzen ahal dituzte. Aurreko ezaugarria oso interesgarria da baliabide eta hiztun gutxi dituzten hizkuntzetarako.

Sistema hauen gakoetako bat abiadura da. Internet erabiltzeko bada denbora errealean lan egin behar du eta nabigazio itzulia (Interneten nabigatu ahala itzulpenak jasoz jatorrizko formatuan) bideratu. Beraz, helburu hori duten sistemetan oinarrizko ezaugarria abiadura izango da.

Domeinua. Sistemaren kalitatea (doitasuna edo *precision* deritzona) hobetzeko bada beste estrategia bat: eremua edo domeinua murriztea. Itzultzailea edozein testutarako prestatu beharrean testu mota zehatz bat itzul dezan prestatuko da. Helburua izango da domeinu horretako itzulpenen kalitatea hobetzea. Domeinua murriztuz, arazoak ere murriztu egingo dira eta abantailak agertuko dira: lexikoa sinpleagoa da, sintaxiaren aukerak finkoagoak dira, semantikoki anbiguotasun gutxiago dago, aurretik itzulitako testuetan antzeko zatiak aurkitzeko probabilitatea asko handitzen da, eta abar. *Meteo* sistema izan da arlo honetan ezagutu den sistema arrakastatsuena, eta ildo hau jorratzeko erreferentzia nagusia. Sistema horrek ingelesaren eta frantsesaren artean eguraldi-iragarpenak itzultzen ditu Kanadan.

Domeinua erabat murriztuz eta domeinu horretarako aurretik egindako itzulpen asko eskuratuz, posible izan daiteke antzekotasun txikia duten hizkuntzen arteko zabalkundeko itzulpena lortzea.

TEKNOLOGIAK

Itzulpen automatikoa egiteko programak bi multzo handitan banatzen dira: erregeletan oinarritutako sistemak (RBMT) eta aurreko itzulpenetan oinarritutako sistemak (analogiaz lan egiten dutela esan ohi da). Azken horiek, era berean, bi azpimultzotan banatzen dira: adibideetan oinarritutako sistemak (EBMT) eta sistema estatistikoak (SMT). Ohiko sistema komertzialak RBMT teknologian oinarritzen dira, baina azken urteetako ikerketen joera SMT sistemak garatzea da. Hala ere, azkenaldian antzematen den irtenbidea hibridazioa da, hau da, teknologia horiek konbinatzea.

RBMT. Sistema hauetan itzulpen-prozesua hizkuntzalariek prestatutako hiztegien eta erregelen bidez kudeatzen da. Hiru mota bereiz daitezke: hitzez hitz itzultzen duten sistemak (zuzeneko itzulpena ere esaten zaio), transferentzia bidezkoak eta *Interlingua* bidezkoak.

Transferentzia bidezkoetan hiru fase bereizten dira: jatorri-hizkuntzan dagoen testuaren analisia, transferentzia edo hizkuntza batetik besterako moldaketa, eta xede-hizkuntzako testuaren sorkuntza. Hiru faseak automatikoki egiten dira, baina arazo handiak daude emaitzak zehatzak izan daitezen, batez ere hizkuntzen arteko antzekotasuna txikia

denean. Hainbat arazo larri daude, baina agian larrienak aurreko atalean aipatutako bi hauek dira: analisi sintaktiko sakona eta hautapen lexikala.

Horrezaz gain, transferentzia hizkuntza parearen mende dagoenez, hizkuntza askoren artean itzulpenak egiteko (Europako Batasunean beharko litzatekeen sistema adibidez) transferentzia-moduluen beharra biderkatu egiten da. Hau da, *n* hizkuntza badugu helburu eta bikote guztien arteko itzulpenak behar baditugu, analisirako zein sorkuntzarako *n* modulu izango dira nahiko, baina transferentziarako *n(n-1)* modulu beharko dira.

Azken eragozpen hori saihesteko oso interesgarriak dira *Interlingua*n oinarritutako sistemak. *Interlingua* erakargarria da ikuspuntu teorikotik ere, hizkuntzatik independentea den adierazpide unibertsala bilatzea oso erronka zaila bezain interesgarria baita hizkuntzalarientzat. *Interlingua* bidezko sistemetan ez dago transferentziarik, eta beraz, analisiak oso sakona izan behar du, hizkuntzatik independentea den adierazpidera pasa behar baita jatorri-esalditik erauzten den informazio guztia. Oso analisi automatiko sakona behar denez erroreak (morfologia, sintaxia, semantika, pragmatika, ...) ere ugari izango dira. Sorkuntza-fasea ere konplexua da, eta anbiguotasunak ebaztea arazo larria izan daiteke. Zoritxarrez hurbilpen honen ildotik orain arte sortu diren tresnak ez dira doitasun handikoak izan, eta ikergai gisa duen interesa handia bada ere, sistema komertzialak eraikitzeko orduan baztertuta dago egun.

EBMT eta SMT. Sistema hauek aurretik egindako itzulpenetan oinarritzen dira, beraz, itzulpen-bildumaren tamaina da berauen kalitatean eragina duen funtsezko faktoreetako bat. Ezaugarri hori partekatzen duten arren, itzulpen-memoriak ustiatzeko garaian desberdin jokatzen dute. EBMTn itzuli behar diren abiapuntu-testuan unitate linguistikoak (sintagmak, esaldiak, patroiak, eta abar) identifikatu behar dira eta horietan oinarrituta bilatzen dira itzulpenak. Beraz, eredu linguistikoa dela esan daiteke.

SMTn oinarria estatistikan dago: itzulpenenetan errepikatzen (abiapuntu-hizkuntzan eta xede-hizkuntzan) diren hitz multzoak (hitz solteak zein multzoak, linguistikoki unitateak direnak edo ez) dira ustiatzen diren elementuak, eta gero elementu horiek ondo konbinatzeko xede-hizkuntzaren hizkuntza-eredu bat erabiltzen da. Ideia sinplea da: hitz bat nola itzuli behar den jakiteko, aurretik itzulitako esaldi guztien artean hitza dutenak aukeratzen ditugu, gero esaldi horien itzulpena lortu eta bilatu zein den esaldi itzuli guzti horietan azaltzen den hitza, gehien azaltzen den hitz hori izango da jatorrizko hitzaren itzulpena. Gero gure esaldiko hitz guztien ordaina zein den jakinda, bigarren etapa batean xede-hizkuntzako hitz horien ordena egokia bilatu behar da, horretarako, hasieran, xede hizkuntzako testu erraldoietan estatistikoki bilatzen da ea hitz horien artean zeintzuk azaldu diren elkarren segidan eta zein maiztasunarekin, eta gero, datu horiek kontuan hartuta hitz ordain horien

Euskararen aldetik hainbat saio egin dira, nagusia OpenTrad proiektuaren barruan. Eusko Jaurlaritzak bultzatutako egitasmoak ere bultzatuko du arlo hau. Edozein kasutan argi eduki behar da egungo baliabideekin ia ezinezkoa dela euskararako zabalkundeko itzulpen automatikoa lortzea epe laburrean.

Itzulpen automatikoaren emaitzak mugatzen dituzten arazoak aski ezagunak dira ikertzaileen artean. Orokorrean esan daiteke arazo nagusia anbiguotasuna (elementu bera modu desberdinetan ulertzeko/analizatzeko aukera) dela, baina ez da arazo bakarra.

konbinazio probableena bilatzen da. Ondorioz bi eredu eraikitzen dira tresna eraikitzean: itzulpen-eredua, non itzulpen-memorietan erlazionatuta agertzen diren zatiak metatzen diren; eta hizkuntza-eredua, non xede-hizkuntzaren ezaugarri sintaktikoak modelatzen diren. Lehen eredua eraikitzeko itzulpen-memorien bildumak behar dira, hizkuntza-eredua eratzeko xede-hizkuntzaren testu-bildumak nahikoak dira, baina analizatuak eta ahalik eta handienak. Arlo honetan laguntzeko software libre garrantzitsua garatu da: GIZA++, Moses, eta abar. Software hau hizkuntza-bikote berrietarako sistemak egiteko oso lagungarria da.

Kontuan hartu behar da emaitza txukunak lortzeko oso bolumen handiak behar direla. *EuroParl* corpusak, Europako Batzordearen itzulpen-bildumak, 30 bat milioi hitz du hizkuntza bakoitzeko. Hortik aurrerako bolumenak ematen du sistema sendo bat eraikitzeko garantia. Euskaraz aritzen garenok, partaide askoren laguntzarekin ere, nekez eskura dezakegu 10 bat milioi hitzeko corpusa itzulpen-eredurako. Gainera, euskara flexio handiko hizkuntza eta hurrenkera askekoa denez, zailtasunak handitu egiten dira hitz itzulien agerkidetzen maiztasuna jaitsi egiten delako. Horren ondorioz, antzeko kalitateko emaitzak lortzeko testu-bilduma handiagoak erabili beharko dira.

Hibridazioa. Aipatutako teknologia bakoitzak aldeko eta aurkako ezaugarriak ditu, RBMTk hizkuntzalarien eta ingeniarien lan handia eskatzen du, eta nekez lortzen da hobekuntza muga batetik aurrera. SMT oso teknologia erakargarria da hasiera batean, itzulpenen bildumak edukiz gero, lan gutxirekin hasierako sistema txukuna azkar egin daitekeelako hizkuntza-bikote batzuetarako. Baina aurretik aipatutako mugekin topatuz gero (itzulpenen bilduma mugatua eta hizkuntzen ezaugarri desberdinak eta hurrenkera librea) lantegi zaila da.

Horren aurrean abian daude metodo bakoitzaren onena hartzea eta eragozpenak saihestea lortu nahi duten ikerketak. Konbinazio hauetan oinarrituta sortzen diren sistemei *hibrido* esaten zaie.

Alde batetik SMT sistemen bilakaera direnak aipa ditzakegu. Sistema hauetan SMT sistemetan itzulpen-bildumak analizatzen dira hizkuntza bakoitzerako dauden tresnekin (sintagmak, esaldiak, etab. identifikatu nahian, eta beraien informazioa itzulpen-ereduan sartu nahian).

Beste aldetik RBMT sistemen aldaerak ditugu. Hauetan itzulpen-bildumetatik ikasitako estatistikak gehitzen dira lexikoan eta erregeletan, ondorioz itzulpenak eta erregelak aplikatzeko probabilitateak hartzen dira kontuan.

Ikerkuntza-gai hauek dira gaur egun puri-purian daudenak, baina emaitzak oraingoz ez dira oso ikusgarriak.

Aplikazioak. Esan bezala merkatuko sistema gehienak RBMT motakoak dira, itzulpena zuzena egiten dutenak antzekotasun handiko hizkuntzen artean edo transferentzia bidezkoak beste hizkuntzen artean. Gaur egungo sistemetan hibridrazio minimo bat erabiltzen da. *Systran*[7] da

ezagunena. Sistema komertziala eta garestia da. 15 bat hizkuntza kontuan hartzen ditu (baina ez konbinazio guztiak). Oso hedatuta dago enpresa handitan eta administrazioan, eta duela gutxi arte *Google*k ere erabiltzen zuen. Aipatutako webgunean egin daitezke probak.

Hala ere, azken urteetan SMTn oinarritutako sistemak izan dira gehien aurreratu dutenak eta komunitate zientifikorako NIST erakundearen bitartez antolatuta sasi-txapelketetan emaitza onenak lortu dituztenak (emaitzak oso eztabaidagarriak izan diren arren)[8]. *Google*k apustu argia egin du sistema hauen alde, eta irabazle izan da lehiaketa horietan. Bere guneetan ere hasi da mota honetako sistemak eskaintzen[9]. Edozein kasutan, kontuan hartu behar da hedadura handiko hizkuntzetarako (itzulpen-bolumen handiko hizkuntza-bikoteetarako zehatz-mehatz esanda) bakarrik lortzen direla emaitza onak, konpetitibo izateko itzulpenen oso bolumen handia behar da (lehiaketa horietan NBEren itzulpenak erabiltzen dira gehienbat).

Euskararen aldetik hainbat saio egin dira, nagusia *OpenTrad* proiektuaren barruan. Eusko Jaurlaritzak bultzatutako egitasmoak ere bultzatuko du arlo hau. Edozein kasutan argi eduki behar da egungo baliabideekin ia ezinezkoa dela euskararako zabalkundeko itzulpen automatikoa lortzea epe laburrean. Baina domeinu jakinetarako eta asimilaziorako oso gauza interesgarriak egin daitezkeelakoan gaude, beti ere modu koordinatuan eta plangintza baten arabera lan egiten bada.

IXA taldean[10], Eleka eta Elhuyar fundazioarekin batera, helburu horiekin lan egiten ari gara eta egungo gure helburu nagusiak honako hauek dira:

- *Opentrad* proiektuan garatutako espainiera-euskara RBMT motako itzultzailea, *Matxin* izenekoa, hobetzea asimilaziorako tresna eraginkorra izan dadin, eta ingelesa-euskara bikoterako hedatzea. Era berean domeinu murriztu baterako moldaketa egiten ari gara, emaitzak ebaluatu ahal izateko.
- Itzulpen-memorien bilduma handitzea sistema estatistiko egoki bat sortu ahal izateko.
- Hibridazioaren bidez aurretik aipatutako bi sistemak hobetu nahi ditugu, egunen batean zabalkuntzako kalitatea lortu ahal izateko.

Bigarren helburua funtsezkoa da, eta ezin du inork bere kabuz ondo bideratu. Beraz, arlo honetan datozen urteetan urrats arrakastatsuak eman ahal izateko gakoa izango da elkarlana eta erakundeen inplikazioa.

Matxin itzultzaile automatikoa Opentrad proiektuari esker garatutako sistema da [3]. Proiektu honen helburua estatu espainiarreko hizkuntza nagusietarako kode irekiko itzulpen automatikoko sistemak sortzea izan da.

ARAZOAK ETA ADIBIDEAK

Itzulpen automatikoaren emaitzak mugatzen dituzten arazoak aski ezagunak dira ikertzaileen artean. Orokorrean esan daiteke arazo nagusia

Itzulpen automatikoa oso eginkizun konplexua da, ingeniaritza-proiektu erraldoia, hizkuntzalariekin eta itzultzaileekin lankidetza estua eskatzen duena. Kalitateko itzulpen automatikoa lortuko bada, ezinbestekoak dira inplikatutako hizkuntzetarako oinarrizko eta kalitatezko tresnak. Beraz, derrigorrezko baldintza da egitasmo hauek testuinguru zabalago batean kokatzea.

anbiguotasuna (elementu bera modu desberdinetan ulertzeko/analizatzeko aukera) dela, baina ez da arazo bakarra. Hona arazo nagusienen zerrenda:

• Anbiguotasun lexikala.
Hizkuntza batetik bestera itzultzerakoan, hitz bat beste hitz batekin (edo batzuekin) ordezkatu behar dugu. Hori egiterakoan, gerta daiteke jatorri-hizkuntzako forma batek, *composición* kasu, hainbat ordain izatea helburu-hizkuntzan: *osaketa*, *osaera*, *idazlan*, *konposizio*, *hitz-elkarketa*. Itzultzaileak horietako bat aukeratuko du, eta beti ez da zuzena izango.
Adibidez, *Realizarán una audición de todas las composiciones* itzuli behar badugu, badakigu itzulpen zuzena dela *Konposizio guztien entzunaldia egingo dute*, baina baliteke sistema automatiko batek beste era honetan itzultzea: *Osaketa guztien entzute bat egingo dute.*
Kasu bertsua gertatzen da preposizio bat itzuli nahi izanez gero. Demagun *por* itzuli nahi dugula euskarara. *Elhuyar* hiztegian dugun lehenengo itzulpena *–(en)gatik* da, baina hori ez da beti *por* horren itzulpen zuzena izango, beste hainbat itzulpen izan baititzake. Hona adibideak:

he ido por verle	*ikusteko joan naiz*
ha sido firmado por el alcalde	*alkateak sinatu du*
está por hacer	*egiteko dago*
ha hecho 100 kilómetros por hora	*100 kilometro orduko egin ditu*
lo ha hecho por la tarde	*arratsaldean egin du*
me lo dijo por teléfono	*telefonoz esan zidan*
lo he hecho por ti	*zuregatik egin dut*
vinieron por otro camino	*beste bide batetik etorri ziren*
vete tú por mí	*joan zaitez nire ordez*
estoy por la paz	*bakearen alde nago*

Itzulpen bat edo beste aukeratzeko hainbat estrategia erabili beharko dira. Batzuetan inguruan dituzten elementuei begiratuko diegu: aditz infinitbo batekin agertzen bada, *aditz izena* gehi *–ko* erabiliko ditugu itzulpena egiteko (*ikusteko joan naiz*, *egiteko dago*). Beste kasu batzuetan esaldiaren formari begiratuko diogu; hala egiten dugu pasiboaren kasuan: *por* daraman sintagma subjektu izango da eta euskaraz *ergatiboa* (*-k*) erabiliko dugu itzultzeko (*alkateak sinatu du)*. Beste kasu batzuetan ondoko hitzaren ezaugarriei begiratuko diegu: neurria adierazten badu, *-ko* lekuzko genitiboaz itzuliko dugu (*100 kilometro orduko egin ditu*); denbora adierazten badu, *-n* inesiboaz (*arratsaldean egin du*); komunikazio-tresna agertzen bada, *-z* instrumentalaz (*telefonoz esan zidan*). Eta, beste batzuetan, berriz, aditzaren azpikategorizazioari erreparatu behar diogu.
• Egiturazko anbiguotasuna edo anbiguotasun sintaktikoa.
Analisi-zuhaitzetan oinarritzen den sistema batek, jatorri-hizkuntzako analizatzailearekiko mendekotasun osoa du. Analisiak anbiguoak

dira, bai mendekotasunei begira, eta baita analisi morfologikoari begira ere. Analisi okerra hautatuz gero, itzulpenean eragin txarra izango du.

Mendekotasunen barruan, garrantzitsua da jakitea ze elementuk modifikatzen duen zer.

Adibidez, *las farmacias dan fichas con consejos sobre enfermedades* segidaren analisia egiterakoan, analizatzaileak analisi-erroreak izan ditzake eta analisi hau eman dezake:

```
dan —  |
       |— las farmacias
       |— fichas
       |— con consejos
       |— sobre enfermedades
```

Analisi honen gainean eraikitako itzulpena hau litzateke: *Farmaziek fitxak ematen dituzte aholkuekin gaixotasunen gainean*.

Analisi zuzena lortuz gero,

```
dan —  |
       |— las farmacias
       |— fichas
            |— con consejos
                 |— sobre enfermedades
```

itzulpen zuzena eraikitzeko moduan geundeke: *Farmaziek gaixotasunen gaineko aholkuak dituzten fitxak ematen dituzte*.

Beste kasuetan pertsonei ere zail izango zaigu jakitea zein den analisi zuzena. *Ha venido el amigo de Bilbao* kasuak bi itzulpen zuzen izan ditzake: *Bilboko laguna etorri da* eta *laguna Bilbotik etorri da*. Analisi-zuhaitzak erabakiko du itzulpena bat den edo bestea den, eta tamalez analisi zuhaitza sortzen duen programak ez du esaldiaren testuingurua ezagutzen edo ulertzen.

Egun ditugun analizatzaileek duten arazo handienetakoa da koordinazioaren analisia, eta analisi horrek baldintzatuko du (aurreko kasuetan bezala) itzulpena.

He venido con Juan y Miren segidaren ondoko analisia jasoz gero,

```
he venido —  |
             |— con Juan
             |— y —     |
                        |— Miren
```

honoko itzulpena emango dugu: *Jonekin etorri naiz eta Miren*.

Itzulpen zuzena lortu ahal izateko (*Jonekin eta Mirenekin etorri naiz*),

hau da behar dugun analisia:

he venido — |
|— con
|— y — |
|— Juan
|— Miren

Mendekotasunen analisi zuzena izateaz gain, oso garrantzitsua da analizatzaileak analisi morfologiko zuzena ematea. Morfologiak ere izugarrizko eragina du itzulpenean. Ez da kontu bera aurreko perpauseko *Miren* horren analisia izen berezia izatea, edo *mirar* aditzaren forma jokatua izatea. Izan ere, analizatzaileak *mirar* aditzaren forma dela markatuko balu, hau litzateke sistemak emango lukeen itzulpena: *Jonekin etorri naiz eta begira bezate*.

• Anbiguotasun semantikoa:

Alderdi semantikoan hainbat alor txerta daitezke, baina arazoa ematen dutenetako bat hitz mota edo hitzen ezaugarriak ondo ez zehaztetik dator. Oso garrantzitsua izan daiteke jakitea ze hitz motaren aurrean gauden. Adibidez, *euskara* hizkuntza dela, edo *lagun* biziduna dela, edo *tren* ibilgailua dela, eta abar.

Hauen ezaugarriak jakiteak eragina du postposizioen aukeraketan. Adibidez, demagun *ha hablado en euskara* itzuli behar dugula. Nola hautatuko dugu *en* preposizioaren itzulpena? *en* preposizioak hainbat itzulpen izan ditzake: *-n, -engan, -z*. Itzulpen orokorrena *-n* dela kontuan izango badugu ere (*está en casa* / *etxean dago*), itzultzaile automatikoari zehazten ahalko zaio biziduna baldin bada ondoan duen hitza *–engan* hautatu beharko duela (*confío en mi amiga* / *nire lagunarengan dut esperantza*), eta hizkuntza baldin bada *–z* beharko duela (*ha hablado en euskara* / *euskaraz hitz egin du*). Hala ere, bereizketa honek ez ditu arazo guztiak konpontzen eta batzuetan inguruko beste elementuei ere begiratu behar zaie: ibilgailu batekin baldin badoa, *-z* erabiliko dugu ondoan doan hitza modifikatu gabe baldin badago (*ha venido en tren* / *trenez etorri da*), baina modifikatzaileren bat baldin badu *–n* erabili beharko dugu (*ha venido en el tren de las once* / *hamaiketako trenean etorri da*).

• Anbiguotasuna pragmatikan.

Hizkuntza guztiek ez dute mundua era berean antolatzen, eta ez dituzte kontzeptuak era berean lexikalizatzen. Espainieraz lexikalizazio bakarra duen elementu batek euskaraz hainbat lexikalizazio izan ditzake. Adibidez, *mi hermana está enferma* bi modutara itzul daiteke euskaraz *ahizpa gaixo dago* edo *arreba gaixo dago*. Esaldiaren testuingurua ezagutu behar da itzulpen egokia zein den erabakitzeko, eta gaur egunean horrelakorik ez da erabiltzen.

Alderantziz, euskaratik espainierara *hura* moduko izenorde bat itzuli nahiko bagenu, arazo bera genuke: *hura eraman zuen* / *lo llevó* / *la llevó*.

Arazo hauek guztiek eragin zuzena dute itzulpen automatikoaren emaitza kaxkarretan, RBMT motako sistemetan *batez* ere. Egindako itzulpenetan oinarritutako sistemetan arazo horietako batzuk saihets daitezke, unitate luzeak harrapatzeko duten gaitasunari esker, baina inguruko informazioa nahikoa ez denean arazoak ez dira ebazten.

MATXIN ITZULTZAILE AUTOMATIKOAREN ADIBIDEAK

Matxin itzultzaile automatikoa *Opentrad* proiektuari esker garatutako sistema da [3]. Proiektu honen helburua estatu espainiarreko hizkuntza nagusietarako kode irekiko itzulpen automatikoko sistemak sortzea izan da. Sistema horien itzulpenak egin ahal izateko abiadura handiko eta kode irekiko bi motor garatu ziren: *Apertium* antzeko hizkuntza-bikoteen arteko itzulpena egiteko (espainiera-katalana, katalana-espainiera, espainiera-galegoa, galegoa-espainiera), eta *Matxin* antzekoak ez diren hizkuntza-bikoteen arteko itzulpena egiteko (espainiera-euskara).

Matxin erregeletan oinarritutako itzultzaile automatikoa da (RBMT).

Egun *Matxin*ek ematen dituen hainbat itzulpen nahiko onak direla esan daiteke, eta hona horietako adibide zenbait:

Le llevé el pan a mi hermano a casa	*Ogia eraman nion nire anaiari etxera*
Viene en coche y vive en esta ciudad	*Automobilaz dator eta hiri honetan bizi da*
Los políticos dicen que demos tiempo al tiempo	*Politikariek esaten dute pazientzia izan dezagula*
Los aviones volaron sobre la muchedumbre	*Hegazkinek jendetzaren gainetik hegan egin zuten*
El libro está sobre la mesa	*Liburua mahaiaren gainean dago*

Bestetan, berriz, nahiz eta itzulpenak erabat zuzenak ez izan, ulergarriak direla esaten ahal da:

Cuatro nuevas sucursales de Correos se abrirán en la capital	*Correos-en 4 sukurtsal berri kapitalean irekiko dira*
El hospital tendrá 48 nuevas habitaciones individuales en 2009	*Ospitaleak 48 banako gela berri izango du 2009tan*

Ondoko lerroetan adibide horiek dituzten itzulpen-arazoak komentatuko ditugu. Azken bi perpaus horien analisia zuzena da. Transferentziari dagokionez, bi perpausetan zenbakiak agertzen dira, eta horiek itzultzeko garaian, oraingoz zenbakia bera jartzen dugu kasu guztietan. Sorkuntzari dagokionez, lehenengo perpausa traketsa dela dirudi ordenaren aldetik; horrela atera da, oro har, espainieraz aditzaren atzetik dagoen elementua aditzaren aurrera pasatzen dugulako. Bigarren perpausean, berriz, data bat agertzen da, baina ez da detektatu data dela: 2009. Zenbaki bat denez, sintagma horri deklinabide mugagabea esleitu zaio transferentzian eta sorkuntza hala egiten du itzultzaile automatikoak. *48 nuevas habitaciones* ere mugagabea baliatuz sortzen du, eta horregatik aditz laguntzaileari ez zaio pluralaren informazioa pasatzen eta *du* laguntzailea jartzen du. Hala eta guztiz ere, uste dugu aurreko itzulpenak ulergarriak direla.

Baina badira euskarazko bertsioa irakurrita ulertzen ez diren itzulpenak ere, noski. Halakoetan espainierazko jatorrizko esaldia irakurri behar da itzulpenek zer esan nahi duten ulertu nahi izanez gero:

Fue entonces cuando escuchó la explosión que se produjo en el primer piso	*Orduan izan zen l eherketa entzun zuenean eragin zen 1 pisuan*
Mientras en la Unión Europea la edad media de independizarse son 22 años, en España supera los 26.	*Europar Batasunean Erdi Aroa banandu bere burua izatera 22 urtetan izan, Espainian 26 gainditzen du.*

Lehenengo perpausak arazoak ditu hiru faseetan.

Analisia ez da zuzena. Analizatzaileak ematen duen analisian *que se produjo* katea *fue* aditzaren mendeko gisa agertzen da, eta era berean *en el primer piso* katea ere *fue* aditzaren mendeko gisa markatzen du. Transferentzian *gertatu*ri esleitutako itzulpena ez da zuzena: *eragin.* Sorkuntzan *primer* ordinalari ez dio ordinalen forma jartzen.

Bigarren perpausak ere arazoak ditu hiru faseetan. Esaldi honetan komaz berezitako bi perpaus ditugu. Bigarrenean hainbat elipsi daude eta horrek analisian eragin handia du. Horrez gain, koma agertzen da eta hainbat esperimentu egin ondoren, koma bat agertzen den bakoitzean bi esalditan banatzea erabaki genuen. Beraz, hemen bi perpausen analisia dugu: komaren aurreko perpausa eta ondorengoa.

Analisi automatikoaren arabera, komaren aurrekoaren burua *mientras* lotura-elementua da. Bere mendeko zuzenak dira lau kate hauek: *en la Unión Europea, la edad media, de independizarse* eta *son*. Eta *son* aditzaren azpian dago *22 años*. Analisi trakets honek ondoko urrats guztiak baldintzatzen ditu. Horretaz gain, *edad media* hitz anitzeko elementu gisa ezagutu du eta horren itzulpen gisa *Erdi Aro* ematen du. Aurrekoaz gain, *independizar* itzultzeko *banandu* erabili du eta *se* hori *bere buru* gisa eman du. *de* preposizioa itzultzeko *partizipioa + izatera* itzulpena hautatu du ondoren aditza duelako[11]. *Mientras* transferentzian ondo itzultzen du (*-n bitartean*), baina sorkuntza egiterakoan sistemak kale egin du eta *izan* aditza soilik jarri du, *diren bitartean* sortu ordez.

Transferentzian morfologia ere transferitzen da, eta *22 años* horri inesiboa esleitzen zaio. Zenbakiak modifikatzen duen elementua denborazkoa baldin bada (kasu honetan *urte*) eta *izan* aditzaren mendekoa baldin bada, horiei inesiboa esleitzen zaie *era el 4 de julio* modukoak itzultzeko. Adibideko kasuan ere inesiboa esleitzen zaio, baina ez da zuzena.

Komaren ondokoari dagokionez, analisia zuzena da. Transferentzian *los 26* horri mugagabea esleitzen zaio (lehen esan dugun bezala zenbaki bat agertzen delako), eta horregatik laguntzailea singularrean agertzen da (*du*).

ERRONKAK

Azalpen teknikoak kontuan hartuta, euskaldunok eta euskal herritarrok gure buruari galde diezaiokegu ea zeintzuk diren gaur egun gure erronkak arlo honetan. Gure ikuspuntutik ondoko puntu hauek azpimarratu nahi ditugu:

1. Itzulpen automatikoa oso eginkizun konplexua da, ingeniaritza-proiektu erraldoia, hizkuntzalariekin eta itzultzaileekin lankidetza estua eskatzen duena. Kalitateko itzulpen automatikoa lortuko bada, ezinbestekoak dira inplikatutako hizkuntzetarako oinarrizko eta kalitatezko tresnak. Beraz, derrigorrezko baldintza da egitasmo hauek testuinguru zabalago batean kokatzea. Testuinguru horretan euskararako analizatzaile/sortzaile zehatzagoak eta baliabide lexikal, morfologiko zein semantiko aberatsagoak bultzatu beharko dira.
2. Ildo beretik arestian aipatu den itzulpen-memorien bilduma lehentasun handiko helburua da, baliabide hori giltza baita etorkizuneko tresnen kalitateari begira.
3. Itzulpen automatikoaren arloan garatzen diren ikerketa-proiektuetan helburu ausartak baina, aldi berean, errealistak ezarri behar dira, eta egungo erabilerez gain, etorkizun hurbilean gerta daitezkeenak aurreikusi. Ildo horretan, asimilaziorako sistemek etorkizun handia dutelakoan gaude.
4. Teknologia berri hauek euskararen garapenerako eta hedapenerako sekulako garrantzia eduki dezakete, batez ere lehen aipatu den esaldi bat buruan izanik: *bestela egingo ez liratekeen itzulpenak* egitea oso mesedegarria izan daiteke euskaraz bizi nahi dugunontzat.
5. Euskal Herrirako erronka ekonomikoa ere bada. Bertako hizkuntza erabiltzeaz eta bultzatzeaz gain, ingurune eleaniztun batean bizi gara, eta espainiera, ingelesa eta frantsesarekin elkarbizitzen ohituta gaude. Horrek aukera handiak ematen dizkigu teknologi berri hauek garatzeko garaian. Beharra eta esperientzia badugu, ausardiarekin konbinatuta emaitza ederra eman dezakeela uste dugu. Irlandan eta Kanadan ausartzen badira, gu zergatik ez.▮

BIBLIOGRAFIA

[1] Wikipedia. Itzulpen automatikoa. http://eu.wikipedia.org/wiki/Itzulpen_automatikoa [Online; 2008ko martxoaren 18an atzitua].
[2] Abaitua J. 2002. Itzulpengintza Automatikoa: hamar orduko sarrera, http://paginaspersonales.deusto.es/ abaitua/konzeptu/ta/mt10h_eu [Online; 2008ko martxoaren 18an atzitua].
[3] Mayor, A. 2007. MATXIN: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz. Doktorego-tesia. Euskal Herriko Unibertsitateko; Donostiako Informatika Fakultatea. http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak
[4] Hutchins, J. MT-Archive. http://www.mt-archive.info/ [Online; 2008ko martxoaren 18an atzitua].

[5] Bernaola I., Morales A. eta Payros I. 2003. Ordenagailuz lagundutako itzulpena eta itzulpenaren kalitatea. Senez 26. http://www.eizie.org/Argitalpenak/Senez/20031210/Bermopa

OHARRAK

1. www.zientzia.net/artikulua_inprimatu.asp?Artik_kod=11907
2. www.opentrad.org
3. http://sustatu.com/1202227533
4. www.bloglines.com/blog/andonisagarna?id=93
5. www.computing.dcu.ie/news/newsitems/sem1_0708/josef/index.html
6. http://sustatu.com/1202722245
7. www.systran.co.uk/
8. www.nist.gov/speech/tests/mt/2008/doc/
9. www.google.com/translate_t
10. http://ixa.si.ehu.es
11. Halako itzulpena zuzena da honelako adibidean: de saberlo antes, no hubiera ido / jakin izatera, ez nintzen joango.