Itzulpen automatikoaren kalitatea ebaluatzeko metrika automatiko neuronalak euskararen zain

Nora Aranberri (HiTZ, UPV/EHU)

Itzulpen automatikoaren garapen eraginkor bat bermatzeko, ezinbestekoa da bide bat izatea, ia kosturik gabe eta denbora-tarte txiki-txikian, sistema bati egindako moldaketek itzulpen-kalitatea hobetzen ala okertzen duten adieraziko diguna. Zenbat eta azkarrago erdietsi aldaketen eragina, orduan eta saiakera bideratu gehiago egin ahal izango dira sistema hobetzeko. Hain justu, horixe da ebaluazio-metrika automatikoen jomuga nagusia. Hortik haratago, bai ikerketan, bai enpresetan, ohikoa da itzultzaile automatikoko (IA) sistemen nolabaiteko kalitate orokorra ezagutzeko ere erabiltzea. Zer esanik ez, lehenengo inpresio arin horren ostean etorri beharko lirateke erabilera-testuinguruan ardaztutako profesionalen ebaluazioak.

Hori horrela, euskararako teknologiak ere aurrera egin dezan, ezinbestekoa da aztertzea ea metrika horiek gure hizkuntza barne hartzen duten eta, hala izatekotan, zer-nolako fidagarritasunez erantzuten duten. Izan ere, baliabide askoko hizkuntzentzat erabilgarri daudenez, ez litzateke harritzekoa batek baino gehiagok hizkuntza txikientzat ere aplikatzea, di-da, bere horretan, ekintza horren eraginaren gaineko inongo hausnarketarik gabe. Zalantzazko erabilera horietatik lortutako emaitzetan oinarrituta hartzen diren erabakietan legoke benetako arriskua.

Metriken azterketan eta garapenean lehen urratsak egiteko, orain dela hilabete batzuk, euskal komunitatearen laguntza eskatu genuen; tartean, baita itzultzaile profesionalena ere, besteak beste ItzuL posta-zerrendaren bidez. Orduan agindu genuen ekimen hartatik ikasitakoak jakinaraziko genituela. Hitzemana zor, artikulu honetan laburbildu ditugu ebaluazio-metriken nondik norakoak, COMET ebaluazio-metrika neuronalak egun euskararekin duen harremana, eta abian jarritako ebaluazio-ekimenaren gaineko zeinbait hausnarketa ildo, interesekoak izango direlakoan.

Ebaluazio-metrika automatikoak

Ebaluazio-metrika automatikoetan, oinarrizkoa da itzulpen automatikoaren kalitatea neurtzeko gai izatea, betiere pertsonen iritziarekin bat etorrita. Hurbilpen desberdineko ebaluazio-metrikak proposatu eta erabili izan dira helburu horretarako.

Zalantzarik gabe, metrikarik hedatuena BLEU deritzona da (Papineni et al. 2002). 2001. urtean proposatu zuten IBM enpresako T.J. Watson ikerketa zentroko kideek, eta berebiziko arrakasta lortu zuen. Gogoratu 1990eko hamarkadan indarra hartzen ari zirela corpusetan oinarritutako IA sistemak, erregeletan oinarritutako sistemen kalterako. Corpusetan oinarritutako lehen sistema estatistiko haiek prest izatea, hau da, entrenatzea, askoz azkarragoa zen. Horrek ekarri zuen hainbat parametro eta estrategia hilabete gutxian probatzeko aukera eta, aldi berean, itzulpenen kalitatea ebaluatzeko behar presazkoagoa, saiakeren ekarpena aztertzea ezinbestekoa baitzen ikerketa bideratzeko. BLEUk ahalbidetu zuen, lehen aldiz, minutu gutxi batzuetan esaldi-zerrenda luze-luzeen kalitatea neurtzea pertsona-ebaluatzaileen beharrik gabe.

Azal-azaletik, BLEUk egiten duena zera da: erreferentziazko itzulpen baten eta itzulpen automatiko baten arteko berdintasun maila kalkulatu, amankomunean dituzten hitz kopurua eta hitz-sekuentzien luzera kontuan hartuta. Zenbaketak esaldi mailan egiten dira. Hipotesia da gainjarritako hitz kopuruak adieraziko digula sorburuko testuko informazioa zer neurritan transmititu den helburuko testura, eta gainjarritako hitz-sekuentzien luzerek emango digutela helburuko hizkuntzaren adierazpen-egokitasunaren berri.

Kritika ugari eta zorrotzak egin zaizkio BLEUri urteetan zehar. Hasteko, sorburuko testu baten baliokideak askotarikoak izan daitezke, eta begi bistakoa da itzulpen automatikoa erreferentzia bakarraren kontra alderatzeak ez duela hori kontuan hartzen. Litekeena da, horrela jokatuz gero, itzulpen egoki asko eta asko txartzat jotzea, adibidez. Soluzioak erraza dirudi: erabil dezagun erreferentzia kopuru handiagoa. Horrek, ordea, itzulpen (profesional) gehiagoren beharra dakar; bestela esanda, diru-inbertsio handiagoa. Erreferentzia bakarrarekin lan egiteko, nahikoa da jada itzulita dauden testuak baliatzea; bi edo hiru bertsio gehigarri biltzeko, aldiz, beharrezkoa da itzultzaile profesionalak kontratatzea, eta hori *garestia* da. Azken batean, ebaluazio-testuak milaka esaldikoak izatea komeni da, eta, gainera, ez da jarraibide zuzena behin eta berriz esaldi-multzo beraren gainean jardutea, baterako edo besterako joerak ekiditeko. Funtsean, bai ikerketan, bai industrian, erreferentzia bakarraren erabilera gailentzen da.

Beste ahulezia bat aipatzearren, jo dezagun euskararen morfologiara. Gure hizkuntza eranskaria da. BLEUk hitz mailan egiten baditu kontaketak, penalizazio bera esleituko die, esaterako, lema eta posposizio okerrez eratutako hitz bati eta lema zuzena baina postposizioak oker dituen hitzari. Horrek zera dakar: alde batetik, euskararentzat metrika gogorragoa izatea gaztelaniarentzat edo ingelesarentzat baino, eta, bestetik, puntuazioaren iturria interpretatzea zailagoa izatea. Izan ere, egun badaude karaktere mailan lan egiten duten metrikak ere; chrF da ezagunena (Popovíc, 2015), eta gero eta hedatuago dago, BLEUren kaltetan.

Azken kezka bat aipatzearren, errepara diezaiogun BLEUren emaitzari. Metrikak zenbaki bat ematen digu. Zenbaki horren esanahia interpretatzean dago koska. Badakigu zenbaki txiki bat kalitate txarraren adierazle dela. BLEUk 5 puntu esleitzen badizkio itzulpen automatikoari, gaizki gabiltza. Ez dakigu zergatik, baina gaizki. Zer esan nahi du, ordea, 15 puntu jasotzeak? Eta 25? Konparaketak egiteko baliagarria zaigu metrika. Aldiz, kalitate absolutuaren berri jakitea ez da batere samurra. Erabilera masiboaren poderioz ikasi dugu hobe dugula 40 puntutik gora lortu erabilgarria izan daitekeen kalitate baterantz hurbiltzeko. Alabaina, atalase hori aldagarria da hizkuntza-parearen, testu-motaren, gaiaren, eta abarren arabera.

BLEUren eta antzeko metriken aurrean, badira bestelako hurbilpenak darabiltzaten proposamenak ere. Azpimarragarria da TER metrika (Snover et al. 2006), itzultzailearen ikuspuntua kontuan hartzen saiatzen dena. TERek itzulpen automatikoaren eta erreferentziazko itzulpenaren arteko distantzia neurtzen du; hain zuzen ere, zenbat hitz ordeztu, gehitu, kendu eta lekualdatu behar diren itzulpen automatikoa erreferentziazko itzulpen bihurtzeko. Eredu horren xedea TERek postedizio-lanaren gaineko informazioa eskaintzea da, hau da, saiatzen da atzematen zenbat lan egin behar den makinaren itzulpenean hura egokia izan dadin.

Profesiora gehiago hurbiltzen da TER metrika. Hala ere, kritikak jaso izan ditu. Besteak beste, produktuan zentratzea eta prozesuaz ahaztea leporatzen zaio. Zenbaketek erreferentziazko itzulpena dute eredu, baina horrek ez ditu beti jasoko itzultzaileak itzulpen horretara iristeko bidean egin eta desegin dituen moldaketa eta aldaketa guztiak. Posible da hainbat saiakera egin ondoren itzultzaileak erabakitzea hitz bakarra aldatzea dela aukera egokiena. Halaber, posible da hitz kopuru nabarmena aldatzea baina kognitiboki esfortzurik apenas eskatzen duten aldaketak izatea. Horrez gain, testu automatikoa eraldatzeko zenbat edizio egin beharko liratekeen kalkulatzean, metrikak ez du pertsonok jardungo genukeen intuizio linguistikorik erabiltzen, optimizazio matematiko hutsa baizik; hortaz, postedizioaren lan-kargaren doitasuna are gehiago lauso daiteke.

Orain arte deskribatutako metrikei doitasun-neurri lexikalak deritze; hain justu, hitz mailako zehaztasunean oinarritzen dituztelako kalkuluak. Urteak dira ikertzaileek metrika berritzaileen beharra aldarrikatzen dutela, maila lexikoan diharduten metrika horien errendimendu eskasa arrazoi. Sormena sustatzeko asmoz, besteak beste, lehiaketa antzeko bat abiarazten da urtero itzulpen automatikoko nazioarteko Conference on Machine Translation (WMT) bilkuraren barruan. Ebaluazio-erronka jakin bat definitzen dute antolatzaileek, eta ikertzaileei dei egiten zaie neurri berriak proposa ditzaten, konparatu eta eraginkorrena irabazle izendatzeko. Inor ez da harrituko jakitean zer noranzko hartu duten metrikek azken hiruzpalau urteetan. Gero eta maizago, sare neuronaletan oinarritutako proposamenak aurkeztu dira, baita, 2020tik, garaile atera ere behin eta berriz, alde gero eta nabarmenagoarekin.

Metrika-familia berri horren oinarrian ez dago itzulpen automatikoaren eta erreferentziazkoaren arteko aldea neurtzea. Horren ordez, pertsona batek itzulpen automatiko bati zer kalitate esleituko liokeen aurreikusten ikasten dute, demagun, otik 100erako eskala batean. Metrika

horietako bat COMET da (Rei et al. 2020), Unbabel enpresako ikerlariek proposatua, eta gero eta zabalduagoa bai ikerketan, bai industrian.

COMET ereduek egun modan dagoen hizkuntza-eredu masibo bat dute oinarrian, hitz bat emanda hurrengoa zein den asmatzeko entrenatzen diren ereduetako bat. Agian ezaguna egingo zaizue XLM-RoBERTa (Conneau et al., 2020). 100 hizkuntza barne hartzen dituen hizkuntza-eredu eleaniztuna da, 2,5 TB testurekin sortua, preseski CommonCrawl deritzon ekimenak eskuragarritako web-testuekin eta Wikipediako testuekin. Datu horiek ardatz harturik, hizkuntza-eredua berrentrenatu egiten du COMETek, hainbat hizkuntza paretako 900.000 esaldi paraleloen multzo batekin itzulpenen kalitatea aurreikusten ikas dezan. Zehatzago esateko, sorburuko esaldiekin, haien itzulpenekin (zeinak kalitate maila askotakoak baitira, COMETek denetik *ikus* dezan) eta itzulpenen kalitate-informazioarekin. Tamalez, datu mota hori ez da erabili 100 hizkuntzetarako eta haien arteko konbinazio guztietarako; hein handi batean, eskuragarri ez dagoelako.

Edonola ere, teorian, COMETek balio beharko luke 100 hizkuntza horietarako, baina ez da azterketa sakonik egin egiaztapen horren gainean. Egindako gutxietan, emaitzak ez dira argiak. Esaterako, Mathur et al. (2020) ikerlanean, ingelesa-inuktitut hizkuntza-parerako fidagarritasuna behatu zen. Ingeleserako datu mordoa du COMETek; izan ere, baliabide gehien duen hizkuntza da. Aldiz, inuktituta ez du ezagutzen, hau da, ez dago ez 100 hizkuntzen zerrenda horretan, ez itzulpenen corpus paraleloan. Hala ere, COMETen aurreikuspenak eta pertsonen iritziak antzekoak izan omen ziren hizkuntza-pare horretarako (0,6 eta 0,8 tarteko Pearson korrelaziodunak). Orain arte egindako azterketa zabalenari erreparatuz gero, ordea, lortutako emaitzek pentsarazten digute metrika ez dela gai 100 hizkuntza horietatik at daudenak fidagarritasunez ebaluatzeko (Kocmi et al. 2021). Gauzak horrela, momentuz, gomendioa da COMETek entrenamenduan ikusi ez dituen hizkuntzetarako ez erabiltzea.

Euskararako COMETen bila

Euskaran arreta jarriz gero, nabarmenak dira COMETek dituen murriztapenak. Txalotzekoa da, munduko hizkuntza-kopurua eta gurearen tamaina kontuan izanda, 100 hizkuntzako zerrenda horren barruan euskara egotea. XLM-RoBERTak euskarazko Wikipediako eta sareko testuak jasotzen ditu. Alabaina, ez gehiegi poztu, euskarazko adibideak ia-ia arbuiagarritzat jotzeko modukoak baitira, eta, proportzioan, gutxi-gutxi. Edonola ere, esan dezakegu COMETek euskara *ikusi* duela. Baina metrikak ez dauka euskararako itzulpen-informaziorik. Ereduan ez da erabili euskararen eta beste hizkuntzen arteko corpus paralelorik, ezta itzulpen-kalitateari buruzko informaziorik ere. Egoera hori dela eta, susmoa dugu COMET metrika baliagarria gerta litekeela euskararako etorkizunean, jada beste hainbat hizkuntzatarako den bezala, baina egun tentagarria bezain arriskutsua dela bere horretan erabiltzea.

Euskara ez da egoera horretan dagoen bakarra, inondik ere. Besteak beste, indo-ariar hizkuntzak ere kinka bertsuan daude. Horren aurrean, ereduek babesten ez dituzten hizkuntzen-

tzat COMET nola garatu aztertu nahian dabiltza zenbait ikerlari (Sai et al. 2023). Lan-ildo horri jarraipena emateko eta euskararako itzulpen automatikoaren kalitatean aurrera egiteko asmotan ekin diogu guk ere lanari.

Gaztelaniatik abiatuta, COMET bide anitzetatik gara genezake euskarazko itzulpen automatikoaren kalitatea neurtzeko, dela hizkuntza-eredu masibo eraginkorrago bat erabilita, dela itzulpen-corpus paraleloa erabilita, adibidez. Zehatz-mehatz, bi aukera horiek probatu nahi izan ditugu.

Proba horiek egiteko, batetik, XLM-RoBERTa ordezkatzeko hizkuntza-eredu masibo adierazgarriagoa behar dugu euskararako. Zehazki, IXAmBERT (Otegi et al., 2020) ereduaz baliatu gara, ingelesa, gaztelania eta euskara bakarrik hartzen dituen eredu bat, 100 hizkuntza beharrean. Eredu horretan, euskararen presentzia orekatua da beste bi hizkuntzekin alderatuta, eta espero dugu horri esker COMETek hizkuntzaren jakintza hobea izatea.

Horrez gain, gaztelania-euskara esaldi paraleloak behar ditugu. Baina ez hori bakarrik: ezinbestekoa da esaldi horiek ebaluatuta egotea. Esaldi paraleloak jada itzulita dauden testuetatik jaso genitzake. Benetako zailtasuna, eta une honetan euskarak ez duena zera da, kalitate desberdineko esaldi pareen corpus bat berariazko ebaluazio-informazioa duena. Hain zuzen ere, datu multzo hori sortzea gertatu da ikerketa-saiakera honen erronka nagusia.

Ebaluazioak biltzeko ekimena

Beharrezko datuak biltzeko, bide eraginkor posiblea iritzi genion gaztelaniatik euskarara itzulitako esaldien kalitatea ebaluatzeko ekimen bat martxan jartzeari. Ez da ikerketa taldetik euskal komunitatera jotzen dugun lehen aldia, eta esperientziak erakutsi digu badagoela laguntzeko prest dagoenik. Horrela, lankidetza irekiko ebaluazio bat edo *crowd-based evaluation* deritzona abian jarri genuen 2023ko udaran.

Ekimenaren diseinuari dagokionez, askotariko erabakiak hartu ziren. Gehien-gehienak itzulpen automatikoko komunitatearen gomendioei jarraituta hartutakoak izan ziren, betiere gure egoerara ekarrita eta gure momentuko baliabideei erreparatuta.

Ebaluazio-esaldiei dagokienez, ekimenean erabilitako multzoak gaztelaniazko 400 esaldi zituen. Konkretuki, FLORES-2002 (Costa-Jussà et al. 2022), TED2020 (Reimers eta Gurevych, 2020), OpenSubtitle (Lison eta Tiedemann, 2016), Elhuyar Corpusa eta Hizkuntzen arteko Corpusa iturrietatik erauzi genituen. Horrek askotariko gaiak eta estiloak biltzea ahalbidetu zigun: web-artikuluak zein azpitituluak eta literatura. Esaldi-multzoa finkatzeko orduan agertu zen zailtasun nabarmenena kopurua zehaztea izan zen. Izan ere, horixe da ikerketa-galera garrantzitsuenetako bat; hain zuzen, zenbat informazio eman behar zaion COMETi doitasunez funtziona dezan. Esan bezala, apustua 400 esaldi-parerekin lan egitea izan zen, behaketa-lana hasteko.

Aipatutako gaztelaniazko esaldien euskarazko itzulpenak iturri anitzetatik lortu genituen. Batetik, hiru IA sistema erabili genituen automatikoki itzultzeko; preseski, HiTZ zentroan

probetarako darabilgun sistema propioetako bat, Eusko Jaurlaritzaren itzuli sistema, eta Meta enpresa buru den No Language Left Behind (NLLB) ekimeneko gaztelania-euskararako itzulpen-eredua (Costa-Jussà et al. 2022). Bestalde, kontrol-esaldi gisa, nahita okertutako IA sistemetako itzulpenak ere gehitu genituen, eta erreferentziazko zenbait itzulpen. Azken horiek pertsonen ebaluazioak egokiak direla ziurtatzeko erabiltzen ditugu nolabait, hau da, gai badira itzulpen okerrei ebaluazio baxua esleitzeko eta erreferentziazkoei ebaluazio altua esleitzeko, fida gaitezke itzulpen automatikoei esleitutako ebaluazioaz. Guztira, 1.500 esaldi-pare inguru prestatu genituen, IA komunitatean baliatzen dituzten kopuruetatik urruti, baina hizkuntza txiki batentzat eta boluntarioei eskatzeko kopuru adigarria.

Ebaluatzaileei honako ariketa hau egiteko eskatu zitzaien: sorburuko esaldia eta haren euskarazko itzulpen bat irakurrita, 0 eta 100 bitarteko eskala jarraituan itzulpenaren kalitatea adieraztea. Parte-hartzaileek nahi adina esaldi ebalua zitzaketen. Ariketa sareko Appraise plataforman (Federmann, 2012) egiteko prestatu genuen, hara iristeko esteka hainbat kanaletatik argitaratuta; besteak beste, ItzuL posta-zerrenda, EHUko Letren eta Informatikako Fakultateetako posta-zerrenda eta sare sozialak.

Izan ere, ebaluatzaile boluntarioak hurbilaraztea ez da erronka makala. Gogoan izan hainbat eta hainbat ebaluazio bildu beharra geneukala. Tamaina horretako esfortzu bati aurre egiteko, ezinbestekoa da jende kopuru handi batek gogoz parte hartzea. Eskakizun horrek ekarri zuen ebaluatzaileen profil malgua definitu izana; zehazki, gaztelaniaz eta euskaraz gutxienez C1 maila izatea edo biak ala biak ama-hizkuntza izatea. Itzulpen-gaitasuna ez genuen aipatu, nahiz horrek ahuleziak ekar ditzakeen. Hori horrela, itzultzaile profesionalek zein hizkuntzetan inolako prestakuntza bereziturik ez duten pertsonek parte har zezaketen ebaluazioan.

44 pertsonak hartu zuten parte ekimenean. Guztira, 1.215 ebaluazio biltzea lortu genuen: 996 itzultzaile automatikoei zegozkienak, 133 okertutako esaldienak eta beste 86 erreferentziazko esaldienak. Kontrol-esaldiek erakutsi ziguten zortzi boluntariok ez zutela ondo erantzun kontrol-esaldietakoren bat, hau da, erreferentziazko esaldiren bati edo gehiagori 50 puntu baino gutxiago esleitu zizkietela edo okertutako itzulpenei 50 puntu baino gehiago esleitu zizkietela. Tamalez, pertsona horien lana baztertu egin behar izan genuen bildutako datuetatik, fidagarritasun-arrazoiak direla medio. Horrela, euskararako itzultzaileetarako bildutako 996 esaldietatik 635 bakarrik baliatu genituen COMETekin lehen probak egiteko. Hala ere, esaldi guztiei etekina ateratzeko proba gehiago egiten ari gara.

COMETen garapeneko lehen urratsen emaitzak

Arestian aipatu dugu gure ikerketaren motibazioa COMETek euskararako eskaintzen duen fidagarritasuna hobetzea dela, haren garapen-ezaugarriek pentsarazten baitigute ez duela doitasun onargarririk izango bere horretan. Haatik, komeni zaigu hori hala den egiaztatzea. Izan ere, giza ebaluazioak ahalbidetuko digu COMETen eta pertsonen arteko balorazioen antzekotasuna (edo haren eza) neurtzea.

1. taulan ageri dira proba horren emaitzak. Giza ebaluazioak adierazten digu, otik 100erako eskala batean, IA sistema bakoitzari batez besteko zer kalitate esleitu dioten lankidetza irekiko boluntarioek orokorrean. Ikusten denez, EHU-IA eta itzuli sistemen kalitatea pare-parekoa da; bigarrena pixka bat aurretik doa, eta oso ona da, antza. NLLB itzultzailea kaskarrago omen dabil, 63,60ko batezbesteko kalitatea lortuta. COMETen emaitzei begiratzen badiegu, ohartuko gara giza ebaluazioarekin bat ez datorren informazioa ematen digutela. Hemen, inportantea da gogoan izatea COMETen puntuazioa 0 eta 1 artekoa izan daitekeela; zenbat eta altuagoa, hobea. Metrika horren arabera, sistemak oso onak dira, hirurak oso-oso berdintsuak. Haren ustean, EHU-IA litzateke onena; itzuli jartzen du bigarren postuan, ia-ia puntuazio berarekin, eta NLLB pixka bat atzerago. Lehen begiratuan giza ebaluazioarekin bat datorrela dirudien arren, NLLB sistemaren emaitzari erreparatzen badiogu, argi dago metrikak ez daukala kalitatea diskriminatzeko gaitasunik. Giza ebaluatzaileek *gainditu* xume bat esleitutako sistemari *oso ondo* bat esleitu dio COMETek. Portaera horrek, gainera, zalantzan jartzen du beste bi sistemei esleitutako kalitatearen kalkulua ere.

	Metrika automatikoak				Giza ebaluazioa
IA sistema	COMET	BLEU	TER	chrF	0-100
EHU-IA	0,8371 #1	15,61 #2	78,50 #2	54,36 #2	82,42 #2
itzuli	0,8367 #2	15,35 #3	79,20 #3	54,17 #3	82,81 #1
NLLB	0,8282 #3	27,19 #1	69,36 #1	56,80 #1	63,60 #3

1. taula. Sistema mailako emaitzak hiru IA sistemetako metrika automatikoen arabera eta giza ebaluazioaren arabera.

Emaitza horien aurrean, bi COMET metrika garatu genituen, gaztelaniatik euskararako itzulpen automatikoak fidagarritasun handiagoz ebaluatzeko gai ote ziren ikusteko. Lehenengoa, COMET-DA, hutsetik sortu genuen, IXAmBERT hizkuntza-eredua oinarri hartuta eta giza ebaluaziotik jasotako itzulpen-informazioa baliatuta. Bigarrena, COMET-22-FT, COMET eredu orokorra itzulpen-informazioarekin berrentrenatuta sortu genuen. Bi kasuetan, itzulpen-corpuseko 635 kasuetatik 535 erabili genituen entrenamendurako, eta 100 baztertu genituen metrikak ebaluatzeko.

Metriken fidagarritasuna neurtzeko, hau da, fidagarritasuna kalkulatzeko eta COMET eredu orokorrarekiko alderik ba ote zeukaten aztertzeko, bai eredu horri bai bi COMET eredu berriei eskatu genien 100 kasuko multzoaren kalitatea emateko. Hiru COMETen aurreikuspenen ta giza ebaluazioen arteko korrelazioa neurtuta, iragarpenen arteko korrelazioa ahula zela behatu genuen (Kendallen Tau 0,2 inguruan) (ikusi 2. taula). Haatik, badaukagu saiakera bertan behera ez uztera bultzatu gaituen emaitzarik ere; izan ere, COMET orokorrarekin alderatuta, COMET-22-FTren doitasuna altuagoa da. Nahiz eta fidagarritasun baxuko metrika sortu dugun, ondorioztatu dugu ezen, euskararako itzulpen-informazioa txertatuz gero, onura

estatistikoki esanguratsua dela. Kontua da erabilitako datu-multzoa ez dela nahikoa zehaztasun handiko metrika bat garatzeko. Izan ere hortxe dago koska, entrenamendurako itzulpendatuen kopuru egokia aurreikustean.[1]

Metrika	τ	ρ	r
BLEU	0,021	0,025	0,042
chrF	0,161	0,236	0,192
TER	-0,023	-0,022	-0,053
COMET-22	0,223	0,326	0,214
COMET-DA	0,119	0,172	0,169
COMET-22-FT	**0,245**	**0,354**	**0,242**

2. taula. Kendall Tau (τ), Spearman (ρ) eta Pearson (r) korrelazio-balioak ebaluazio-kasuetan. Letra lodiz ageri dira estatistikoki esanguratsuak diren emaitzak ($p > 0,05$).

Begirada sakonagoa ebaluazio-ekimenari

Ebaluazio-metrika automatikoek zeresana izango badute IA sistemen garapenean, ziurtatu beharko dugu metrika horiek entrenatzeko erabiltzen ditugun datuak —hots, oinarriko hizkuntza-eredu masiboa eta, bereziki, itzulpen-informazioa biltzen duen corpusa— kalitatezkoak direla. Gogoeta horrek eraman gintuen lankidetza irekiko ebaluazio-ekimena sakonago aztertzera, haren balioaz eta ahuleziez hausnartzeko.

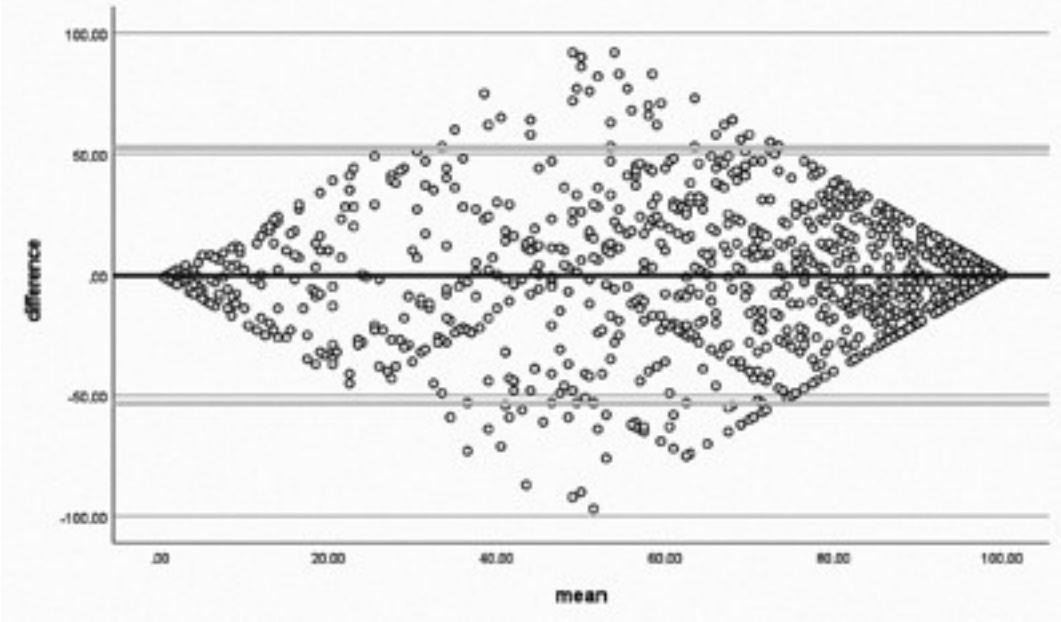
Ebaluatzaile-profil murritzen aurrean parte-hartze zabalagoko boluntario-profilak baliatzeak nahi gabeko ondorioak ekar ditzake; txarrenean, gerta daiteke bildutako informazioa erabat doia ez izatea. Desbideratze hori behatzeko premiaz, ebaluatzaile profesional batek berrebaluatu zuen deskribatu berri dugun ebaluazio-ekimenean bildutako esaldi-pare multzo osoa. Nahiz eta profesional bakar baten iruzkinak ez diren nahitaez esaldi-pareen egiazko kalitatearen adierazle egoki bakarra, argudia genezake hark emandakoak aukera on posibleak direla eta emaitza koherentea izango dugula multzo osoan zehar.

Boluntarioen ia baldintza beretan (Appraise erabili beharrean kalkulu-orri bat erabilita), ebaluatzaile profesionalak 782 esaldi-pare ebaluatu zituen. Multzo horretan, 40 esaldi-pare errepikatuak ziren, ausaz multzoan zehar barreiatuta, ebaluatzailearen lanaren barne-sendotasuna neurtzeko, hau da, behatu nahi genuen ea esaldi-pare bera erakutsiz gero zer neurritan esleitzen zion itzulpen-kalitate bera. Hala kalkulatu genuen 0,896ko intrakorrelazio koefizientea (ICC,

1. COMET euskararako eta malterarako doitzeko egindako proben xehetasun gehiago aurkitzeko, ikus artikulu hau: Júlia Falcão, Claudia Borg, Nora Aranberri eta Kurt Abela. 2024. COMET for Low-Resource Machine Translation Evaluation: A Case Study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (COLING-LREC), maiatzak 20-25, Turin, Italia.

ingelesezko siglak erabilita), ia *bikaina* deritzona, hau da, ebaluatzaileak erakutsi zuen bere lanaren barne-korrelazio oso altua zuela.

Hori horrela, profesionalaren eta boluntarioen ebaluazioen arteko korrelazioa kalkulatu genuen, ikusteko ea zenbateraino zetozen bat. Haien ICCa 0,768koa zen, hots, korrelazio *ona* zeukaten. 1. irudian ikusten da korrelazio hori, Bland-Altman grafiko batean irudikatuta. Batetik, ikusten dugu erdiko marra, batezbestekoen diferentziaren ordezkaria, zerotik oso hurbil dagoela eta puntu gehien-gehienak ausaz sakabanatuta daudela goiko eta beheko marrek osatzen duten espazioan. Horrek adierazten digu profesionalak eta boluntario-taldeak egindako balorazioak bereziki antzekoak direla.



1. irudia. Bland-Altman grafikoa, ebaluatzaile profesionalaren eta boluntarioen arteko adostasuna islatzen duena. Y ardatzean, haien ebaluazio zehatzen arteko desberdintasuna adierazten da, eta X ardatzean, berriz, bien ebaluazio zehatzen batezbestekoa.

Izan ere, batak eta besteek IA sistemei esleitutako kalitatea elkarren ondoan jarriz gero, ikusten dugu ia berdina dela (ikusi 3. taula). Biek ere itzuli sistemari deritzote onena, eta gertugertu dauzka ikerketako EHU-IA sistema. EHU-IAri eta itzuli sistemei esleitutako batezbesteko kalitateari erreparatuz gero, puntu bakarreko aldea dago, 0-100 eskalan. Atzerago geratzen da NLLB sistema, 61,20 eta 58,78 punturekin, hurrenez hurren. Kalitate kaskarragoa esleitu zaio azken horri, eta profesionala boluntarioak baino zorrotzagoa izan da.

IA sistema	Lankidetza irekiko ekimena		Ebaluazio profesionala	
	Batez besteko kalitatea	Desbideratze estandarra	Batez besteko kalitatea	Desbideratze estandarra
EHU-IA	77,81	23,73	78,05	23,07
itzuli	78,45	23,32	79,89	20,94
NLLB	61,20	29,17	58,73	25,26

3. taula. Lankidetza irekiko ekimeneko parte-hartzaileek eta ebaluatzaile profesionalak IA sistemei esleitutako kalitate-emaitzen batezbestekoa eta desbideratze estandarra.

Emaitza horiek erakusten digute ekimeneko boluntario taldeak profesional jakin horren pare-parean ebaluatu duela kalitatea. Badakigu, ondo jakin ere, boluntario-multzo horretan badagoela parte-hartzaile profesionalik. Tamalez, etika-auziak direla bide, ez daukagu jakiterik profesional eta ez-profesionalen proportzioa zein den, eta, egia esan, ez dakigu ea horrek eraginik ba ote duen emaitzak pareko(ago)ak izateko orduan. Izan ere, horixe dugu ikergai garrantzitsua hurrengo saiakerarako. Emaitzek erakusten dute ezen, sistemen kalitate orokorra zein den jakiteko, berdin dela profesional bakarraren lan mardula baliatzea edo lankidetza irekiko ekimena martxan jartzea.

Kontua da, baina, COMETen gisako metrikak garatzeko, ez dugula sistema mailako balorazioa baliatzen, baizik esaldiz esaldi emandako puntuazioa. Eta aldeak nabarmenagoak dira hor. Boluntario bakoitzaren eta profesionalaren arteko korrelazioa kalkulatuz gero, ICC emaitzen arabera, hiru pertsonak oso korrelazio ahula dute, hamabi pertsonak erdi mailakoa, hemeretzik ona eta bederatzik bikaina. Hortaz, metrikentzat kalitatezko informazioa bildu nahi izanez gero, mereziko luke gai izatea profesional batekin korrelazio ahula (eta, agian, erdi mailakoa ere bai) duten pertsonen lana identifikatu ahal izateko eta baztertzeko. Une honetan, ez daukagu horretarako biderik, eta horixe da etorkizunerako bigarren ikergai garrantzitsua.

Esaldi mailako itzulpen-kalitatearen balorazioen fidagarritasuna identifikatzearen garrantziaz jabetzeaz gain, diseinuari lotutako zenbait erabaki zalantzan jartzera ere eraman gaitu ebaluazio-ekimenaren gaineko hausnarketak. Hona hemen, labur-labur, gogoetarako ideia batzuk:

- Testuinguruaren eragina: itzultzaile profesional oro jabetzen da sorburuko esaldi baten baliokide egokia proposatzeko ezinbestekoa dela haren testuingurua ezagutzea, bai testu barrukoa, bai testuz kanpokoa. Hori horrela, ebaluazio-diseinuari egin diezaiokegun kritiketako bat litzateke esaldi-pareak testuingururik gabe ematea. Nola jakin ea itzulpena egokia den dagokion inguruan? Bestalde, ezin diogu itzulpengintzan trebatu ez denari halako kontzientziarik eskatu, eta horrek eragina izan lezake haren kalitatearen pertzepzioan.
- Sorburuko esaldiaren eragina: mota honetako ebaluazioetan, sorburuko esaldiari, haren lexikoari eta egiturei dagokien baino garrantzi gehiago ematen zaie. Testuinguru ezak

bultzatuta, litekeena da ebaluatzaileek sorburuko esapideak gertuagotik jarraitzen dituzten itzulpenei kalitate altuagoa esleitzea, sormen eta naturaltasun handiagoa duten baina gehiago aldentzen diren proposamenei baino.

- Sorburuko esaldien kalitatea: ekimenean erabilitako esaldi-multzoan, behin baino gehiagotan behatu dugu sorburuko esaldien kalitatea ez zela espero genezakeen bezain ona; besteak beste, hizkuntza-kalitateagatik. Horrelako adibideen aurrean, ebaluatzaileek aukera bat baino gehiago izaten dute; esaterako, okerra oker itzultzeagatik penalizaziorik ez esleitzea itzulpenari edo okerra ez zuzentzeagatik itzulpena kaskartzat jotzea. Jarraibide gehigarririk gabe, gerta daiteke kalitate bereko esaldiei balorazio desberdina esleitzea.
- Adierazpen-egokitasunaren eragina: egungo sistema neuronalen itzulpen-proposamenen bereizgarrietako bat da hizkuntza aski zuzena eta naturala erabiltzea. Halaber, mezuaren doitasun falta ezkuta dezake ezaugarri horrek, edo, are okerrago, gerta daiteke mezu okerra islatzea. Horren arriskua zera da: sorburuko esaldia konplexua edo domeinu oso zehatzekoa bada eta ebaluatzailea profesionala ez bada, kalitate ona eslei dakioke zuzena ez den itzulpenari.
- Ebaluazio-eskalaren doitasun maila: itzulpen automatikoaren komunitatean adostasunik lortzen ez duen alderdietako bat ebaluazio-eskalaren irismena da. Badago 4 puntukoa izan behar dela dioenik, edo 5 puntukoa; 7 puntukoa ere erabilia da, eta, ebaluazio-kanpaina handietan, azken hiruzpalau urteetan, 100 puntuko barra jarraitua gailendu da. Azken aukera horrek, zeina guk geuk ere aplikatu baitugu, kalitatea doitasun handiz baloratzea ahalbidetzen du. Haatik, ez da aztertu gehiegizko eskakizuna ez ote den lankidetza irekiko ekimen baterako; izan ere, ez da definitzen ebaluazio-lan minimorik, eta, hortaz, boluntarioak ez du beti izaten eskalara ohitzeko betarik.
- Penalizazioen zorroztasuna: ebaluazio mota honetako zailtasun azpimarragarriena itzulpen oker edo kaskarrei aplikatu beharreko penalizazioa erabakitzean datza. Gehien-gehienetan, jarraibideek ez dute horren gaineko argibiderik ematen, eta ebaluatzailearen esku uzten da kasu bakoitzak merezi duen puntuazioa adieraztea. Profesionala izan zein ez, ez da erraza horretan bat etortzea. Gainera, ez da zehazten ea errore garrantzitsu batengatik ebaluatzaileak zuzenean esleitu behar ote duen 50 puntutik beherako kalitatea edo ea itzulpen-zati zuzenek orekatzen ote duten azken puntuazioa. Alegia, balorazio sendoak jasotzeko aukerak murriztu egiten dira.
- Ebaluatzaileen ustezko itzulpen-gaitasuna: ebaluatzaileen profilarekin zerikusi zuzena duen alderdi gehigarri bat da zer eragin izan dezakeen ebaluatzaileek beren itzulpen-gaitasunaz duten pertzepzioak. Litekeena da itzulpen-proposamen bat irakurtzerakoan norberaren hizkuntza eta itzulpen gaitasuna islatzea emaitzan eta horren araberako kalitatea esleitzea. Hobea bada, altua; ahulagoa bada, baxua. Profesionalen kasuan, joera horrek ez luke aparteko eragin kezkagarririk ekarriko, baina kontuan hartzekoa izan liteke ebaluatzaile ez-profesionalak ditugunean.

- Postedizio-efektua: inoiz ebaluatzaileek adierazi izan dute proposamena hobetzeko egin beharreko edizio-lanarekin erlazionatzen dutela itzulpen-kalitatea. Horrek zera ekar dezake, postedizio-lana baloratzea, eta ez itzulpen-kalitatea, posible baita errore handi bat erraz-erraz konpontzea eta, alderantziz, oker txiki bat zuzentzeko lanak hartu beharra.[2]

Ez dugu pentsatu behar ebaluazioaren diseinua egiterakoan aurrekoak kontuan hartu ez zirenik. Erabakiak ikuspuntu praktikotik hartu zirela esatera ausartuko naiz; alegia, momentuko baliabideen araberakoa izan zela (datuak, denbora eta esfortzu pertsonala), eta, esan bezala, ahal izan zen neurrian itzulpen automatikoko komunitatearen gomendioei jarraituta. Hala ere, ekimeneko emaitzak aztertu ondoren, diseinuko zenbait aukera berrikusi behar direlakoan gaude, horrelako egitasmo batek berez dakarren esfortzuari (bai ikertzaileen, bai boluntarioen aldetik) ahalik eta etekin handiena ateratzeko. Tamalez, oro har baliabide askoko hizkuntza eta testuinguruei begira egin ohi direnez horrelako ebaluazioak, gutxi hitz egiten da zer eraginkortasun duten baliabide urriko testuinguruetan. Bada, agian, guri dagokigu horren eredu izatea.

Etxera eramatekoak

Euskararako itzulpen automatikoaren garapenean aurrera egiteko eta haren erabilera ahalik eta bidezkoena bultzatzeko, argi dago ebaluazio-metrika automatiko fidagarriak behar direla. Egun gailentzen ari diren metrikak sare neuronaletan oinarritutakoak dira, eta, ondorioz, entrenatu beharrak dira. Baliabide askoko hizkuntzak aurreratuta dabiltza lan horretan, eta emaitza txukunak lortu dituzte.

Aurreko metrika-moten aldean, COMET eta antzeko metrikak euskararako doitzeko, batetik, hizkuntza-eredu masibo eraginkor bat behar dugu, eta, bestetik, sorburuko testuekin, haien itzulpenekin eta kalitateari buruzko informazioarekin osatutako corpus handi-handi bat. Hizkuntza-eredu masiboetan gabiltza lanean HiTZ zentroan, eta hala sortu dira, besteak beste, BERTeus (Agerri et al. 2020) eta IXAmBERT (Otegi et al. 2020) eredu eleaniztun orekatuak eta Latxa (Etxaniz, et al. 2024) eredu elebakarra. Itzulpenen kalitatea jasotzen duten esaldi-pareen corpusa biltzea dugu orain erronka nagusia. Egindako lehen saiakeratik ikasi dugu milaka adibide beharko ditugula eta gomendagarria litzatekeela ebaluazio-ekimenaren diseinuan zenbait berrikuntza txertatzea, datuen doitasuna bermatzeko.

Datu horien bilketan ikusiko gaituzue aurki berriro. Alabaina, frogatu dugunez, itzulpenen kalitate-ebaluazioak lortzea eginkizun zaila, handia eta neketsua da, eta ikerketa-zentroetatik elkarlana eskatzen diegu euskal gizarteari eta, nola ez, euskararen profesionalei. Eta, ahaztu gabe, eskerrik asko orain arte erakutsitako prestutasunagatik eta emandako laguntzagatik.

2. Lankidetza irekiko ekimenaren gaineko analisi eta hausnarketa sakonagoak aurkitzeko, ikus artikulu hau: Nora Aranberri (2024). Analysis of the Annotations from a Crowd MT Evaluation Initiative: Case Study for the Spanish-Basque Pair. *Proceedings of the 25th Annual Conference of The European Association for Machine Translation*, ekainak 24-27, Sheffield, Erresuma Batua.

ERREFERENTZIAK

Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa eta Eneko Agirre (2020). Give your Text Representation Models some Love: the Case for Basque. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4781–4788 or., Marseilla, Frantzia. European Language Resources Association.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer eta Veselin Stoyanov (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Sarean. Association for Computational Linguistics.

Costa-Jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard… eta Jeff Wang (2022). No Language Left Behind: Scaling human-centered machine translation. arXiv: 2207.04672

Etxaniz, Julen, Sainz, Oscar, Perez, Naiara, Aldabe, Itziar, Rigau, German, Agirre, Eneko, Ormazabal, Aitor, Artetxe, Mikel eta Soroa, Aitor (2024). Latxa: An Open Language Model and Evaluation Suite for Basque. arXiv preprint arXiv: 2403.20266.

Federmann, Christian (2012). «Appraise: an open-source toolkit for manual evaluation of mt output», *Prague Bulletin of Mathematical Linguistics*, 98: 25–36.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita eta Arul Menezes (2021). To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. *Proceedings of the Sixth Conference on Machine Translation*, 478–494. Sarean. Association for Computational Linguistics.

Lison, Pierre eta Jörg Tiedemann (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 923–929. Portorož, Slovenia. European Language Resources Association (ELRA).

Mathur, Nitika, Wei, Johnny, Freitag, Markus, Ma, Qungsong, & Bojar, Ondřej. (2020). Results of the WMT20 metrics shared task. *Proceedings of the Fifth Conference on Machine Translation*, 688–725. Sarean.

Otegi, Arantxa, Aitor Agirre, Jon Ander Campos, Aitor Soroa eta Eneko Agirre (2020) Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 436–442. Marseilla, Frantzia. European Language Resources Association.

Papineni, Kishore, Salim Roukos, Todd Ward eta Wei-Jing Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Filadelfia, AEB. Association for Computational Linguistics.

Popović, Maja (2015). chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395., Lisboa, Portugal. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha eta Alon Lavie (2020). COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Sarean. Association for Computational Linguistics.

Reimers, Nils eta Iryna Gurevych (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525. Sarean. Association for Computational Linguistics.

Sai, Ananya B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra eta Raj Dabre (2023). IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 14210–14228. Toronto, Kanada. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla eta John Makhoul (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. Cambridge, Massachusetts, AEB. Association for Machine Translation in the Americas.