Ikasketa automatikoko tekniken erabilgarritasun-azterketa euskararako postediziorako gomendio-sistema eraikitzeko

(Evaluation of the suitability of machine learning techniques to build a post-editing recommendation system for Basque)

Nora Aranberri[1], *Jose A. Pascual*[2]

[1] IXA taldea, Euskal Herriko Unibertsitatea (UPV/EHU)

[2] School of Computer Science, The University of Manchester

nora.aranberri@ehu.eus
jose.pascual@manchester.ac.uk

DOI: 10.1387/ekaia.19700

Jasoa: 2018-06-26 Onartua: 2018-07-16

Laburpena: Gaztelania-euskara bikotearekin lan egiten duten itzultzaileentzat eskuragarri dagoen itzulpen automatikoaren kalitatea nahiko baxua da, eremu profesionalean erabiltzea oztopatzen duena. Lan honek posteditatzea edo itzultzea, bietatik eragingarriagoa zein den erabakitzen laguntzeko gomendio-sistema bat eraikitzearen egingarritasuna ikertzen du, datu-multzo mugatu batekin ikasketa-algoritmoen jokaera aztertuz. Lehenengo, postedizio-esfortzua aurreikusten duten erregresio-ereduak eraiki ditugu kalitate orokorrean, denboran eta edizioetan oinarritzen direnak. Bigarrenik, edizio-metodo eraginkorrena gomendatzen duten sailkapen-ereduak eraiki ditugu oinarrizko ezaugarriei eta ezaugarri linguistikoei postedizio-esfortzuko ezaugarriak gehituta. Emaitzek korrelazio altuak erakusten dituzte erregresio-ereduek aurreikusitako HTERen, denboraren eta edizio kopuruaren eta errealen artean. Era berean, sailkatzaileen emaitzek erakusten dute doitasun altuz iragarri dezakegula itzultzea edo posteditatzea, bietatik zein den eraginkorragoa.

Hitz gakoak: Itzulpen automatikoa, postedizioa, gomendio-sistema, euskara.

Abstract: The overall machine translation quality available for professional translators working with the Spanish-Basque pair is rather poor, which is a deterrent for its adoption. This work investigates the plausibility of building a comprehensive recommendation system to speed up decision time between postediting or translation from

scratch using the very limited training data available. First, we build a set of regression models that predict the postediting effort in terms of overall quality, time and edits. Secondly, we build classification models that recommend the most efficient editing approach using postediting effort features on top of linguistic features. Results show high correlations between the predictions of the regression models and the expected HTER, time and edit number values. Similarly, the results for the classifiers show that they are able to predict with high accuracy whether it is more efficient to translate or to postedit a new segment.

Keywords: Machine translation, post-editing, recommendation system, Basque.

1. **SARRERA**

Egunean-egunean itzulpen automatikoa (IA) hobetuz doan arren, euskararako itzulpenak ez du eremu profesionalerako jauzia egin oraindik eta nekez erabiltzen da itzultzaileen artean [14]. Izan ere, gaztelania-euskara hizkuntza bikoterako, esaterako, IAren kalitatea baxua da [2, 4], doan eskuragarri dauden sistemetan behintzat, adibidez, *itzultzailea*[1] edo *Google Translate*[2]. Testuinguru horretan, ikertu nahi dugu posible ote den eraikitzea itzultzaileentzako informatzaileak izan daitezkeen IAren kalitatea estimatzeko ereduak. Itzultzaileek esaldiz-esaldi IA erabili edo ez erraz erabaki ahalko balute, teknologia horren abantailak ustiatzen hasi ahalko lirateke.

Ereduak eraikitzeko, postedizio ikastaro batean bildutako datu-multzo txiki bat baliatuko dugu. Bertan, postedizioak, hau da, itzulpen automatikoaren orrazketak, produktibitatea hobetzen du batzuetan, hutsetik itzultzearekin alderatuz. Lehenengo helburutzat, itzultzaileei edizio-lanaren estimazio-adierazle batzuk ematea dugu, IA proposamen bat posteditatu edo hutsetik itzultzea hobe den erabakitzen laguntzeko. Horretarako, postedizio-esfortzua estimatzen duten zenbait adierazlerentzako erregresio-eredu multzo bat eraiki dugu (IA kalitate orokorra, denbora eta edizio-lana). Emaitzek erakusten dute 0,70eko korrelazioa dagoela benetako adierazleen eta estimatutakoen artean.

Hala ere, ezin uka daiteke eraginkorragoa litekeela metodorik azkarrena zein litzatekeen, posteditatzea edo itzultzea, zuzenean balioetsiko lukeen eredu bat. Eredu honen gomendioa erabil liteke itzultzaile batek, lanean diharduela, zein edizio-metodo erabili erabakitzeko, edo edizio fasea hasi aurretik IA proposamen desegokiak baztertzeko. Horretarako,

[1] http://www.itzultzailea.euskadi.eus
[2] https://translate.google.com

sailkatzaile batzuk eraiki ditugu, edizio-metodorik eraginkorrena zein den gomendatzen dutenak. Eredu horien doitasun baxua dela-eta, emaitzak hobetzen saitu gara postedizio-esfortzuarekin erlazionatutako ezaugarriak erabilita. Hala ere, informazio hori ez dagoenez eskuragarri edizioa bukatuta egon arte, aurreko erregresio-ereduak erabilita estimatu dugu. Emaitzek erakusten dute sailkatzaileen doitasuna nabarmen hobetzen dela, nahiz eta erregresio-ereduetatik lortutako datuek doitasuna galtzea ekarri.

2. **AURREKARIAK**

Atal honetan kalitatearen eta postedizio-esfortzuaren adierazleei buruz literaturan landutakoaren laburpena egingo dugu. 2004an, konfiantza-estimazioko teknikak, ordura arte hizketa-ezagutzan erabiliak, IAren esparrura ekarri zituzten, itzulpenak postediziorako egokiak ziren IA proposamenak aukeratzeko balio lezaketelakoan [11]. Esaldi-mailan lan egiten zuten erregresio eta sailkapen-ereduak entrenatu ziren NIST [12] eta WER [18] ebaluazio metrika automatikoen balioak iragartzeko. Atazek interesa piztu zuten arren, esperimentuek erakutsi zuten iragarritako metrika automatikoen balioak ez zetoztela bat pertsonek esaldiei esleitutako kalitatearekin eta postedizio-esfortzuarekin.

Antzeko emaitzak aurkeztu ziren ondorengo lanetan [22]. Horietan IA sistemaren independente eta dependente ziren hainbat ezaugarri erabili ziren ereduak eraikitzeko. Ondo zebiltzan pertsonen lana balioesteko orduan baina, berriz ere, metrika automatikoen emaitzekin ez zuten korrelazio alturik. Orduan, estimazio-ereduak eraikitzeko itzulpen automatikoaren eta erreferentziazko itzulpen baten arteko edizio-distantzian oinarritutako TER eta haren moldaketa den HTER [21] metrikak erabili ziren [23], zeinek itzultzaileek egiten duten postedizio lana era zuzenagoan kontuan hartzen duten. Oraingoan ereduak bat etorri ziren pertsonen esleipenekin. HTER finkatu zen kalitatearen estimazioa (KE) izeneko atazetan kalitate orokorraren adierazle bezala, eta hala jarraitzen du gaur egun, nahiz eta zenbait saiakera agertu diren kalitatea neurtzeko bestelako bideak aztertu dituztenak [24].

Ordutik, zenbait autore itzultzaileentzako erabilgarriak izan litezkeen informazioa iragarten duten ereduak eraikitzen saiatu dira. Batzuk postedizio-denbora aurreikusten saiatu dira [27], beste batzuk IA proposamen multzo batetik onena aukeratzen [8], edo esaldi berri bat itzultzeko IA sistemaren proposamena edo itzulpen memoria bateko hautagaiak erabili beharko liratekeen gomendatzen [16]. Hala ere, esan genezake kalitatearen balioespenaren arloko ikerketa gehienak 2012tik urtero antolatzen den KE ataza partekatuak bideratzen dituela. Lehenengo urtean,

parte-hartzaileek bost mailatan definitutako eskuzko kalitatea aurreikusten jarri zuten arreta [15, 17]. 2013 eta 2014 urteetan, ataza desberdinak definitu ziren, hala nola HTER eta postedizio-denbora iragartzea eta IA proposamen desberdinak sailkatzea [9, 10]. Ordutik, baina, esfortzu gehiena HTER iragartzean egin da eta, nahiz eta bestelako adierazleei buruzko lanak onartu (postedizio-denbora, pultsazio-kopurua), ez da horrelako lanik argitaratu.

Profesionalei itzulpen lana erraztuko dien adierazle multzo zabal bat eskaintzeko asmoz, lan honetan postedizio-esfortzuko adierazleen kopurua handitu nahi dugu. Zehazki, proposatzen duguna da gomendio-sistema bat garatzea non (1) IAren proposamenen kalitatea iragartzen den HTER metrikan oinarrituta, (2) postedizio-esfortzua iragartzen den denbora eta edizio-mota eta kopuruaren arabera, eta (3) edizio-metodo egokiena gomendatzen den esaldi bakoitzerako, postediziorako edo hutsetik itzultzeko egokiagoa den sailkatuz.

3. **DATU-BILKETA ETA PROZESAMENDUA**

Hizkuntza handientzat ez bezala, gaztelania-euskara bikotearentzat estimazio-ereduak eraikitzeko ez dago datu-multzorik eskuragarri. Hori dela eta, 2015ean itzultzaile profesionalekin egindako postedizio ikastaro batean bildutako datuak moldatu ditugu. Atal honetan datuak eta ezaugarri linguistikoak aurkezten ditugu.

3.1. **Datu-multzoak**

Aipatutako ikastaroan, itzultzaileek hainbat postedizio eta produktibitate ataza egin zizuzten zazpi astetan zehar (ikusi 1. taula). Postedizio atazetan, itzultzaileek IAko proposamenak orraztu zituzten eta bitartean horretarako erabiltzen zuten denbora neurtu genuen, esaldiz esaldi. Horrela posteditatzean zein abiaduratan dabiltzan jakin genezake. Produktibitate atazetan, aldiz, itzultzaileek esaldiak txandaka hutsetik itzuli edo posteditatu egin zituzten. Itzultzaileak bi multzotan banatu genituen, batzuek posteditatzen zituzten esaldiak besteek hutsetik itzul zitzaten eta alderantziz, horrela esaldi guztietarako bai posteditatuko bertsioak eta bai hutsetik itzulitakoak biltzeko. Horrela, esaldi bera itzultzeko eta posteditatzeko zenbat denbora behar den neurtu ahal izan genuen. Informazio horrek ahalbidetzen digu jakitea posteditatzea edo hutsetik itzultzea azkarragoa den.

1. taula. Itzultzaile profesionalek egindako ataza guztien zerrenda.

Ataza zenbakia	Ataza mota	Itzultzailek	Testua	IA sistema	Esaldiak	Iturburuko hitzak
1-4	postedizioa	10	1	itzultzailea	60	1467
5	produktibitatea	10	1	itzultzailea	21	495
6-9	postedizioa	10	1	itzultzailea	81	1958
10	produktibitatea	9	1	itzultzailea	16	506
11-14	postedizioa	8	1	itzultzailea	82	2043
15	produktibitatea	8	1	itzultzailea	22	366
16-19	postedizioa	8	1	itzultzailea	80	1964
20	produktibitatea	8	1	itzultzailea	29	516
21-24	postedizioa	8	1	itzultzailea	138	2045
25	produktibitatea	8	1	itzultzailea	26	515
26-29	postedizioa	6	1	Google Translate	121	2082
30	produktibitatea	6	1	Google Translate	24	508
31-34	postedizioa	5	2	itzultzailea	187	2012
35	produktibitatea	5	2	itzultzailea	60	486

Emakundek argitaratutako txosten bat, «Sexismoa joko eta jostailuen 2013ko publizitate-kanpainan» deritzona (1. testua), eta mugikor baten eta ikuzgailu baten erabiltzaile-gidak (2. testua) landu zituzten. Gaztelaniaz idatzitako jatorrizko testuak Eusko Jaurlaritzak atzigarri daukan *itzultzailea*[3] sistemarekin (Lucy enpresak garatua) itzuli ziren. IAren kalitate orokorra nahiko baxua zen (50,7 HTER) eta profesionalek edizio ugari egin zituzten IAren proposamenak esaldi onargarri bihurtzeko.

Esperimentuetarako (ikusi 4.1. atala), bildutako datuak bi multzotan banatu genituen, postedizio (PE) multzoa eta produktibitate (PR) multzoa. Lehenak postedizio atazetako esaldiak hartzen ditu barne eta bigarrenak produktibitate atazetakoak. Erabaki genuen lehenengo asteko atazak (1-5 atazak) baztertzea hori baitzen itzultzaileek postedizioarekin zuten lehen hartu-emana, eta beraz, egindako lana fidagarria ez zelako. Halaber, erabaki genuen 26-30 atazak baztertzea, beste IA sistema batekin itzuli genituelako esaldiak eta kasu horretan itzultzea posteditatzea baino azkarragoa zelako (ikusi 2. taula). Ondorioz, 568 iturburu esaldirekin egindako lana bildu genuen (10.022 hitz) postedizio atazetatik eta 153 esaldirena (2.389 hitz) produktibitate atazetatik. Itzultzaile guztiek ataza berak egin zituztenez, datu-multzoek iturburuko esaldi bakoitzarentzat hainbat itzulpen dituzte.

[3] http://www.itzultzailea.euskadi.eus

2. taula. Batezbesteko postedizio (PE) eta itzulpen (IT) denbora milisegundotan (ms) hitzeko (h) itzultzaileek egindako produktibitate ataza bakoitzeko.

Ataza	Postedizio denbora	Itzulpen denbora
5	4576,73	**4353,46**
10	3058,86	3882,97
15	2920,31	4400,37
20	3454,05	4224,66
25	3174,79	3520,80
30	3523,23	**2974,36**
35	3054,51	291,58

Azkenik, bildutako datu multzoetan ereduak eraikitzeko beharrezkoa den informazioa gehitu genuen. Alde batetik, postedizio adierazleak iragarri nahi ditugu. Horretarako, PE multzoa erabili genuen, hori baita postedizio lana biltzen duen multzo handiena. Bertan itzultzaileen postedizio bertsioak geneuzkan eta esaldi bakoitzerako edizio-denbora ere bai. Beraz, horretaz gain, HTER balioak, edizio-mota bakoitzaren kopurua eta edizio-kopuru totala gehitu genizkion, automatikoki kalkulatzen direnak IA proposamena eta itzultzaileen lana alderatuta. Bestetik, itzultzea edo posteditatzea gomendatzen duten sailkapen ereduak nahi ditugu. Horretarako, PR multzoa erabili behar dugu, nahiz eta oso mugatua izan, horretan baino ez baitaukagu postedizio eta itzulpen denbora. Esaldi bakoitzarentzat eraginkorrena zen edizio-metodoa zehazteko, denbora-irabazian oinarritutako estrategia erabili genuen: produktibitate ratioa (itzulpen denbora/postedizio-denbora). Ratio hori kalkulatzeko, iturburuko esaldi bakoitzarentzat itzultzaileek erabilitako denboraren batezbestekoa kalkulatu genuen, horrela aldagarritasuna kontuan izanez. Batetik gorako balioek adierazten dute posteditatzea eraginkorragoa dela eta batetik beherakoek itzultzea azkarragoa dela.

Ratioa kontuan izanda, hiru etiketa multzo erabili genituen esaldiei informazio hau gehitzeko, E2, E3 eta E5, non bi, hiru eta bost ekiteka erabiltzen diren hurrenez hurren. E2k zuzenean esleitzen die *posteditatu* etiketa batetik gorako ratiodun itzulpenei (esaldien %63ren kasua) eta *itzuli* etiketa batetik beherakoei (esaldien %37ren kasua). E3an, itzultzaileen arteko aldagarritasuna dela-eta, batetik gertu dauden balioak guztiz zehatzak ez direla onartzen da eta metodo batekin zein bestearekin lortzen den denbora-irabazia nabarmena ez dela. Horrela, 0,90 eta 1,10 bitarteko ratioei *edozein edizio-metodo* etiketa esleitzen zaie. Azkenik, E5ek bi etiketa gehitzen dizkio E3 multzoari metodo batekin eraginkortasuna argi eta garbi hobetzen deneko kasuak identifikatzeko. Hala, 0,70etik beherako eta 1,30etik gorako ratioak bereizi egiten dira.

3.2. **Ikasketa automatikorako ezaugarriak**

Ikasketa automatikoko algoritmoei aurreikusi nahi dugun informazioa ikasten laguntzeko, esaldi bakoitzerako zenbait ezaugarri bildu genituen. *Quest++* [25] erabilita, WMT12-17 KE atazetan ematen diren oinarrizko 17 ezaugarri berak erauzi genituen, IA sistemarekiko independenteak diren azaleko ezaugarriak. Gehienak datu-multzoetako esaldiak entrenamenduko corpus handi bateko esaldiekin konparatuta lortzen dira, adibidez, hizkuntza-ereduko probabilitateak, n-grama maiztasunak eta hitz bakoitzeko itzulpen posibleak.

Horretarako erabili genituen gaztelania eta euskarako corpusek 38 eta 44 milioi esaldi dituzte, hurrenez hurren. Gaztelaniakoak WMT atazetan atzigarri jarritako datuak ditu (Europarl corpusa, NBko corpusa, News Commentary corpusa). Euskarakoak, iturri desberdinetatik hartutako testuak ditu, hala nola *Egunkaria* koak, *EITB* koak [13], *Elhuyar Web Corpus*[4]ekoak eta administrazioko itzulpen memorietakoak. GIZA++ [19] entrenatzeko erabilitako corpus elebiduna txikiagoa da eta 7,8 milioi esaldi ditu. Orokorrean, corpusak, eta batez ere atal elebakarrak, tamaina onekoak dira hizkuntzak modelatzeko. Hala ere, gure datu-multzoetako domeinua ez dago bertan errepresentatuta eta horrek eragin negatiboa izan lezake erauzitako ezaugarrien doitasunean.

Eragozpen horri aurre egiten saiatu ginen gure datu-multzoetako esaldietatik zuzenean erauzitako informazio linguistikoa baliatuz. Hala ere, azpimarratu nahi dugu lan honen helburua ez dela ingenieritza linguistikoa egitea [26] eta [7] lanetan bezala. Gaztelaniako datuak erauzteko, esaldiak *ixa-pipes* tresnekin [1] prozesatu genituen eta euskarakoak *IxaKat*ekin [20]. Zehazki, maiztasunak erauzi dira, kategoria gramatikalena, ezaugarri morfologikoen eta dependentzia-erlazioena, hau da, esaldiko buruak (aditza izaten da normalean) mendeko elementuekin dituen erlazio gramatikalena, bai iturburuko bai IA proposameneko esaldietarako (10,185 eta 42 ezaugarri gaztelaniarako, hurrenez hurren, eta 10,316 eta 28 euskararako). Aurretiko froga batzuetan ikusi genuenez ezaugarri morfologikoek ez zutela hobekuntza nabaririk ekartzen, horiek baztertzea erabaki genuen.

Beraz, gure esperimentuetan, PE eta PR multzoak ezaugarriok gehituta erabili ditugu. Zehazki, lau datu-multzotan banatu ditugu. Alde batetik, oinarrizko ezaugarriak baliatuta sortutako multzoak bai postedizio multzorako eta bai produktibitate multzorako (PE-17 eta PR-17), eta bestetik, oinarrizko ezaugarriez gain ezaugarri linguistiko gehigarriak ere baliatzen dituztenak (PE-107 eta PR-107).

[4] http://webcorpusak.elhuyar.eus

4. **LAN ESPERIMENTALA**

Atal honetan IAren kalitatea eta postedizio-esfortzua iragartzeko egindako esperimentuak azalduko ditugu.

4.1. **Esperimentuak**

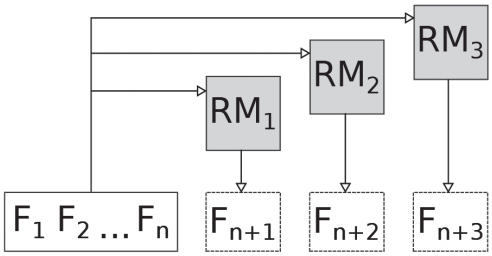
Esperimentuak hiru multzotan banatu ditugu. Lehenengoan, bost erregresio-algoritmok postedizio-esfortzuko zenbait adierazle iragartzeko duten gaitasuna ebaluatu genuen. Adierazle horiek aurreikusten jakiteko, PE multzoak erabili ditugu (PE-17 eta PE-107), nahikoa baita postediziolana biltzen duten datuekin, eta hau PR baino handiagoa delako. Hona hemen aurreikusi nahi ditugun adierazleak:

— **HTER:** Itzultzaile profesionalei IAren proposamenari buruzko kalitate orokorra emateko erabiliko da. Edizio-distantzia neurtzen du eta, beraz, IAren proposamenari beharrezko kalitate mailara iristeko egin beharreko edizioak hartzen ditu kontuan erreferentziako itzulpeneko hitz kopurura normalizatuta.
— **Postedizio-denbora:** Itzultzaile profesional batek IA proposamenak beharrezko kalitate maila izan dezan eraldatzeko behar duen denbora zehazten du.
— **Edizio-motak:** IA proposamenari egin beharreko edizio-mota bakoitzeko informazioa eskeintzen du, hau da, txertatzeak, ezabatzeak, ordezkapenak eta mugimenduak, HTE-Rek kalkulatuta. Nahiz eta HTERek erabiltzen duen hurbilpen matematikoa ez datorren beti bat profesionalen sen linguitikoarekin, uste dugu baliagarria gerta litekeela postedizio-esfortzuaren konplexutasuna agertzeko.
— **Edizio kopuru totala:** IAren proposamenari beharrezko kalitate mailara eramateko egin beharreko edizio kopuru gordinean datza, HTERek kalkulatuta. Edizio-motek informazio zehatzagoa ematen duten bitartean, adierazle honek edizio kopuru globalaren informazio gordina ematen du.

Bigarren esperimentu multzoa iturburuko segmentu bat hutsetik itzuli edo posteditatu egin beharko litzatekeen gomendatzen duten sailkatzaileak eraikitzera eta ebaluatzera bideratuta zegoen. Ataza honetarako komenigarria da bai postedizio eta bai itzulpen lana alderatzen duten datuak izatea. Beraz, kasu honetan, PR multzoak (PR-17 eta PR-107) erabili genituen.

PR multzoetako datu murritzak direla-eta, espero genuen ereduek doitasun baxua izatea. Horregatik, hirugarren esperimentu-multzo bat ere proposatu eta ebaluatu genuen, non PR datu-multzoan postedizio-esfortzuko adierazleen informazioa gehitzen zen; zehazki, postedizio-denbora, HTER eta edizio kopurua. Horretarako, benetako postedizio adierazleekin entre-

natu genituen ereduak, nahiz eta esaldi berrientzako informazio hori ez dagoen eskuragarri. Ondoren, lehenengo esperimentu multzoan eraikitako ereduak erabili genituen esaldi berrientzako, ebaluatu aurretik, ezaugarri gehigarri horiek iragartzeko (ikusi 1. irudia). Postedizio-esfortzuko ezaugarri gehigarri horiekin sortutako datu-multzoak PR-20 eta PR-110 izendatzen dira.



1. irudia. Aurretik entrenatutako erregresio-ereduak erabilita ezaugarriak *n*-tik *n* + 3-rako hedapenaren irudikapena.

Esperimentu guztietan, ikasketan eta ebaluazio prozesuan, 10-aldiko balidazio gurutzatua egiten da datu-multzoekin. Erregresio-ereduen doitasuna (ρ) korrelazio koefizientearekin neurtzen da, zeinek bi aldagairen arteko erlazio linealaren indarra eta norabidea neurtzen duen. Bestetik, sailkatzaileen doitasuna ROC kurbaren azpialdeko azalera erabilita neurtzen da. Esperimentu bakoitza 10 aldiz errepikatzen da eta (μ_ρ eta μ_{ROC}) batezbestekoak eta desbideratze estandarra (σ_ρ eta σ_{ROC}) ematen ditugu. Era berean, binakako t-testak ($p < 0{,}05$) egin ditugu algoritmo onenen emaitzen esanguratsutasun estatistikoa egiaztatzeko. Egiaztatu dugu, halaber, algoritmo bakoitzarentzat, ea ezaugarri linguistikoek ekarpen esanguratsurik egiten duten (†).

4.2. **Erregresio eta sailkapen algoritmoak**

Erregresio algoritmoek, aurrejakintzan oinarrituta, etorkizuneko antzeko egoera baterako balio bat aurreikusten duten bitartean, sailkapen algoritmoek erantzun posibleen artean probableena aurreikusten dute. Ikasketa automatikoko hainbat algoritmo daude erregresio eta sailkapen-ereduak eraikitzeko. Gure esperimentuen helburua gomendio-sistema entrenatzeko algoritmo horiek duten gaitasuna aztertzea denez, bost algoritmo erabilienak aukeratu ditugu bakoitzaren portaera analizatzeko. Zehazki, Hurbileneko k auzokideak (k-NN), Sailkapen eta erregresio zuhaitzak (CART), Sostengu-bektoreen makina (SVR erregresioetarako eta SVM sailkatzaileetarako) eta Geruza anitzeko pertzeptroia (MLP) erabili ditugu

bai erregresio, bai sailkatzaileak sortzeko, bi aukerak ematen baitituzte; eta Erregresio lineala (LR) erregresiorako eta Erregresio logistikoa (LG) sailkapenerako.

Azpimarratu nahi dugu lan hau iragarpenen kalitatea neurtzeko lehen saiakera dela eta etorkizuneko lantzat jotzen dugula algoritmoen doitzea legozkikeen optimizazioekin.

5. KALITATERAKO ETA POSTEDIZIO-ESFORTZURAKO ERREGRESIO-EREDUEN EMAITZAK

Atal honetan postedizio-esfortzua iragartzen duten erregresio-ereduak aurkeztuko ditugu. Adierazle bakoitzaren emaitzak erakutsiko ditugu, hau da, kalitate orokorra, denbora eta edizioak, bakoitza bere aldetik, PE-17 eta PE-107 datu-multzoetarako. Algoritmo onenen emaitzak letra lodiz ematen dira esanguratsutasun estatistikoa adierazteko.

5.1. Kalitate orokorra HTERekin

Hasteko, analiza ditzagun esaldien kalitatea (HTER) iragartzen duten emaitzak PE-17 datu-multzoan (ikusi 3a. taula). Emaitzek erakusten dute k-NN algoritmoarena dela korrelazio koefiziente altuena, 0,71koa. LR eta SVR algoritmoek lortzen dituzte emaitza baxuenak 0,35eko eta 0,32ko koefizienteekin hurrenez hurren. Horrek aditzera ematen du ez LR ez SVR ez direla gai ezaugarrien eta HTER balioen arteko erlazioa modelatzeko. Hori egiaztatzeko, ezaugarrien eta HTER balioen arteko korrelazioa neurtu dugu, eta hiru kasutan izan ezik, korrelazioak 0,1 baino baxuagoak direla ikusi dugu. Izan ere, algoritmo horiek egokiagoak dira erlazio linealak atzemateko. Azterketa azkar batean, SVRko kernel ez-lineala erabilita egindakoan, korrelazioaren batezbestekoaren gorakada ikusi genuen: 0,68±0,04ra PE-17n eta 0,70±0,04ra PE-107n.

3. taula. Erregresio-ereduen emaitzak.

Alg	PE-17		PE-107	
	$\mu\rho$	$\sigma\rho$	$\mu\rho$	$\sigma\rho$
LR	0,3499	0,0399	0,4509†	0,0373
k-NN	**0,7146**	**0,0220**	**0,7144**	**0,0218**
CART	0,6704	0,0367	0,6685	0,0347
SVR	0,3211	0,0415	0,4126†	0,0335
MLP	0,4704	0,0517	0,5870†	0,0456

(a) HTER eredua

Alg	PE-17		PE-107	
	$\mu\rho$	$\sigma\rho$	$\mu\rho$	$\sigma\rho$
LR	0,7137	0,0402	**0,7238**	**0,0372**
k-NN	0,7106	0,0362	0,7131	0,0366
CART	0,7081	0,0392	0,7092	0,0383
SVR	0,7135	0,0405	**0,7265**	**0,0388**
MLP	0,7122	0,0380	0,6955	0,0436

(b) Denbora eredua

Aurreko emaitzak PE-107 datu-multzokoekin konparatuz gero ezaugarri linguistikoek ikasketan duten eragina aztertzeko, PE-17an hobeto zebiltzan algoritmoetan, k-NN eta CART, ezaugarri berrien eraginik ez dela nabari ikusten dugu. Alabaina, beste hiru algoritmoek hoberantz egiten dute esanguratsuki. Badirudi, beraz, korrelazio lineal indartsuagoa dagoela ezaugarrien eta HTER balioen artean. HTER balioekiko 0,1 baino korrelazio altuagoa zuten ezaugarriak 25era igotzeak egiaztazu zuen hori.

5.2. **Postedizio-denbora**

PE-17 datu-multzoko emaitzei erreparatuz gero (ikusi 3b. taula) ikus genezake, HTERekin ez bezala, algoritmo guztiek jokatzen dutela antzera, eta 0,71 inguruko korrelazio koefizientea lortzen dutela. Kasu honetan ezaugarrien eta denboraren arteko korrelazioa 0,2 baino altuagoa da 8 ezaugarrirentzat eta horrek LR eta SVR algoritmoen portaera ona azaltzen du.

PE-107 datu-multzorako emaitzak aztertzen baditugu ikusten dugu ezaugarri linguistiko gehigarrien ekarpenak ez duela garrantzirik. Ikus daiteke LR eta SVR direla gehien baliatzen dituztenak eta emaitzarik altuenak horrela lortzen dituztela. Ezaugarrien eta denboraren arteko erlazioa analizatuz gero, ikus daiteke 0,2 baino korrelazio altuagoa dutela 47 ezaugarrik eta 0,1 baino altuagoa, beste 21ek.

5.3. **Edizio-motak eta kopuru totala**

Ez da argitaratu IA proposamen bat beharrezko kalitate-mailara eraldatzeko edizio-mota desberdinak estimatzen saiatzen den lanik. Avramidisek (2014; 2017) edizio-mota bakoitzerako ereduak entrenatu zituen HTER ereduaren doitasuna hobetzeko saiakera batean. Hala ere, eredu bakoitzaren balioaren inguruan ez zuen hausnartu. Gure emaitzen doitasunari erreparatzen badiegu, ikus genezake gomendio-sistemaren parte izan litezkeela postedizio-esfortzuari buruzko informazioa itzultzaileei emateko.

Azter ditzagun edizio-mota desberdinak PE-17 datu-multzoan lehenik eta behin (ikusi 4. taula). Eredu guztientzat k-NN da algoritmo onena baina edizio-mota bakoitzerako lortzen den doitasun maila nabarmen aldatzen da. Ordezkapenak iragartzen dituen eredua da emaitzarik altuenak lortzen dituena, 0,80ko korrelazio koefizientearekin. Mugimenduek eta txertatzeek ere emaitza onak lortzen dituzte. Hala ere, ezabatzeak iragartzen dituzten ereduek doitasun baxua dute, 0,52ko korrelazioarekin. Horrek erakusten digu algoritmoak ez direla gai ezaugarrien eta adierazle horren arteko erlazioa antzemateko.

4. taula. Edizio-mota bakoitzarentzako eta edizio totalerako erregresio algoritmoen emaitzak.

		Txertatzeak		Ezabatzeak		Ordezkapenak		Mugimenduak		Edizio Totalak	
Datu-multzoa	Alg	μ_ρ	σ_ρ	μ_ρ	σ_ρ	μ_ρ	σ_ρ	μ_ρ	σ_ρ	μ_ρ	σ_ρ
PE-17	LR	0,5685	0,0427	0,4537	0,0421	0,7336	0,0180	0,6167	0,0266	0,8029	0,0180
	k-NN	**0,7011**	**0,0687**	**0,5214**	**0,0435**	**0,8035**	**0,0182**	**0,7422**	**0,0238**	**0,8660**	**0,0168**
	CART	0,6556	0,0930	0,4896	0,0459	0,7876	0,0187	0,7164	0,0252	0,8550	0,0176
	SVR	0,5625	0,0273	0,4517	0,0415	0,7325	0,0179	0,6147	0,0274	0,8020	0,0187
	MLP	0,6633	0,0740	0,4731	0,0488	0,7404	0,0205	0,6344	0,0268	0,8125	0,0198
PE-107	LR	0,6167	0,0610	0,4840	0,0394	0,7566	0,0187	0,6710†	0,0257	0,8247	0,0175
	k-NN	**0,7010**	**0,0688**	**0,5214**	**0,0435**	**0,8035**	**0,0182**	**0,7423**	**0,0238**	**0,8660**	**0,0167**
	CART	0,6571	0,0893	0,4874	0,0452	0,7872	0,0190	0,7184	0,0249	0,8558	0,0182
	SVR	0,5750	0,0306	0,4723	0,0373	0,7505	0,0190	0,6567	0,0274	0,8193	0,0179
	MLP	0,6664	0,0758	0,4746	0,0466	0,7364	0,0332	0,6674	0,0370	0,8210	0,0268

Edizio-mota desberdinek postedizio lana estimatzerakoan, itzultzaileentzako izan dezaketen balioari erreparatzea merezi du. Txertatzeek eta ordezkapenek lan neketsua dakarte, non itzultzaileek faltan dagoen informazioa gehitu edo IA proposamenean gaizki dagoena ordezkatu behar duten. Mugimenduak esfortzu gutxiagoko edizioak direla argudia genezake, non informazio egokia hortxe dagoen, baina toki desegokian. Hiru edizio-mota horiek 0,70etik gorako korrelazio koefizienteak lortzen dituzte eta nolabaiteko fidagarritasunarekin erakuts geniezazkieke itzultzaileei optimizazio froga batzuen ondoren. Ezabatzeak, aldiz, esfortzu gutxiko edizio-motatzat har ditzakegu, itzultzaileek soberan dauden elementuak baztertzen baitituzte. Beraz, baliteke postedizio-lan neketsuena ez izatea. Aztertzeke dago itzultzaileentzat mota horietako zeintzuk diren informazio baliagarriena daukaten edizioak postedizio-esfortzua kalkulatzerakoan.

Edizio kopuru totalari erreparatzen badiogu, mota edozein dela, ikusten dugu iragarpen ereduek bereziki ondo jokatzen dutela. Berriz ere k-NN da algoritmorik onena 0,86ko korrelazio koefizientearekin, baina bost algoritmoen korrelazio balioak daude 0,80tik gora.

Aurreko erregresio-ereduetan gertatzen zen bezala, hemen ere ikusten dugu PE-107 datu-multzoko ezaugarri linguistiko gehigarriek ez dutela ia inongo ekarpenik egiten ikasketa prozesuan eta behin ere ez dituztela algoritmo onenen emaitzak hobetzen.

6. EDIZIO-METODORAKO SAILKATZAILEEN EMAITZAK

Atal honetan, itzultzea edo posteditatzea, metodorik egokiena zein den aurreikusten duten sailkatzaileen emaitzak erakusten ditugu. Emaitzak ematen ditugu, ezaugarri linguistikoekin zein gabe. Era berean, hasierako emaitza baxuei aurre egiteko asmoz, postedizio-esfortzuko adierazleak ezaugarritzat hartzeak duen eragina erakusten duten sailkatzaileen emaitzak ere ematen ditugu. Eredu horiek PR datu-multzoarekin eraiki dira, nahiz eta esaldi kopuru mugatua izan, itzulpen lana eta postedizio lana alderatzeko aukera ematen digun datu-multzoa delako.

6.1. Oinarrizko sailkapen-ereduak

ehenik eta behin, oinarrizko sailkapen-ereduen emaitzak aurkeztuko ditugu. Eredu hauek produktibitatearen inguruko datu-multzoan eskuragarri genituen segmentu guztiekin entrenatu genituen.

Bi etiketako (2L) atazaren emaitzak aztertuko ditugu lehenengo (ikusi 5. taula). Bai PR-17 eta bai PR-107 multzoentzako algoritmo guztiek lortzen dituzte emaitza baxuak μ_{ROC} 0,60 azpitik, SVM atzean geratzen delarik. Hala ere, PR-107 datu-multzoari erreparatzen badiogu, ikusten dugu ezaugarri linguistiko gehigarriei esker SVM beste algoritmoen parera iristen dela.

Azter ditzagun hiru etiketentzako (3L) emaitzak (ikusi 5. taula). Kasu honetan ez dago desberdintasun estatistikoki esanguratsurik algoritmoen artean. Aurrekoan bezala, ezaugarri linguistikoek SVRrentzako baino ez dute ekarpenik egiten.

5. taula. Bi etiketarentzako sailkapen algoritmoen emaitzak.

Alg	PR-17 (2E)		PR-107 (2E)		PR-17 (3E)		PR-107 (3E)		PR-17 (5E)		PR-107 (5E)	
	μ_{ROC}	σ_{ROC}	μ_{ROC}	σ_{ROC}	μ_{ROC}	σ_{ROC}	μ_{ROC}	σ_{ROC}	μ_{ROC}	σ_{ROC}	μ_{ROC}	σ_{ROC}
LG	**0,56**	**0,15**	0,60	0,15	0,53	0,17	0,50	0,17	0,57†	0,25	0,40	0,24
k-NN	**0,58**	**0,11**	0,56	0,11	0,56	0,11	0,55	0,13	0,57	0,15	**0,56**	**0,15**
CART	**0,51**	**0,11**	0,49	0,09	0,50	0,11	0,49	0,10	0,54	0,13	**0,55**	**0,13**
SVM	0,50	0,02	0,57†	0,11	0,48	0,06	0,57†	0,13	0,57	0,22	**0,52**	**0,22**
MLP	0,59	0,13	0,58	0,17	0,58	0,14	0,56	0,16	0,69	0,22	**0,55**	**0,22**

Bost etiketentzako (5L) ere emaitzen joera baxua dela ikus dezakegu (ikusi 5. taula). Hala ere, kasu honetan, ezaugarri linguistiko gehigarriek ez dute inolako ekarpenik egiten. Izan ere, estatistikoki esanguratsua den bakarra LGrentzako atzerapausoa da.

Orokorrean, ondoriozta dezakegu sailkapen-ereduen eraginkortasuna testuinguru erreal batean erabiltzeko beharrezkoa litzatekeen zehaztasuna

izatetik urruti dagoela. Nahiz eta algoritmoen konfigurazioa aldatuz hobetzeko aukerak egon, uste dugu ezaugarriek ez dietela informazio egokia ematen algoritmoei sailkapen atazarako.

6.2. **Postedizio-esfortzuko adierazleak darabiltzaten sailkapen-ereduak**

Sailkatzaileen errendimendua hobetzeko asmoz, ikasketa prozesuan postedizio-esfortzuko adierazleak ezaugarri bezala erabiltzea proposatzen dugu, ezaugarri orokorrez eta linguistikoez gain. Ustea da adierazle horiek zehatzago islatzen dituztela itzultzaile batek edizio-metodo bat bestearen gainetik aukeratzeko arrazoiak. Agian, horrela, algoritmoek ikasteko duten informazioa informagarriagoa izango da. Horretarako, aurreko sailkatzaileek HTER, edizio kopurua eta denbora errealak ezaugarri bezala erabilita emaitza hobeak lortzen dituzten analizatu dugu lehenengo. Bigarrenik, hiru ezaugarri horiek itzulpena bukatutakoan baino ez daudenez eskuragarri, sailkatzaile berriak erabili ditugu testeko esaldiei aurreikusitako postedizio-adierazleak gehituta (ikusi 4.1. atala).

6. taulan laburtu ditugu postedizio-esfortzuko hiru adierazleak gehituta lortutako emaitzak. μ_{ROC} en bidez ematen dira eta entrenamendu multzoaren batezbestekoari dagozkio. Espazio-mugak direla eta ikasketa prozesuko desbideratze estandarra ez dugu aipatu. Ikus dezakegu PR-20 multzoan ereduen doitasuna nahiko ona eta bikaina tartean dagoela, etiketa kopurua dena dela. Hobekuntza nabarmena da oinarrizko sailkatzaileekin alderatuta eta HTER, denbora eta edizio kopurua ezaugarritzat erabiltzearen onura erakusten du. k-NN, CART eta MLP dira emaitza onenak lortzen dituzten algoritmoak multzo guztietarako eta LG eta SVM baxuenak PR-20 multzoan. Hala ere, erregresio esperimentuetan gertatu den bezala, ikusten dugu ezaugarri linguistikoak gehitutakoan (PR-110), LGren eta SVRren eraginkortasuna hobetu egiten dela. Interesgarria da ikustea k-NNrentzako ere eragin positiboa dutela.

6. taula. Sailkapen algoritmoen emaitzak postedizio-esfortzuko ezaugarriak erabilita.

Alg	PR-20			PR-110		
	2L	3L	5L	2L	3L	5L
LG	0,76	0,74	0,80	**0,99**†	**1,00**†	**0,99**†
k-NN	**0,99**	**0,99**	**0,98**	**1,00**†	**1,00**†	**1,00**†
CART	**0,98**	**0,99**	**1,00**	**0,99**	**1,00**	**1,00**
SVM	0,64	0,63	0,77	0,91†	0,93†	0,96†
MLP	0,97	0,96	0,95	1,00	0,99	0,96

(a) μ_{ROC}-ren arabera emaitzak.

Alg	PR-20			PR-110		
	2E	3E	5E	2E	3E	5E
LG	0,66	0,69	0,75	0,70	0,73	0,75
k-NN	**0,89**	0,89	0,90	**0,89**	**0,99**	0,90
CART	0,88	**0,90**	0,89	**0,89**	0,90	0,90
SVM	0,56	0,55	0,60	0,83	0,85	0,87
MLP	0,88	0,83	0,92	0,88	0,90	0,99

(b) T_{ROC}-ren arabera emaitzak.

Postedizio-esfortzuko adierazleak ezaugarritzat erabilita lortutako emaitza onek bultzatuta, produkzio egoera erreal bateko baldintzetan probatu nahi izan genuen hurbilpen hau. Produktibitateko datu-multzoa entrenamendu eta ebaluazio multzoetan banatu dugu. 153 iturburu esaldietatik, ausaz %80 entrenamendu multzoan eta %20 ebaluazio multzoan sartu ditugu. Entrenamendu multzoak esaldi bakoitzerako eskuragarri daukagun informazio guztia dauka. Ebaluazio multzoan, baina, esaldi bakoitzaren instantzia bakarra erabili dugu aurreko atalean sortutako erregresio-eredu onenek aurreikusitako HTER, denbora eta edizio-kopuru totalari dagozkion ezaugarriak gehituta, hasierako ezaugarriez gain (PR-20 eta PR-110 ditugu 3 ezaugarri berriekin). Emaitzak, T_{ROC}en arabera emanak, hau da, ebaluazio multzoari ikasitako ereduak aplikatu ostean lortutako ROC balioa, 6b. taulan ematen dira.

Esperimentu honetan eredu onenen emaitzak on eta bikain tartekoak dira bai PR-20 zein PR-110 multzoetarako eta etiketa-multzo guztietarako. Kontuan izan gure aurreikuspenek errore tarte bat dakartela erregresio-ereduetatik, eta hala ere, doitasun maila altua dutela. PR-20 multzoan, MLP da algoritmo eraginkorrena eta LG eta SVM dira emaitza baxuenak lortzen dituztenak. PR-110 multzoan ere algoritmo berdinek darraite emaitza onenak lortzen baina k-NN eta CART bereziki eraginkorrak dira 2 eta 3 etiketetako multzoetan eta MLP 5 etiketetakoan. Garrantzitsua da SVMk datu-multzo honetan lortzen duen hobekuntza azpimarratzea. Orokorrean esan dezakegu sailkapen-ereduen emaitzak nahiko onak direla eta testuinguru erreal batean erabiltzeko doitasun maila egokia erakusten dutela. Are gehiago, espero dugu sailkatzaileen doitzearen ondorioz eta erregresio-ereduen doitzearen ondorioz emaitza hobeak lortzea.

7. **ONDORIOAK**

Artikulu honetan IAren kalitateaz harago doazen estimazio eredu batzuk entrenatzeko aukera aztertu dugu itzultzaile profesionalei esaldi berri bat itzultzea edo posteditatzea eraginkorragoa den erabakitzen lagunduko dien sistema bat eraikitzeko. Zehazki, analizatu dugu ea emaitza onargarriak lortzerik genuen gaztelania-euskara bikoterako, zeinentzat IA kalitatea baxua den, eta beraz, ez dagoen profesionalki hedatua. Esperimentuak postedizio eta produktibitate ataza batzuetatik bildutako datu-multzo txiki batzuk baliatuta egin ditugu.

HTER, denbora eta edizio-motak eta kopurua aurreikusten dituzten erregresio-ereduak entrenatu ditugu postedizio-esfortzuaren adierazle bezala. Erakutsi dugu korrelazio koefiziente nahiko altuak, 0,70etik gorakoak, lortzen ditugula ia adierazle guztientzat. Edizio-kopuru totalak dirudi aurreikusteko errazena 0,86ko korrelazio koefizientearekin, eta edizio-mota

desberdinak, zailenak, bereziki ezabatzeak. Emaitzek erakusten dute orobat kategoria gramatikalen eta dependentzia erlazioen maiztasuna ezaugarri bezala erabiltzeak ez dituela gure eredu gehienak hobetzen. k-NN izan da algoritmo eraginkorrena, HTER eta edizioak barne hartzen zituzten ereduak entrenatzeko batez ere. Denbora aurreikusteko LR eta SVR izan dira eraginkorrenak, ezaugarri linguistikoetatik hobekuntza lortuta.

Sailkatzaile bat ere entrenatu dugu, edizio-metodo eraginkorrena zein den gomendatzen duena. Oinarrizko ereduak emaitza baxuak lortzen zituen arren, 0.58 inguruan, erakutsi dugu postedizio-esfortzuko adierazleak ezaugarritzat erabilita emaitzak nabarmen hobetzen direla, 0.99 ingurura. Informazio hori esaldi berrientzat eskuragarri ez dagoenez, aurretik sortutako erregresio-ereduak erabili ditugu ezaugarri horiek esaldi berrietan aurreikusteko. k-NN, CART eta MLP izan dira behin eta berriz emaitza oneak lortu dituzten algoritmoak datu-multzo guztietarako.

Ikasketarako erabilitako datu-multzoen tamaina mugatuak erakusten du, alde batetik, domeinu zehatz baterako eredu fidagarriak lortzeko ez dela beharrezkoa datu-multzo handi bat izatea. Hala ere, eredu horien eramangarritasuna, hau da, beste datu-multzo edo testu baterako duen baliagarritasuna frogatzeke dago.

Lortutako emaitza onak medio, hurrengo pausoa ereduon parametro desberdinak erabiliz eraginkortasun optimoa lortzea eta datu-multzo berrietan probatzea litzateke. Gure helburua da ikertzea ea HTER, postedizio-denbora eta edizio-motak zein punturaino izan daitezkeen mesedegarri itzultzaile profesionalentzat.

8. **ESKER ONAK**

Ikerketa hau partzialki finantzatua izan da Espainiako gobernuko TADEEP proiektuaren (*TIN2015-70214-P*) eta Eusko Jaurlaritzako QUALES proiektuaren (*KK-2017/00094*) bitartez.

9. **ERREFERENTZIAK**

[1] AGERRI, RODRIGO, JOSU BERMUDEZ eta GERMAN RIGAU. 2014. «IXA pipeline: Efficient and Ready to Use Multilingual NLP tools». *LREC2014, 9th Language Resources and Evaluation Conference*, Reykjavik, Iceland. 3823-3828.

[2] ARANBERRI, NORA, GORKA LABAKA, ARANTZA DIAZ DE ILARRAZA eta KEPA SARASOLA. 2014. «Comparison of postediting productivity between professional translators and lay users». *Third Workshop on postediting Technology and Practice*, Vancouver, Canada. 20-33.

[3] ARANBERRI, NORA. 2017. «What Do Professional Translators Do when postediting for the First Time? First Insight into the Spanish-Basque Language Pair». *HERMES-Journal of Language and Communication in Business*, (56):89-110.

[4] ARANBERRI, NORA, GORKA LABAKA, ARANTZA DIAZ DE ILARRAZA eta KEPA SARASOLA. 2017. «Ebaluatoia: crowd evaluation for EnglishBasque machine translation». *Language Resources and Evaluation*, 51(4):1053-1084.

[5] AVRAMIDIS, ELEFTHERIOS. 2017. «Sentence-level quality estimation by predicting HTER as a multi-component metric». *WMT-2017, Conference on Machine Translation, Shared Task Papers*, Copenhagen, Denmark. 534-539.

[6] AVRAMIDIS, ELEFTHERIOS. 2014. «Efforts on Machine Learning over Human-mediated Translation Edit Rate». *WMT-2014, Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland. 302-306.

[7] AVRAMIDIS, ELEFTHERIOS. 2012. «Quality estimation for machine translation output using linguistic analysis and decoding features». *WMT-2012, Seventh workshop on statistical machine translation*, Montreal, Canada. 84-90.

[8] AVRAMIDIS, ELEFTHERIOS, MAJA POPOVIC, DAVID VILAR eta ALJOSCHA BURCHARDT. 2011. «Evaluate with confidence estimation: machine ranking of translation outputs using grammatical features». *WMT-2011, Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland. 65-70.

[9] BECK, DANIEL, KASHIF SHAH eta LUCIA SPECIA. 2014. «SHEF-Lite 2.0: Sparse Multi-task Gaussian Processes for Translation Quality Estimation». *WMT-2014, Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland. 307-312.

[10] BICICI, ERGUN eta ANDY WAY. 2014. «Referential Translation Machines for Predicting Translation Quality». *WMT-2014, Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland. 313-321.

[11] BLATZ, JOHN, ERIN FITZGERALD, GEORGE FOSTER, SIMONA GANDRABUR, CYRIL GOUTTE, ALEX KULESZA, ALBERTO SANCHIS eta NICOLA UEFFING. 2004. «Confidence Estimation for Machine Translation». *ACL-2004, 42th Annual Meeting of the Association for Computational Linguistics*, Geneva, Switzerland. 315-321.

[12] DODDINGTON, GEORGE. 2002. «Automatic Evaluation of Machine Translation Quality using N-gram Co-ocurrence Statistics». *HLT-2002, 2nd Human Language Technologyies Conference*, San Diego, California. 128-132.

[13] ETCHEGOYHEN, THIERRY eta ANDONI AZPEITIA. 2016. «A Portable Method for Parallel and Comparable Document Alignment». *Baltic Journal of Modern Computing*, 4(2):243255.

[14] GARMENDIA, LIERNI, NAROA LASARTE eta MAIALEN PINAR. 2017. «Situación actual y viabilidad de la TA en euskera: posedición y análisis de los resultados de un motor de TABR español-euskara». *Master Thesis*, Universitat Autònoma de Barcelona.

[15] HARDMEIER, CHRISTIAN, JOAKIM NIVRE eta JORG TIEDEMANN. 2012. «Tree kernels for machine translation quality estimation». *WMT-2012, Seventh Workshop on Statistical Machine Translation*, Montreal, Canada. 109-113.

[16] HE, YIFAN, YANJUN MA, JOSEF VAN GENABITH eta ANDY WAY. 2010. «Bridging SMT and TM with Translation Recommendation». *ACL-2010, 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden. 622-630.

[17] MOREAU, ERWAN eta CARL VOGEL. 2012. «Quality estimation: an experimental study using unsupervised similarity measures». *WMT-2012, Seventh Workshop on Statistical Machine Translation*, Montreal, Canada. 120-126.

[18] NIEEN, SONJA, FRANZ JOSEF OCH, GREGOR LEUSCH eta HERMANN NEY. 2002. «An evaluation tool for machine translation: Fast evaluation for MT research». *LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece. 39-45.

[19] OCH, FRANZ JOSEF eta HERMANN NEY. 2003. «A Systematic Comparison of Various Statistical Alignment Model». *Computational Linguistics*, 29(1):19-51.

[20] OTEGI, ARANTXA, NEREA EZEIZA, IAKES GOENAGA eta GORKA LABAKA. 2016. «A Modular Chain of NLP Tools for Basque». *TSD 2016, 19th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic. *Lecture Notes in Artificial Intelligence*, 9924:93-100.

[21] SNOVER, MATTHEW, BONNIE DORR, RICHARD SCHWARTZ, LINNEA MICCIULLA eta JOHN MAKHOUL. 2006. «A Study of Translation Edit Rate with Targeted Human Annotation». *AMTA-2006, 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts. 223-231.

[22] SPECIA, LUCIA, MARCO TURCHI, NICOLA CANCEDDA, MARC DYMETMAN eta NELLO CRISTIANINI. 2009. «Estimating the Sentence-Level Quality of Machine Translation Systems». *EAMT-2009, 13th Conference of the European Association for Machine Translation*, Barcelona, Spain. 28-37.

[23] SPECIA, LUCIA eta ATEFEH FARZINDAR. 2010. «Estimating Machine Translation postediting Effort with HTER». *AMTA-2010, Workshop Bringing MT to the User: MT Research and the Translation Industry*, Denver, Colorado. 33-41.

[24] SPECIA, LUCIA, NAJEH HAJLAOUI, CATALINA HALLETT eta WILKER AZIZ. 2011. «Predicting machine translation adequacy». *Machine Translation Summit XIII*, Xiamen, China. 19-23.

[25] SPECIA, LUCIA, GUSTAVO HENRIQUE PAETZOLD eta CAROLINA SCARTON. 2015. «Multi-level Translation Quality Prediction with QuEst++». *ACL-IJCNLP 2015 System Demonstrations*, Beijing, China. 115-120.

[26] SPECIA, LUCIA eta MARIANO FELICE. 2012. «Linguistic features for quality estimation». *WMT-2012, Seventh workshop on statistical machine translation*, Montreal, Canada. 96-103.

[27] SPECIA, LUCIA. 2011. «Exploiting objective annotations for measuring translation postediting effort». *EAMT-2011, 15th Conference of the European Association for Machine Translation*, Leuven, Belgium. 73-80.