Ekaia, 2021, 40, 315-328
https://doi.org/10.1387/ekaia.22123

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Euskarazko on-line artikuluetan aipatutako izendun entitate nabarmenen identifikazioa denbora errealean

(Real-Time Identification of Named Entities in Online Basque-language Media)

Joseba Fernández de Landa, Rodrigo Agerri*

HiTZ Zentroa-Ixa, Euskal Herriko Unibertsitatea (UPV/EHU)

LABURPENA: Lan honen helburu nagusia, hedabideetako euskarazko edukian aipatzen diren izendun entitate nabarmenen identifikazioa da, identifikazioa denbora errealean eginez. Horretarako, euskaraz argitaratutako albisteetatik izendun entitateak automatikoki jaso eta etiketatzeko sistema garatu da, artearen egoerako Ikasketa Sakoneko ereduak erabiliz. Izendun entitateen identifikadoreari esker, denbora errealean jasotako albisteetako izendun entitateak etengabe identifikatu eta jasotzen dira, erregistro bat osatuz. Bukatzeko, identifikatutako izendun entitate nabarmenak astero publikatzen dira Wikipediako orri batean, Euskarazko Wikipedian artikulurik ez daukaten entitate nabarmenak erakusteko asmoz.

HITZ GAKOAK: Euskarazko Hedabideak, Izendun Entitateen Erauzketa, Hizkuntzaren Prozesamendua.

ABSTRACT: *Names referring to people, institutions, or places may be defined as named entities. Extracting named entities from news texts can help to identify the most commented topics talked about in news media. The main objective of this work is to identify in real-time those named entities that are most commented upon on Basque-language online media. In order to do so, we develop a system to automatically collect and annotate the named entities appearing in news written in Basque language. The annotation of named entities is performed using state-of-the-art deep learning models. Finally, the most frequent identified entities are published weekly in a Wikipedia page to display which entities do not currently have an article in the Basque Wikipedia.*

KEYWORDS: Basque-language Media, Named Entity Recognition, Natural Language Processing.

**** Harremanetan jartzeko / Corresponding author:*** Joseba Fernández de Landa. HITZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU, 20018, Donostia. – joseba.fernandezdelanda@ehu.eus – https://orcid.org/0000-0001-6067-3571

Nola aipatu / How to cite: Fernández de Landa, Joseba; Agerri, Rodrigo (2021). «Euskarazko on-line artikuluetan aipatutako izendun entitate nabarmenen identifikazioa denbora errealean». *Ekaia*, 40, 2021, 315-328. (https://doi.org/10.1387/ekaia.22123).

Jasotze-data: 2020,urriak 9; Onartze-data: 2021, urtarrilak 13.

ISSN 0214-9753 - eISSN 2444-3581 / © 2021 UPV/EHU

1. **SARRERA**

Hizkuntzaren Prozesamenduko (HP) atazetako bat informazio-erauzketa dugu, testuetatik nahi den informazioa lortzeko aukera ematen duena. Horrela, informazio erauzketaren hastapenetan, ataza nagusia egitura-gabeko testuetatik egituratutako informazioa erauztean zetzan, hau da, testuan oinarritutako datu multzo bat egituratzea, nahi den informazioa lortzeko helburuarekin. Ataza hori bideratzen ari zirelarik, konturatu ziren izendun entitateak oso garrantzitsuak zirela informazio unitateak identifikatzeko, hala nola, pertsona, erakunde edota lekuen izenak [1]. Pertsona, erakunde edota kokalekuak definitzeko erabiltzen diren hitz multzo bezala definitu daitezke izendun entitateak. Termino hauen identifikazioaz automatikoki arduratzen den ataza Izendun Entitateen Identifikazio (ingelesez, *Named Entity Recognition*-NER) izenez ezagutzen da, testuetatik pertsona, erakunde edota tokien izenak automatikoki identifikatzeko gai izanik. Ataza horretarako, sekuentzia-etiketatze izenez ezagututako teknika baliatuz, esaldi batean azaltzen diren izenak etiketatzen dira. Esaterako, 1. taulan ikus dezakegu sekuentzia etiketatzea nola egiten den izendun entitateak identifikatzeko. Izendun entitatearen lehen tokena hasiera bezala markatzen da (B) eta barneko (I) beste token guztiak ere, soberan dauden tokenak markatu gabe utziz (O). Honi esker, token batez baino gehiagoz osatutako izendun entitateak identifika daitezke. Bestalde, pertsona (PER), erakunde (ORG) eta tokiak (LOC) ere identifikatzen dira, izendun entitate bakoitzaren kategoria iradokiz.

Hala ere, euskarazko izendun entitateen identifikazioa burutzeko, lematizazioaren beharra dugu. Lematizazioaren bidez hitzaren forma kanonikoa lortzen da, hau da, hitzaren forma flexionatu guztien ordezkaria edo hiztegi batean topatuko genukeen hitza. Euskararen inguruko hizkuntza gehienek (gaztelania, frantsesa, ingelesa, galiziera, katalana...) ez dituzte izen bereziak deklinatzen, euskarak ordea bai. Arrazoi horregatik, beharrezkoa da ataza honetarako izendun entitateen forma kanonikoa lortzea, *Trumpek, Trumpen, Trumpengana* bezalako flexioen atzean *Trump* izendun entitatea identifikatzea baita gure asmoa. 1. taulan ikusi dezakegun moduan, *Trumpek, Washingtonetik, AEBko* eta *Gobernuak* izendun entitateetatik *Trump, Washington, AEB* eta *Gobernu* lortzea interesatzen zaigu.

1. taula. Izendun entitateen sekuentzia etiketatzea, adibidea.

B-PER,	B-LOC,	O	O	O	B-ORG	I-ORG	O	O.
Trump**ek**,	Washington**etik**,	gogor	kritikatu	du	AEB**ko**	Gobernu**ak**	egindako	lana.
Trump	Washington	gogor	kritikatu	*edun	AEB	Gobernu	egin	lan

Izendun entitateek informazio ugari eskaintzen dute, hitz-multzo sinple batekin testuinguru zabal eta aberatsa barnebiltzen baitute. Informazio ugari barnebiltzen duten hitz hauek garrantzitsuak dira testu mota edota gaia identifikatzeko orduan. Testu bakoitzean azaltzen diren izendun entitateak erlazionatzean, aukera egon daiteke testuak sailkatzeko, izendun entitateen erlazioak ezagutzeko, izendun entitate aipatuenak ezagutzeko eta abar. Era honetan, izendun entitateen identifikazio automatikoa ataza garrantzitsutzat jotzen dugu, testuan oinarritutako analisiari esker, ikerkuntza sozialerako baliagarria den metodologia berri bati ateak irekitzeko aukera izan daitekeelako.

Lanaren helburua, izendun entitate aipatuenen artean berrienak lortzea denez, euskal hedabideen euskarazko artikuluetako testuak baliatuko dira. Euskal hedabideak oso aktiboak dira euskarazko edukia argitaratzeko orduan, datu-iturri emaritsu eta jarraitu bat izanik euskarazko edukia aztertu nahi duen edonorentzat. Euskaraz argitaratzen duten hainbat hedabide digitaletatik albisteak etengabe jasoko dira, 8 hedabide ezberdinetatik, ikerketarako beharrezkoak diren testuak bilduz.

Testuan oinarritutako datu multzoa prozesatzeko, artearen egoerako teknologiak aplikatuko dira, euskararako ere garatuta baitaude teknologia hauek. Teknologia hauetako aipatuenen artean sare neuronal artifizialak ditugu, HPko hainbat atazatan hobekuntza handiak ekarriz, baita euskarazko edukia prozesatzeko orduan ere. Euskararen arlora ekarrita, zeresan handia eman duen Itzulpen Automatiko Neuronala daukagu, erregelan oinarritutako itzultzaile zein itzultzaile estatistikoak baino emaitza hobeak lortzen dituena [2]. Beste arlo batzuetan ere hobekuntza handiak ekarri ditu neurona-sareetan oinarritutako teknologiak, hala nola, gaien zein sentimenduen sailkapenean, kategoria morfosintaktikoa iradokitzean (POS-tagging) eta baita izendun entitateen identifikazioan ere [3].

Aldi berean, aipatutako ataza hauek guztiak aurrera eramateko, corpus bereziak behar dira, Ikasketa Automatikoaren parte diren neurona-sare artifizial hauek, entrenamendu zein ebaluaketarako datuak behar baitituzte. Datu horiek hizkuntzarekiko menpekoak dira eta kasu gehienetan hizkuntza gutxituetarako datuak topatzea oso zaila da. Euskararen kasua berezia da, baliabide ugari baitaude hizkuntzaren prozesamenduko hainbat ataza aurrera eramateko. Lan honetarako, Ixa taldearen baitan garatutako hainbat lanabes erabili dira [3, 4], terminoen lematizazioa zein izendun entitateen identifikazioa burutzeko.

Euskaraz idatzitako artikuluetan pertsonaia aipatuenak eta berrienak identifikatzeko asmoa duen lan hau aurrera eramateko, hainbat pauso eman dira. Lanaren edukian zentratuz, datu-bilketa zein prozesaketa konbinatu dira lan honetan, hasieratik bukaerara arteko pauso guztiak guk geuk garatu eta kontrolatuz. Datu-bilketaren atalean hedabide digitaletan euskaraz

argitaratzen diren zein artikulu jaso eta nola egiten den azalduko da. Behin lanerako datuak jasota, datu hauen prozesaketa nola egin azalduko da hurrengo atalean, izendun entitateak nola identifikatu eta nabarmenenak nola aukeratu diren azalduz. Emaitzen atalean, lan honen emaitzekin bistaraketa zein on-line publikazio dinamikoa erakutsiko da. Azkenik, lan honen ondorioak aurkeztuko dira, eta lanarekin lortutako helburuak zein etorkizunerako ireki daitezkeen aukerak azalduko dira.

Lan honen ekarpenen artean, sarean argitaratzen diren euskarazko artikuluen denbora errealeko monitorizazioa daukagu. Bestalde, jaso diren euskarazko albiste horietan izendun entitateak identifikatzeko, Ikasketa Sakonean (*Deep Learning*) oinarritutako teknikak baliatuko dira, puntako aurrerapen teknologikoak euskal hizkuntzarako egokituz eta lan honetan aplikatuz [3]. Horrez gain, jasotako emaitzak modu publikoan argitaratu eta denbora errealean eguneratuak izango dira Wikipediako orrialde bat baliatuz[1].

2. **METODOLOGIA**

Argitaratutako euskarazko artikuluetan izendun entitateak identifikatu eta hauetatik nabarmenak direnak aukeratzeko bi prozesu ezberdin konbinatuko dira, alde batetik datu-bilketa propio bat abiaraziz, eta bestetik, bildutako datu horien prozesaketa gauzatu datu-meatzaritza erabiliz:

— **Datu-bilketa**: atal honetan, lehenik eta behin, euskarazko hedabideen eskuzko identifikazioa egungo da, eta lan honetarako egokienak diren hedabideak aukeratuko dira. Behin hedabidek identifikatuta, hauen albiste bakoitzetik datuak jasoko dira denbora errealean, MSM [5] programa baliatuz.
— **Datuen prozesaketa**: behin lanerako datu multzoa prest egonda, albiste bakoitzaren izendun entitateen identifikazioa eta erauzketa eta izendun entitate berrienen aukeraketa egingo da; pauso hau ere denbora errealean burutuko da:

- *Izendun entitateen identifikazioa eta erauzketa*: Izendun entitateen identifikazioa egiteko bi sistema ezberdin frogatu eta ebaluatu dira lan honetan. Alde batetik IXA *pipes* [4] eta, bestetik, Flair-BMC [3]. Horrela, albiste bakoitzaren edukia tokenizatzeko eta izendun entitateen identifikazioa burutzeko orduan metodo egokiena aurkeztuko da. Izendun entitateen errekonozimenduaren atazarako sistema onena zein den erabaki ostean, albiste bakoitzaren izendun entitateak gordetzen dira.

[1] https://eu.wikipedia.org/wiki/Wikiproiektu:Euskarazko_albisteetako_Izen_Entitateak

- *Lematizazioa*: izendun entitateen forma kanonikoa lortzeko asmoarekin lematizazioa erabili da. Euskararen aberastasun morfologikoa dela medio, izendun entitateen forma flexionatuei deklinabidea kendu eta lema erauzteko baliagarria da. Euskarazko lematizatzaileak garatzeko, IXA *pipes* eta Flair-BMC entrenatu genituen, UD 2.2 corpusa erabiliz [6].
- *Izendun entitate berrienen aukeraketa*: izendun entitate berrienen aukeraketa egiteko, termino maiztasun-alderantzizko dokumentu-maiztasuna (TF-IDF) erabiliko da. Izendun entitate guztien artean soilik nabarmenak aukeratu ahalko dira, oso ohikoak edo gutxi aipatuak direnak baztertuz.

3. **DATU-BILKETA**

Esan bezala, ikerketaren lehen pausoa izango da euskaraz argitaratzen duten hedabide digitaletatik datuak biltzea. Horretarako, beren edukia edo edukiaren zati bat euskaraz argitaratzen duten hedabide digitaletatik, euskarazko albisteak jasoko dira. Lehenik eta behin, euskaraz edukia publikatzen duten hedabideen artean, eskuzko aukeraketa burutu da. Aukeraketarako hedabideen euskarazko edukiaren kantitatea zein irakurle kopurua aintzat hartzen saiatu da. Era honetan, *Berria*, *Argia*, *Zuzeu*, *Sustatu*, *EiTB*, *Naiz*, *Grupo Noticias* eta *Vocento* hedabideak entzunak izango dira.

Aipatutako hedabideetatik albisteen edukia jasotzeko, MSM crawlerra [5] erabili da. Honi esker, etengabeko entzuketa bat egiten da hedabideen RSS loturetatik albisteak jasotzeko. Jasoketan, hedabide ezberdinetatik berri-jarioa jaso eta garbitzen da. Albistearekin batera, analisi sozialerako baliagarriak diren hainbat meta-datu jasoko dira. 2. taulan ikus daitekeen moduan, jasotako albiste bakoitzetik hedabidea, data, hizkuntza, titularra, edukia eta lotura erabiltzen dira. Garatutako datu-bilketa sistemari esker, astero 1.000 artikulu inguru jasotzen dira euskaraz, sistema etengabe hornituz denbora errealean.

2. taula. Jasotako albisteen datuen egituraketaren adibidea.

hedabidea	berria
data	2020-02-05
hizkuntza	eu
titularra	Bizitza xumeago, egiazkoago baten alde
edukia	Badira zenbait aste ospakizun eta jai...
lotura	https://www.berria.eus/paperekoa/1876/016...

4. **DATUEN PROZESAKETA**

Behin datuak lortuta, bildutako albisteetatik nabarmenak diren izendun entitateak jasotzeko asmoarekin, bi pauso ezberdin eman beharko dira. Lehenik eta behin, albiste bakoitzean izendun entitateen identifikazioa burutuko da, albiste bakoitzean aipatzen diren izendun entitate guztiak batuz. Bigarren pauso moduan, albisteetako izendun entitateetatik nabarmenak eta berrienak aukeratuko dira.

4.1. **Izendun entitateen identifikazioa**

Izendun entitateen identifikazioa pertsonen, erakundeen edota toki izenak identifikatu eta sailkatzean datza. Ataza hau automatikoki burutzea ez da batere erraza, hainbat ataza desberdinen kateaketa eskatzen baitu, pauso bakoitzean doitasuna galtzen delarik. Horrela, testu baten tokenizazioa, edo hitzen zein puntuazioaren zatikatzea, egiten da. Hirugarren pauso bezala, izendun entitateen identifikazioa eta sailkapena burutzen da. Azkenik, lematizazioa izango genuke, izendun entitate bakoitzaren forma kanonikoa jasotzen duena. Gainera, euskararena bezalako baliabide mugatutako ingurune batean, zailagoa da horrelako teknologiak garatzea, beharrezkoak diren baliabideak oso mugatuak baitira.

Era horretan, euskarako testuetan izendun entitateen identifikaziorako, aurretik garatutako bi sistema erabiliko dira, bien jardunaren arteko konparaketa egin ahal izateko. Horretarako, berriki garatutako Flair-BMC [3] eta, orain arteko sistema onenetarikoa, IXA *pipes* [4] konparatuko dira. Aurrez egindako ebaluaketa kuantitatiboaz gain, kasu errealetan aplikatu eta emaitzen eskuzko konparaketa kualitatibo bat burutu da.

— **IXA *pipes***: sistema honek eredu gainbegiratuak ikasten ditu, horretarako Perzeptroiaren algoritmoa [7] erabiltzen du. Sailkapena, etiketatutako corpusarekin eta etiketatu gabeko testutik erauzitako ezaugarrien clusterren konbinaketari esker burutzen da. Era horretan, informazio lokala, hitzen irudikapenen ezaugarrien hiru mota ezberdinekin konbinatuko da: Brown [8], Clark [9] eta Word2Vecen [10] oinarritutako clusterrak. Horrela, hitzen irudikapenak burutzeko, aipatutako hiru teknika ezberdinetatik (Brown, Clark eta W2V) eratorritako clusterrak konbinatzen dira.
— **Flair-BMC**: Flair-ek ikasketa sakoneko sistema eta karaktereetan oinarritutako testuingurudun hitz-bektoreei egiten die erreferentzia [11]. Flair hitz bektoreek hitzak jasotzen dituzte karaktere segidak izango balira bezala. Gainera, hitz baten bektorean oinarritutako irudikapena gertuko testuinguruaren araberakoa izango da. Flair, hitz-bektoreek zein sistemak, oso emaitza onak lortu izan dituzte sekuentzia-etiketatze atazetan, ingelesezko izendun entitateen identi-

fikazioan zein, kategoria morfosintaktikoaren iradokizunean, besteak beste. Horregatik, Flair-BMC euskararako eredua entrenatzeko, Flair hitz-bektore propioak entrenatu ziren, Basque Media Corpusa (BMC) baliatuta [3]. BMC, 224.6 milioi tokenez osatutako euskarazko corpusa da, euskarazko hedabide digitaletako zein euskarazko Wikipediako edukiz hornitua.

Bi sistemen entrenamendu zein ebaluaketarako Euskararako Izendun Entitateen Corpusa (EIEC) erabili da, euskarazko esaldiez osatutako corpusa, izendun entitateak eskuz etiketatuak dituena [12]. Horrela, 3. taulan, bi sistema ezberdinen ebaluazioko emaitzak ikus daitezke, sistemen arteko konparaketa ahalbidetuko duten metrikekin. Ikus daitekeen moduan, bi identifikatzaileen arteko aldea handia da, Flair-BMC ereduan oinarritutako sistemak 6 puntu baino gehiagoko F1 balioa lortzen baitu IXA *pipes* sistemarekin alderatuta. Horrek esan nahi du Flair-BMC identifikadoreak askoz zehaztasun handiagoa daukala izendun entitateak identifikatzeko orduan [3].

3. taula. Euskarazko Izendun Entitateen Identifikaziorako ebaluaketa emaitzak EIEC corpusean.

	Doitasuna	Estaldura	F1
IXA *pipes*	80.66	73.14	76.72
Flair-BMC	84.32	82.66	**83.48**

Ebaluaketa kuantitatiboaz gain, ataza zehatz honetan bi identifikadoreen ebaluaketa kualitatiboa burutu da, ataza zehatz honetan bakoitzaren errendimendua zuzenean ebaluatzeko. Bi identifikadoreak kualitatiboki ebaluatzeko, bakoitzaren emaitzak aztertu eta konparatu dira, testu zati berdinetan emandako irteerak alderatuz. Horrela, emaitzen eskuzko azterketa eginda, esan beharra dago bi sistemek nahiko emaitza antzekoak ematen dituztela orokorrean, baina kasu batzuetan hobekuntza nabaria dela. Emaitza ezberdinak izan dituzten bi kasu zehatz ausaz aukeratu dira corpusa osatzen duten albiste guztien artean. Era honetan, identifikadore bakoitzaren errendimenduen arteko konparaketa zehatzago bat egin ahal izan da.

1. «Egia berdaderoa eta fede absolutua duten eredu kolektibista bortitz askoren aurrean, mundua bakoitzaren askatasunetik eraiki daitekeen biderako ere balio digu, atzean uzteko [**Platon**], [**Aristoteles**] eta [**Hegel**] ditxosozkoak irudikatutako gizarte ideal eta perfektuak...»

   — IXA *pipes*: [Hegel]
   — Flair-BMC: [Platon, Aristoteles, Hegel]

2) «[**Troy Price**] [**Iowa**]ko [**Alderdi Demokrata**]ren presidenteak barkamena eskatu du, eta emandako datuak zuzenak direla ziurtatu du [**CNN**] telebista-katean egindako adierazpenetan.»

— IXA *pipes*: [iowa alderdi demokrata, cnn]
— Flair-BMC: [Troy Price, Iowa, alderdi demokrata, CNN]

Alde batetik, 1. adibidean ikus daiteke IXA *pipes* identifikadoreak soilik izendun entitate bat detektatzen duela (*Hegel*), Flair-BMC ereduak 3 detektatzen dituen bitartean: *Platon, Aristoteles eta Hegel*. Bestalde, 2. adibidean Flair-BMC eredua *Troy Price* identifikatzeko gai izateaz gain, *Iowa* eta *Alderdi Demokrata* banatu eta bakoitza izendun entitate propiotzat identifikatu ditu. Beraz, Flair-BMC eredua, ataza zehatz honetarako ere sailkatzaile egokia dela iruditzen zaigu, ebaluaketa kuantitatiboaz gain, corpus zehatz honetan kualitatiboki hobeto badabilela frogatua geratu baita.

Behin erabakita Flair-BMC identifikadorea hobea dela, sistema horrekin erauzitako izendun entitateak erabiliko dira hurrengo ataletan. Bestalde, izendun entitate mota guztietatik (pertsona, erakunde eta lekuak) soilik pertsonenak aukeratuko dira, hauen monitorizazioa egitea baita asmoa.

4.2. **Izendun entitate berri eta aipatuenen aukeraketa**

Azken pausoa pertsonen izendun entitate nabarmenen aukeraketa izango da, ohikoenak direnak alboratuz eta azkeneko egunetan agertzen ari diren berriak aukeratuz. Pertsonei dagokien izendun entitate berrienen aukeraketa egiteko, termino maiztasun-alderantzizko dokumentu-maiztasuna (ingelesez *Term Frequency–Inverse Document Frequency* edo TF-IDF) erabiliko da. Era horretan, artikulu berriak eta zaharrak konparatuko dira, berrienak eta berezienak diren pertsonak aukeratzeko. Banaketa horri esker, azkeneko asteko pertsonen izendun entitateak beste guztiekin konparatzen dira, ohikoak diren pertsonen izendun entitateak alboratu eta azkeneko astean nabarmenak direnak jasotzeko asmoz.

TF-IDF neurria, terminoen maiztasuna eta alderantzizkoa dokumentu maiztasunaren arteko biderketa dugu. Termino-maiztasunari (TF) dagokionez, dokumentu bakoitzean termino bakoitza zenbat aldiz azaldu den neurtzen du. Alderantzizkoa dokumentu maiztasunari (IDF) dagokionez, termino bakoitza dokumentu guztietan ohikoa edo arraroa den zehazten du. TF-IDF neurriak, TF eta IDF neurriak konbinatuz, dokumentu zehatzetan termino bereziak identifikatzen lagunduko digu. Dokumentu-bilduman gutxitan azaldu den termino bat dokumentu zehatz batean maiz agertuz gero, TF-IDF puntuazio altua lortuko du. Dokumentu zehatz horretako termino horren TF-IDF puntuazio altuak esan nahi du, termino zehatz hori lagungarria izan daitekeela dokumentua definitzeko orduan. Hori, baliagarria zaigu

dokumentu bilduma batean dokumentu zehatzen gaiak zehazteko, besteak beste. Kasu zehatz honetan, albiste zahar eta berrien dokumentu multzoan izendun entitate berri eta nabarmenak identifikatzeko baliatuko da.

Hurbilpen honetan, soilik bi dokumentuk osatuko dute dokumentu bilduma, bata albiste zaharrez osatua eta bestea albiste berriez osatua. Iragan asteko izendun entitate nabarmenak identifikatzeko, iragan astean argitaratutako albiste guztien izendun entitateekin dokumentu bat sortzen da, albiste berrien dokumentua izango dena. Aste zehatz hori baino lehenago publikatu diren albisteetako izendun entitate guztiekin, berriz, albiste zaharren dokumentua sortuko da. Bi dokumentuz osatutako dokumentu-bilduma horren gainean TF-IDF teknika aplikatuko da, modu horretan, albiste berrietan nabarmenak diren izendun entitateak identifikatuz.

4. taulako adibidean ikus daitekeen modura, albiste berrien bilduman, Arantxa Tapia, Iñigo Urkullu eta Puigdemont aipatuenak izan arren, ez dira izendun entitate nabarmen bezala kontsideratuko, albiste zaharren bilduman ere aipatuenen artean baitaude. Denboran zehar errepikatzen diren izendun entitateetan erreparatu ordez, berriki atentzioa sortzen ari diren izendun entitateak identifikatzea izan da gure asmoa lan honetan. Abagune zehatzean zeresana sortzen ari diren pertsonak identifikatzeko asmoa dago horren atzean. Horrela, TF-IDF teknikari esker, Martín Villa eta Maria Servini izendun entitateak identifika ditzakegu abagune zehatzeko pertsona nabarmentzat. Horrela, TF-IDF teknika baliatuz, aste zehatz bakoitzean nabarmenak diren izendun entitateak identifikatzeko gai gara.

4. taula. Dokumentu talde bakoitzean aipatuenak diren pertsonen izendun entitateak.

Albiste zaharrak	Albiste berriak
Arantxa Tapia	Arantxa Tapia
Iñigo Urkullu	Iñigo Urkullu
Carles Puigdemont	Carles Puigdemont
Jose Luis Zumeta	Martín Villa
Lucio Urtubia	Maria Servini

5. EMAITZEN ARGITALPEN DINAMIKOA WIKIPEDIAN

Jasotako datuak prozesatu ostean, pertsonen izendun entitate nabarmenak identifikatuta edukiko ditugu, azkeneko astean esanguratsuenak direnak erakutsiz. Baina, abagunearen argazki estatiko hori ez da gai gizartearen ezaugarri den dinamikotasunaren ñabardura islatzeko. Horregatik, etengabe eguneratzen diren emaitzak argitaratzeko modua dinamikoa izango da, pertsonen izendun entitate nabarmenak gertakarien arabera aldatuko dira. Era horretan,

emaitza hauek modu ezohiko batean erakutsiko dira, uztargarriak diren bi metodo dinamiko ezberdin erabiliz. Aukeretako batek, emaitza gordinak Wikipediako orri batean argitaratuko ditu, astero eguneratuz. Bigarren aukerak bistaraketa berritzaile bat proposatzen du, eta izendun entitateak grafiko interaktibo batean argitaratuko ditu, hori ere etengabe eguneratzen egongo dena.

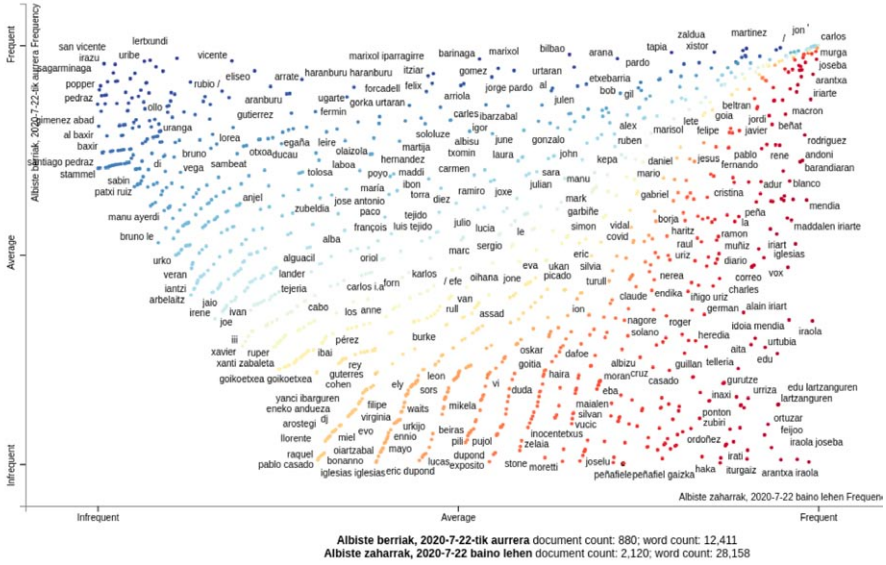
Alde batetik, lehen aukera bezala, Euskal Wikipediako orri bat baliatuko da astero lortzen diren izendun entitate nabarmenak erakusteko asmoarekin. Astelehenero, iragan berri den asteko datuak zein aurreneko datu guztiak alderatuko dira, iragan berri den astean aipatu diren izendun entitate nabarmenak lortu eta argitaratzeko. Emaitzen bistaraketarako modu honen helburua informazioa komunitatearekin konpartitzea izango da, kontsultatzen duen edonork aukera izango baitu emaitza ikusteko. Orriak, erabilitako metodologiaren azalpen txiki bat edukitzeaz gain, azkeneko lau asteetan nabarmenak izan diren izendun entitateak erakutsiko ditu, aste bakoitzean 10 izeneko zerrenda bat erakutsiz. Gainera, jaso diren aste guztietako izen nabarmenak kontsultatu ahalko dira, informazio guztia gordeta eta erakusgarri egongo baita. Bukatzeko, aipatu beharra dago, emaitzetan erakusten diren izendun entitateak esteka bezala erakusten direla, izendun entitate hauek Wikipediako beren orriekin lotuz, existituko balira. Eguneraketa dinamiko hori guztiz automatikoa izango da, *mwclient* euskarria erabiliz, astero emaitzak publiko eginez.

Bestalde, bigarren aukera bezala, astero adierazgarrienak diren izendun entitateak aurkitzeaz gain, emaitzen bistaraketa ezberdin bat proposatzen da, izendun entitate maiztasuna eta berritasuna aintzat hartzen dituena. Scattertext [13] teknikari esker, izendun entitate berriak konparatu ahalko dira zaharrenekin. Horrez gain izendun entitate aipatuenak eta gutxien aipatuenen arteko konparaketa egin ahalko da aldi berean. Adierazpen grafiko hau egunero berrituko da, iragandako 7 egunetako datuak eta azkeneko hilabeteko datuak konparatuta, izendun entitate sailkatu eta agerpenen iturria ikusteko aukera emanez.

Bistaraketaren adierazpen grafikoan (1. irudia) izendun entitate banaketa topa dezakegu, denboraren eta agerpen kopuruaren arabera. Era horretan, bi dimentsioetako grafikoaren goiko erdian entitate berri eta ohikoenak topa ditzakegu. Aldi berean, grafikoaren eskuineko aldean entitate zahar ohikoenak topatuko dira. Halaber, entitate berri aipatuenak grafikoko goiko eskuineko koadrantean aurkitu ahal izango ditugu: azkeneko astean aipatuenak izan diren entitateak izango dira horiek. Bestalde, ezkerreko goiko koadrantean beti aipatuak diren entitateak kokatuko dira, hau da, ohikoenak. Eskuineko beheko koadrantean, ostera, albiste zaharretan ohikoak izan diren eta albiste berrietan agerpen txikia daukatenak azaltzen dira.

Bistaratze sistema honek izendun entitate bilatzaile bat ere badauka, entitatea grafikoan kokatzeaz gain bere agerpen guztiak emango dizkiguna. Agerpenetan hedabidea, eguna, albistera lotura eta albistean agertzen diren

bestelako entitateak azalduko dira. Era horretan, entitate bakoitzaren informazio ahalik eta osatuena lortuko da, eta bere agerpenen testuingurua erakutsiko duen bistaratze bat eskainiko du.



1. irudia. Bistaraketa scattertext erabilita.

Laburbilduz, emaitzak bi modutara publikatuko dira etengabe, nahieran kontsultatu eta erabiltzeko. Batetik Wikipediako web orria baliatuz[2], bertan astero nabarmenak diren izendun entitateak kontsultatzeko aukera emanez. Bestetik, berritasunaren eta maiztasunaren araberako adierazpen grafiko batean erakutsiko dira emaitzak, Scattertext baliatuta[3]. Gainera, prozesu guztian zehar erabilitako kodea eskuragarri dago nahi duenarentzat[4].

6. **ONDORIOAK**

Euskararako diseinatutako hizkuntza-teknologiak aplikatuz, testuetatik oinarrizko informazioa erauztea izan da lan honen muina. Izendun en-

[2] https://eu.wikipedia.org/wiki/Wikiproiektu:Euskarazko_albisteetako_Izen_Entitateak
[3] http://ixa2.si.ehu.es/josebafdl/nortzuetaz.html
[4] http://ixa2.si.ehu.es/josebafdl/NE_extraction_basquemedia-master.tar.xz

titateak identifikatzeko orduan orain arteko sistemarik onena eta proposamen berri bat konparatu dira enpirikoki, neurona-sareetan oinarritutako sistema berriaren nagusitasuna erakutsiz. Honekin, teknika berrietara egokitzearen garrantzia azpimarratzearekin batera, hizkuntza bakoitzean oinarritutako modelo zehatzen beharra erakutsi da, euskararen adibidea erabiliz.

Izendun entitateak euskarazko artikuluetan identifikatuz, testuetatik oinarrizko informazioa erauztea lortzen da. Kasu zehatz honetan, hedabide digitalek euskaraz idazten dutenean zer pertsonaiaz idazten duten identifikatu ahal izan da. Gainera, albiste berri eta zaharren arteko denboraren araberako konparaketari esker, aipatu diren pertsonen artean berrienak zeintzuk diren lortzen da. Era honetan, hedabideek aktore berriei buruz idazten dutenean, posible izango da hauek identifikatzea.

Lan honen berritasunen artean, emaitzen automatizazioa daukagu, astero emaitzak berrituko baitira. Emaitza dinamikoen publikotasunari esker, euskarazko albisteen izendun entitate berrienak automatikoki argitaratuko dira astero, komunitateak nahieran erabili dezan informazio hori.

Etorkizuneko lanetarako ere hainbat ate irekitzen dira egindako lanarekin, hainbat arlo ezberdin jorratzeko aukera emanez. Hobekuntza interesgarri bat korreferentzia ebazpena izango litzateke, pertsona berdinari aipamena egiten dioten izendun entitate ezberdinak bateratzeko asmoarekin (EAEko lehendakaria = Iñigo Urkullu = Urkullu), errealitatearen isla argiago bat edukiz. Bestalde, izendun entitateen desanbiguazioa aplikatzea ere interesgarria izango litzateke, era berdinean idazten diren izendun entitateak ezberdintzeko aukera emanez. Emaitzen argitalpenari begira, identifikatutako izendun entitate nabarmenenen batek euskarazko Wikipedia orria edukiko ez balu, berau sortzea Wikipediako bestelako orriak euskarara automatikoki itzuliz. Horrez gain, lanean garatutako sistemak beste hizkuntzetara moldatzea ere interesgarria izango litzateke, albisteen monitorizazio zein izendun entitateak identifikatzeko sistemak garatzeari esker.

Lan honetatik eratorritako teknika eta emaitzak beste analisietarako erabili ahal izango dira, Wikipediako artikulu, artikulu zientifiko edota Twitterreko edukietan, adibidez. Horrela, izendun entitate garrantzitsuenak zeintzuk diren identifikatuz, testuen gaiak edota antzekotasunak identifikatzeko aukera emango digu. Etorkizunean ikerkuntza sozialerako baliagarria izan daitekeen metodologia bat ere aurkeztu da, hedabide eta izendun entitateen arteko erlazioak aztertuz informazio ugari lor baitaiteke. Horrela, Twitter bezalako datu iturrietan ikusitako ikerketa sozialerako aukerak [14] hedabide digitaletatik erauzitako datuekin ere hornitzeko aukera emango du metodologia honek.

7. ESKER ONAK

Udako Euskal Unibertsitateari (UEU) eta Euskal Wikipediako Kultur Elkarteari lan hau diruz laguntzeagatik, *Humanitate digitalen inguruko euskarazko ikerketa sustatzeko 2019 deialdia* bitartez. Aldi berean, Rodrigo Agerri RYC-2017-23647 diru-laguntzaren jasotzailea da.

BIBLIOGRAFIA

[1] NADEAU, D. eta SEKINE, S. 2007. «A survey of named entity recognition and classification». *Lingvisticae Investigationes*, **30**,1, 3-26.

[2] ETCHEGOYHEN, T. *et al*. 2018. «Neural machine translation of basque». *Proceedings of the 21st AnnualConference of the European Association for Machine Translation*, 139-148.

[3] AGERRI, R. *et al*. 2020. «Give your Text Representation Models some Love: the Case for Basque». *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4781-4788.

[4] AGERRI, R. eta RIGAU, G. 2016. «Robust multilingual named entity recognition with shallow semi-supervised features». *Artificial Intelligence*, **238**, 63-82.

[5] SAN VICENTE, I. *et al*. 2018. «Real Time Monitoring of Social Media and Digital Press». *arXiv preprint arXiv:1810.00647*.

[6] ZEMAN, D. *et al*. 2018. «CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies». *In Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, 1-21.

[7] COLLINS, M. *et al*. 2002. «Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms». *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, **10**, 1-8.

[8] BROWN, P. *et al*. 1992. «Class-based n-gram models of natural language». *Computational linguistics*, **18(4)**, 467-479.

[9] CLARK, A. *et al*. 2003. «Combining distributional and morphological information for part of speech induction». *In Proceedings of the tenth conference on European chapter of theAssociation for Computational Linguistics-*, **1**, 59-66.

[10] MIKOLOV, T. *et al*. 2013. «Distributed representations of words and phrases and their compositionality». *In Advances in neuralinformation processing systems*, 3111-3119.

[11] AKBIK, A. *et al*. 2018. «Contextual String Embeddings for Sequence Labeling». *27th International Conference on Computational Linguistics*, 1638-1649.

[12] ALEGRIA, I. *et al*. 2004. «Design and development of a named entity recognizer for an agglutinative language». *In First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*.

[13] KESSLER, J. 2017. «Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ». *Proceedings of ACL 2017, System Demonstrations*, 85-90.

[14] FERNANDEZ DE LANDA, J. *et al*. 2019. «Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case». *Information, MDPI*, 10, 6, 212.