eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Neurona-sareetan oinarritutako korreferentzia-ebazpen automatikoa

Egilea

Gorka Urbizu Garmendia

Zuzendariak

Ander Soraluze eta Olatz Arregi

informatika fakultatea
facultad de informática

2018ko irailaren 5a

Laburpena

Neurona-sareetan oinarritutako euskararako lehenengo korreferentzia-ebazpenerako sistema aurkezten da GrAL honetan. Horretarako polonierarako eraiki berri den sistema hartu da oinarritzat eta euskararen ezaugarrietara moldatu da. Korreferentzia anotatuta duen EPEC (euskarazko erreferentziazko corpusa) corpusaren zatia, 45.000 hitzetakoa, erabili da neurona-sarea entrenatu eta ebaluatzeko. Ondoren, sistema hobetzeko asmotan, hainbat proba egin dira: hitz-embeddingen dimentsioak handitu, ezaugarri berriak gehitu eta neurona-sarearen parametroak aldatu. CoNLL metrikan lortu den emaitzarik onena % 54,66 puntuko F-measurea izan da urrezko aipamenekin eta % 41,20 puntuko F-measurea aipamen-detektatzaile automatikoarekin.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Hizkuntzaren prozesamendua (NPL - *Natural Language Processing*) informatika, adimen artifiziala eta hizkuntzalaritza diziplinen arteko arloa da, pertsonek, makinekin edo beste pertsona batzuekin duten hizkuntzaren bitarteko komunikazioa errazteko tresna konputazionalak ikertzeaz arduratzen dena. Hizkuntzaren prozesamendua eta hizkuntzalaritza konputazionala sinonimotzat har daitezke, nahiz eta lehenak ikuspuntu teknologikoa eta bigarrenak hizkuntzalaritzaren ikuspuntua azpimarratzen dituztela esan ohi den.

Euskal Herriko Unibertsitateko Ixa ikerketa taldeak, 1987an sortu zenetik, 30 urte daramatza hizkuntzalaritza konputazionalaren arloan lanean hizkuntzaren tratamendu automatikoan lan egiten. Euskararen gaineko ikerketa aplikatua du xede nagusitzat, eta horretarako UPV/EHUko Informatika Fakultatean nagusiki informatikariz eta hizkuntzalariz osatutako diziplinarteko taldea dabil hizkuntzaren inguruko ikerketa eta produktuen garapenean lanean.

Hamarkada horietan, Ixa taldea proiektu askotan aritu da, kasu batzuetan enpresekin edo erakundeekin elkarlanean, eta jardun honek hainbat tresna, aplikazio, tesi eta artikulu utzi ditu. Euskararako beharrezko diren tresnen artean, besteak beste, Xuxen zuzentzaile ortografikoa (Agirre et al., 1992), Euskal Wordnet (Pociello et al., 2011), sintaktikoki etiketatuta dagoen euskarazko EPEC corpusa (Aduriz et al., 2006) eta euskararen analisi linguistiko konputazionalerako analisi-kate sendo bat (Aduriz et al., 2004; Otegi et al., 2016) garatu ditu.

Gradu amaierako lan honek aurrez hizkuntzalaritza konputazionalaren arloan, Ixa taldearen barnean euskarazko testuen korreferentzia-ebazpenean egindako lanei (Soraluze, 2017) teknika berriekin jarraipena ematea du helburu.

Korreferentzia-ebazpena honela defini genezake (Grishman and Sundheim, 1995):

"Testu bateko bi espresio testualek objektu berbera adierazi edo erreferentziatzen dutenean, bi espresio horien artean korreferentzia erlazio bat dagoela esan ohi da. Testu batean ager daitezkeen espresio testual horien arteko korreferentzia-erlazioak ebaztea helburu duen atazari korreferentzia-ebazpena deritzo".

Adibidez:

(1) Nazio Batuen Erakundea izan zen bitartekari eta hark hartu zuen prozesuaren ardura.

Nazio Batuen Erakundea, *bitartekari* eta *hark* espresio testualek NBEari erreferentzia egiten diotenez, korreferenteak direla edo korreferentzia-erlazioa dutela esan dezakegu.

Korreferentzia-ebazpena terminoa *Message Understanding Conference* (MUC-6, 1995) konferentzian zehaztu zen ikuspuntu konputazionaletik. Eta hortik aurrera, ataza horri erantzuteko sistema automatikoen garapenean aurrerapausuak eman dira.

Aipatzeko moduko beste bi termino *entitatea* eta *aipamena* dira, ataza honetan sarri erabiltzen direnak. Entitatea mundu errealeko pertsona, objektu edo erakunde bat litzateke; aipamena berriz, entitate bati erreferentzia egiten dion hitz multzo edo espresio testuala da.

Termino horiek hobeto ulertzeko, lehengo adibidera itzul gaitezen:

(2) [Nazio Batuen Erakundea] izan zen [bitartekari] eta [hark] hartu zuen [prozesuaren ardura].

Adibide honetan kortxete artean ikus ditzakegunak aipamenak izango lirateke, lehenengo hirurak, [Nazio Batuen Erakundea], [bitartekari] eta [hark] entitate berdinari egiten diote erreferentzia, beraz, korreferenteak dira eta korreferentzia kluster edo multzoa osatzen dute. Bestalde, [prozesuaren ardura] aipamenak entitate ezberdin bati egiten dio erreferentzia eta horregatik ez da gainerako aipamenekin korreferentea. Korreferentzia klusterrik osatzen ez duten aipamen hauei *singleton* deritze.

Korrefentzia-ebazpena bi azpi-atazatan banatu ohi da, alde batetik aipamen-detekzioa, eta beste aldetik erreferentzien ebazpena (Pradhan et al., 2011). Lehenik aipamen-detekzioa egiten da, testuan zehar entitateren bati erreferentzia egiten dioten espresio testualak identifikatuz. Ondoren, aipamenak multzokatu behar dira, entitate berdinari erreferentzia egiten diotenak elkartuz.

Korreferentzia-ebazpen automatikoa garrantzitsutzat jotzen da, oro har, testu ulermen sakona dakarren *Lengoaia Naturalaren prozesamenduko* (NLP) ataza oro burutzeko (Clark, 2015). Besteak beste, informazio erauzketan, testuen laburpenean, galdera-erantzun sistema automatikoetan, sentimenduen analisian eta itzulpen automatikoan aplikatzen da. Hori dela eta, korreferentzia-ebazpenean, teknika ezberdinak erabiliz, hainbat saiakera egin da. Lehenik erregela bidezko sistema adimendunak erabiliz, gerora korreferentzia-ebazpenerako corpus handiagoak sortzearekin batera ikasketa automatikoa erabiliz, eta berriki honen barruan neurona-sareekin.

Euskararen kasuan ere, gainerako alorretan duen aplikagarritasunagatik azken urteotan korreferentzia-ebazpenerako tresnen garapenean jardun da UPV/EHUko Informatika Fakultateko Ixa taldea, besteak beste, Ceberio et al. (2008), Arregi et al. (2010) eta bereziki Soraluze (2017) lanetan ikus daitekeen moduan.

Lan honen helburua, euskararako korreferentzia-ebazpenean egindako lanari segida emanez, euskararako neurona-sareetan oinarritutako lehen korreferentzia-ebazpenerako sistema eraikitzea da, baliabide askoko zein ertainetako hizkuntzetan artearen egoeran dauden emaitzak lortu baitira ikasketa sakona erabiliz.

Gradu amaierako lanaren memoria honek ondorengo eskema jarraitzen du, sarrera-ren ondoren, Proiektuaren Helburuen Dokumentua atalean (2. kapitulua), proiektuaren irismena, planifikazioa eta kudeaketa finkatuko dira. Aurrekariak atalean (3. kapitulua) korreferentzia-ebazpenaren arloan orain arte egindako aurrerapenak bilduko dira. Ondo-ren Garapena atalean (4. kapitulua) erabilitako corpusa eta aurreprozesaketa zein neurona-sareen inguruan hartutako erabakiak azalduko dira. Bukatzeko, lortutako emaitzak eta ateratako ondorioak aipatuko dira (5. eta 6. kapituluak), bibliografiarekin memoria hau amaituz.

2. KAPITULUA

Proiektuaren Helburuen Dokumentua

Aurkibidea

Proiektuaren Helburuen Dokumentuan, proiektuaren deskribatzeaz eta helburuak azaltzeaz gain, lanaren deskonposaketa egitura (LDE) eta landutako atazak zerrendatuko dira kapitulu honetan. Gainera, emangarriak eta mugarriak azalduko dira eta proiektuan zehar egindako lana islatuko duen Gantten diagrama eta dedikazioaren zenbatespena aurki daitezke.

2.1 Proiektuaren deskribapena

Proiektuaren helburua, euskarazko testuen korreferentzia-ebazpena neurona-sareetan oinarrituz automatikoki egiteko gai den sistema garatzea eta ebaluatzea da. Horretarako eskuragarri dagoen, aipamenak eta korreferentzia erlazioak anotatuta dituen euskarazko EPEC-KORREF (Ceberio et al., 2018) corpusetik abiatu eta aurreprozesaketa bat egingo da, ezaugarri batzuen erauzketa eginez neurona-sarearen sarrera lortzeko. Ondoren neurona-sarea entrenatu eta ebaluatu egingo da. Azkenik hobekuntza batzuk egin, emaitzak konparatu eta ateratako ondorioak memoria batean bilduko dira.

2.2 Proiektuaren plangintza

2.2.1 Lanaren deskonposaketa egitura

2.1 irudiko diagraman proiektuan zehar landutako atal ezberdinak ikus daitezke lan-paketetan antolatuta. GrAL hau osatzen duten ataza ezberdinak 4 multzotan banatu dira: ikasketa, garapena (inplementazioa eta probak), dokumentazioa eta kudeaketa.

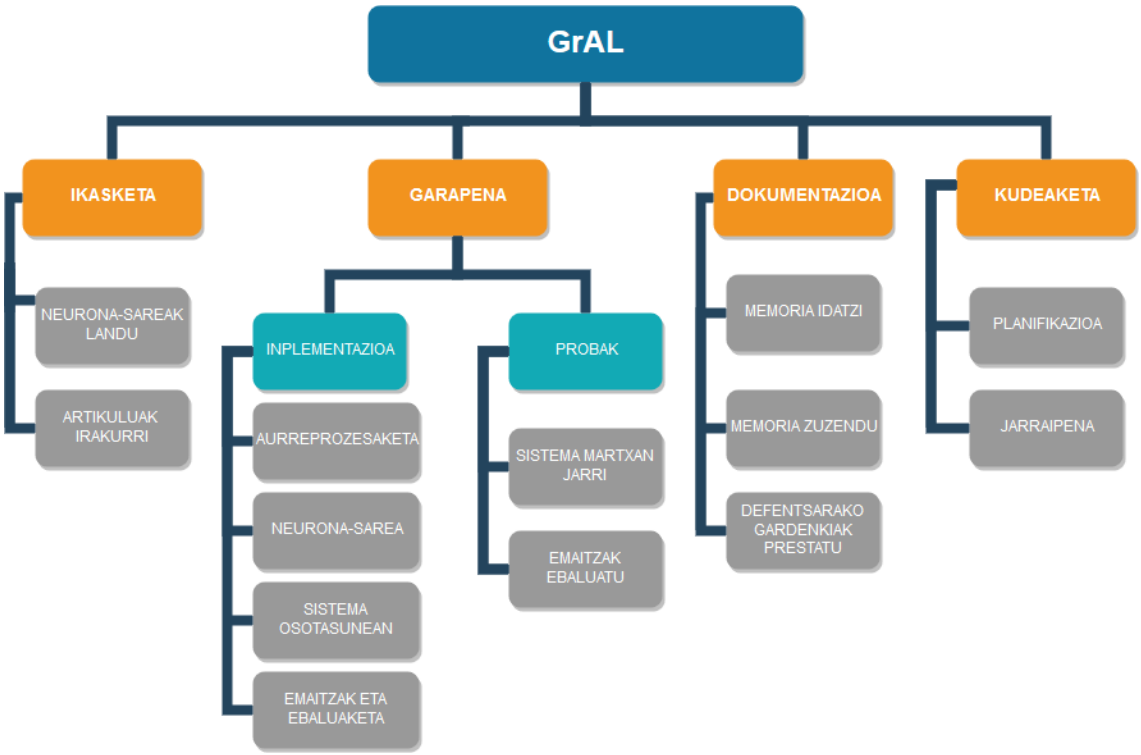
2.2.2 Atazak

Jarraian, proiektua garatzeko beharrezkoak izan diren atazak zerrendatu dira, aurreko ataleko lanaren deskonposaketa egiturako lan-paketeak jarraituz.

A1. Ikasketa

A1.1 Neurona-sareak landu

A1.2 Artikuluak irakurri



2.1 Irudia: Lanaren deskonposaketa egitura (LDE diagrama)

A2. Garapena

A2.1 Inplementazioa

A2.1.1 Aurreprozesaketa

A2.1.2 Neurona-sarea

A2.1.3 Sistema osotasunean

A2.1.4 Emaitzak eta ebaluaketa

A2.2 Probak

A2.2.1 Sistema martxan jarri

A2.2.2 Emaitzak ebaluatu

A3. Dokumentazioa

A3.1 Memoria idatzi

A3.2 Memoria zuzendu

A3.3 Defentsarako gardenkiak prestatu

A4. Kudeaketa

A4.1 Planifikazioa

A4.2 Jarraipena

2.2.3 Emangarriak

Proiektu honetan hainbat eramangarri sortuko dira, nagusienak proiektuaren memoriaren txosten hau eta sortutako sistema bera izanik. Txostenean proiektuaren helburuak, kudeaketa, garapena eta ateratako ondorioak bilduko dira, *LaTeX* erabiliz idatziko da, eta estilo zuzen eta zaindu bat izango du. Bestalde, proiektuaren defentsararen egunean aurkezpenean lagungarri izango diren gardenkiak sortuko dira.

Dokumentazio idatziaz gain, sortutako neurona-sareetan oinarritutako euskararako lehenengo korreferentzia-ebazpenerako sistema ere emangarri bat izango da. Aurreprozesaketarako erabilitako kodea, neurona-sarea bera eta ebaluatzeko erabilitako kodea eskuragarri jarriko dira *github* plataforman [1].

2.2.4 Mugarriak

Proiektu honek hainbat mugarri administratibo ditu finkatuta:

- 2018ko apirilaren 20an GrALa erregistratu.
- 2018ko uztailaren 24ean ikasleak GrALaren matrikula, eta defentsa eskaera.
- 2018ko uztailaren 24ean tutoreak defentsa baimena.
- 2018ko irailaren 6an lana ADDI plataformara igo.
- 2018ko irailaren 17-20an lanaren defentsa.

Finkatuta dauden mugarri administratiboez gain, hainbat mugarri akademiko finkatu dira:

- 2018ko otsailaren 15erako neurona sareak landu eta artikuluak irakurri.

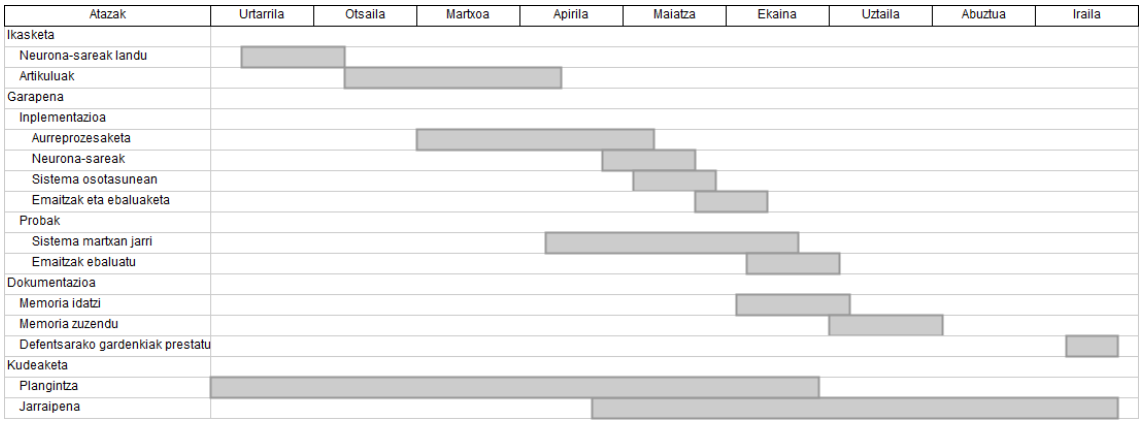
[1] https://github.com/gorka96/gorka96-EUSKORREF-NN

- 2018ko apirilaren 30erako oinarri lerroaren ezaugarrien erauzketa inplementatu.
- 2018ko maiatzaren 31rako korreferentzia-ebazpanerako sistema martxan jarri.
- 2018ko ekainaren 10erako memoriaren lehen bertsioa amaitu.

2.2.5 Gantt

Gantt diagrama:

2.2 irudiko diagraman proiektuaren atal ezberdinak ikus daitezke denboran zehar banatuta.



2.2 Irudia: Gantt diagrama

2.2.6 Dedikazioaren zenbatespena

2.1 taulan proiektua amaitzeko egindako lanaren dedikazioaren zenbatespena ikus daiteke, lan-paketeka antolatuta.

2.2.7 Arriskuen plana

Luzera handiko proiektuetan ohikoa da, lana amaitu bitartean gerta litezkeen arazoak ekiditeko edo hauen eragina baretzeko, arriskuak identifikatzea eta erantzun bat ematea.

Proiektu honetan arrisku nagusi bat identifikatu da: Ikerketa lan bat egiten dudan lehen aldia izanik, eta erabili behar ditudan neurona-sareei buruzko ezagutzarik aurretiaz ez dudanez, kodearen arazoei aurre egitean edo lehen diseinuan aurreikusi gabeko atazaren bat

Lan-paketea	Dedikazioaren zenbatespena (ordutan)
A1. Ikasketa	**70**
A1.1. Neurona-sareak landu	50
A1.1. Artikuluak irakurri	20
A2. Garapena	**140**
A2.1. Inplementazioa	**110**
A2.1.1. Aurreprozesaketa	40
A2.1.2. Neurona-sareak	40
A2.1.3. Sistema osotasunean	20
A2.1.4. Emaitzak eta ebaluaketa	10
A2.2. Probak	**30**
A2.2.1. Sistema martxan jarri	15
A2.2.2. Emaitzak ebaluatu	15
A3. Dokumentazioa	**75**
A3.1. Memoria idatzi	50
A3.2. Memoria zuzendu	20
A3.3. Defentsarako gardenkiak prestatu	5
A4. Kudeaketa	**15**
A4.1. Planifikazioa	10
A4.2. Jarraipena	5
GUZTIRA	**300**

2.1 Taula: Dedikazioaren zenbatespena

sortu delako, implementazio garaian atzerapenak pilatzea gerta liteke. Hori gertatzeko aukerak badirenez, hori ekiditen saiatzeko planifikazioan neure buruari jarritako tarteko epe eta mugarriak, ezustekoi aurre egiteko denbora utziz finkatuko dira. Hori nahikoa izango ez balitz, azken asteetan lanordu kopurua handituko da, eta hala ere, azken mugarrira iritsiko ez banintz, gradu amaierako lana atzeratu eta irailean aurkeztea erabakiko nuke.

Bestalde lan luze guztietan bezala, arazo tekniko zein pertsona akatsak tarteko, egindako lanaren zati bat edo lan osoa galtzeko arriskua identifikatu dut. Aldez aurretik neurriak hartzearren, inplementazioaren garaian, aldaketa nagusiak edo funtzionalitate berriren bat garatzen den aldiro, atal funtzional bakoitzaren kodearen segurtasun-kopia egingo dut sarean. Bestalde, hilabeteroko maiztasunarekin proiektu osoa biltzen duen direktoriaren segurtasun-kopia egingo dut sarean, aurreprozesatutako corpusak, kodea, eta bestelakoak biltzeko. Azkenik, memoriaren txostenaren babes-kopia egingo dut egunero.

2.2.8 Lan metodologia

Proiektu hau banakako lana denez, lan handiena etxetik egin da. Lauhilekoa hasi aurretik, urtarrilean neurona-sareak landu ditut courserak eskaintzen duen ikasketa sakonari buruzko kurtsoa eginez. Ikasturteak dirauen bitartean ikasketan eta garapenean lanean aritu naiz, ikasturtea amaitu ondoren, apirilean, denbora gehiago eskaini zaio gradu amaierako lan honi. Tutoreekin bilerak sarri egin dira, proiektuaren irismena zehazteko, gomendioak emateko eta zalantzak argitzeko. Bilerak kurtsoa amaitu arte hilabetean behin edo bitan egin dira, hortik aurrera astean behin. Aurreprozesaketak 30-120 minutu behar ditu eta sarea entrenatu eta emaitzak ebaluatzeak 5-20 minutu, denborak sarrerarako erabilitako bektorearen tamainaren arabera aldatzen direlarik. Sistema *Python3* lengoaian kodetu da, eta memoria idazteko LaTeX editorea erabili da.

2.3 Jarraipena eta kontrola

2.2.6 atalean proiektuari dedikatzea espero zen denbora ikusi dugu. Orain, behin proiektua amaituta egindako lan guztia islatuko dugu 2.2 taulan.

Lan-paketea	Dedikazioaren zenbatespena (ordutan)	Benetako dedikazioa (ordutan)
A1. Ikasketa	**70**	**83**
A1.1. Neurona-sareak landu	50	51
A1.2. Artikuluak irakurri	20	32
A2. Garapena	**140**	**156**
A2.1. Inplementazioa	**110**	**126**
A2.1.1 Aurreprozesaketa	40	71
A2.1.2 Neurona-sareak	40	11
A2.1.3 Sistema osotasunean	20	28
A2.1.4 Emaitzak eta ebaluaketa	10	17
A2.2. Probak	**30**	**30**
A2.2.1 Sistema martxan jarri	15	22
A2.2.2 Emaitzak ebaluatu	15	8
A3. Dokumentazioa	**75**	**95**
A3.1. Memoria idatzi	50	49
A3.2. Memoria zuzendu	20	41
A3.3. Defentsarako gardenkiak prestatu	5	5
A4. Kudeaketa	**15**	**27**
A4.1. Planifikazioa	10	21
A4.2. Jarraipena	5	6
GUZTIRA	**300**	**371**

2.2 Taula: Benetako dedikazioa

3. KAPITULUA

Aurrekariak

Aurkibidea

Kapitulu honetan gradu amaierako lan honen aurrekariak aipatuko dira, korreferentzia-ebazpenean aurrerapausoak emateko asmoz erabili diren teknika ezberdinak azalduz. Baliabide ugariko hizkuntzetan (nagusiki ingelesean) zein euskaran emandako pausuak aztertuko dira.

3.1 Korreferentzia-ebazpena

Korreferentzia-ebazpen automatikoaren hastapena 60-90eko hamarkadetan kokatzen da, lan esanguratsuenen artean Hobbs-en (1978) "Resolving Pronoun References", Carterek (1987) proposatutako "Shallow Processing Anaphor Resolver"(SPAR) eta Lappin and Leass-en (1994) "Algorithm for Pronominal Anaphora Resolution"artikuluak azpimarra ditzakegu (Hobbs, 1978; Carter, 1986; Lappin and Leass, 1994).

Korreferentzia-ebazpena seigarren eta zazpigarren *Message Understanding Conference* (Grishman and Sundheim, 1995; Hirschman, 1997) konferentzietan hasi zen lantzen espresuki, ordura arte itzulpen-automatikoko edo beste problema batzuen azpiatazatzat hartzen baitzen. Ordutik hainbat teknika aplikatu dira ataza ebazteko, erregela bidezko sistemak, ikasketa automatikokoak eta ikasketa sakonekoak.

Aurrez aipatu bezala, korrefentzia-ebazpena bi azpi-atazatan banatzen da, aipamen-detekzioan eta erreferentzien ebazpenean (Pradhan et al., 2011). Ondoren, bi azpi-atazetan eman diren aurrekariak aurkeztuko dira.

3.1.1 Aipamen-detekzioa

Testuan zehar entitateren bati erreferentzia egiten dioten espresio testualak identifikatzeari aipamen-detekzioa deritzo eta sekulako garrantzia du korreferentzia ebazpenean, azpi-ataza honetan lortutako emaitzek korreferentzia-ebazpenean lortutako emaitzak baldintzatzen baitituzte (Soraluze, 2017). Aipamen-detekzioan hobekuntzak lortuz gero, erroreak hurrengo atazetara hedatzea saihestuko litzateke eta korreferentzia-ebazpenaren eraginkortasuna handituko zen. Hori dela eta, oinarrizkoa da korreferentzia-ebazpenerako sistemetan lortutako emaitzak aipamen-detekzio automatikoarekin, edota urre-patroia (*gold standard*) baliatuz lortu diren adieraztea, kasu batzuetan biak erabili eta konparatzen direlarik.

Aipamen-detekziorako erabilitako teknikei dagokionez, erregelatan oinarritutakoak

eta ikasketa automatikoan oinarritutakoak bereizi genitzake (Pradhan et al., 2011). Orokorrean erregelatan oinarritutako aipamen-detektatzaileek emaitza hobeak ematen dituzte, baina artisautza lan handia eskatzen dute, eta beste hizkuntza batzuetara egokitzeak zailtasun ugari ditu eta emaitzak kaxkarrak dira.

Aipamen-detekzioan malgutasun handiarekin jokatu ohi da, ahalik eta estaldura handiena lortzeko, aipamenak baztertzea saihestu asmoz. Lortutako doitasuna oso ona ez izan arren, gerora korreferentzia-ebazpenerako sistemak soberan dauden aipamenak baztertzen ikas dezake (Wiseman et al., 2015; Clark, 2015).

3.1.2 Korreferentzia-ebazpenerako teknikak

Korreferentzia ebazpenerako lehen sistemak, aurrez izenordainen ebazpenerako sistemak bezala, erregelatan oinarritutakoak ziren. Orain urte gutxi batzuk arte, artearen egoera finkatu duten sistemek teknika hori erabili dute, ikasketa automatikoko lehen sistema arrakastatsuak (Soon et al., 2001) aspaldi sortu ziren arren. MUC-6 eta MUC-7 konferentzietatik, zein ondorengoetatik, sortutako anotazio kopuru esanguratsua izanik, bazirudien ikasketa automatikoko teknikek erregelatan oinarritutakoei aurrea hartuko zietela, baina erregelatan oinarritutakoak gailendu ziren oraindik urte luzez, horien artean bat aipatzearren, *CoNLL 2011 Shared Task*ean (Pradhan et al., 2011) lehenengo postua lortu zuen Stanfordeko korreferentzia-ebazpenerako sistema determinista (Lee et al., 2013).

Stanfordeko unibertsitatean garatutako korreferentzia-ebazpenerako erregelatan oinarritutako sistema bahe multzo batez dago osatuta. Sistemaren arkitektura modularrarengatik erraz integra daitezke bahe berriak, edo sistema ingelesa ez den beste hizkuntza batetarako egokitu (Chen and Ng, 2012; Fernandes et al., 2012).

Ikasketa automatikoari dagokionez, ikasketa gainbegiratua soilik azalduko da, ikasketa ez gainbegiratuak corpusen beharrik ez izatearen abantaila duen arren, artearen egoeran lortu diren emaitzak ez direlako oso onak izan (Ng, 2008). Aipamen eta erlazio korreferentez anotatutako corpusen tamaina handitzeak ikasketa automatikoaren bidea zabaldu zuen, ordutik sistema ugari plazaratu da, instantziak sortzeko metodo, ezaugarri linguistiko eta ikasketa algoritmo ezberdinak konbinatuz (Soraluze, 2017).

Aipamenetatik instantziak sortzeko honako 4 ereduak dira nagusi:

- Aipamen-bikote eredua (*mention-pair model*).

Eredu honetan sailkatzaile bat entrenatzen da aipamen bikote bat korreferente den edo ez erabakitzeko. Lehenik aipamen bikoteak korreferente edo ez-korreferente gisa sailkatzen dira, eta gero aipamen bikoteak multzokatzen dira algoritmo ezberdinekin. Testu batean dauden aipamen-bikote posible guztietatik gutxiengoak direnez korreferenteak, sailkatzailea entrenatzeko garaian bikote negatiboak gutxitze aldera, metodo ezberdinak aplikatu izan dira, nabarmenenak Soon et al. (2001) eta berriki Sapena et al. (2011) izanik.

- Entitate-aipamen eredua (*entity-mention model*).

  Eredu honetan aipamen bat aurretik sortutako aipamen multzo batekin korreferentea den edo ez erabakitzen da. Ikasketarako, aipamenak, klusterrak eta klaseak (ea korreferenteak diren edo ez) osatutako hirukoteak erabiltzen dira (Luo et al., 2004; Yangy et al., 2004).

- Aipamen-mailakatze eredua (*mention-ranking model*).

  Entrenamendu garaian aipamen bakoitza aurreko bi aipamenekin lotzen da, bata korreferentea eta bestea ez-korreferentea izanik. Instantziaren klaseak bi hautagaietatik onena zein den adierazten du, ebaluazio garaian berriz, bi hautagaietatik aurrekaria izateko zein den probableena aukeratzen da, bestea baztertuz, txapelketa (*tournament*) eredua deritzonean (Connolly et al., 1997; Yang et al., 2003).

- Multzo-mailakatze eredua (*cluster-ranking model*).

  Eredu hau aurreko bi ereduen (entitate-aipamen eta aipamen-mailakatze) arteko konbinazio bat da, eta bakoitzak dituen abantailak konbinatzen ditu (Rahman and Ng, 2009).

Korreferentzia-ebazpena gauzatzeko ikasketa automatikoa darabilten sistemen artean, BART (*Beautiful Anaphora Resolution Toolkit*) (Versley et al., 2008) sistema nabarmendu behar da, aurreprozesaketarako hainbat metodo, ikasketarako ezaugarri ezberdinak eta errore analisirako tresnak biltzen dituena. BARTen izaera modularrarengatik corpus, ezaugarri multzo eta hizkuntza ezberdinetara egokitu da.

Azken bizpahiru urteetan hizkuntzaren prozesamenduko atazetan neurona-sareek izan duten arrakasta dela eta, korreferentzia-ebazpenerako ere ikasketa sakona erabiltzen duten sistemak azaldu dira. Sistema gehienek ikasketa automatikoko sistemen antzera funtzionatzen dute, ikasteko eta sailkatzeko atalak neurona-sareekin ordezkatuz.

Gaur egun, neurona-sareen bitartez ikasitako hitz-embeddingek hitzen antzekotasun semantikoa biltzeko gai dira. Korreferentzia sistemetan ezagutza semantikoa integratzeko saiakerak egin dira lehenago ere, ezaugarri semantikoak erabiliz, baina ez da lortu esperotako hobekuntzarik. Hitz-embeddingekin, ordea, korreferentzia-ebazpenerako sistemei ezagutza semantikoa arrakastaz integratzea lortu da, emaitzetan hobekuntza nabarmenak lortuz.

Neurona-sareetan oinarritutako korreferentzia-ebazpenerako lehenetariko sistema esanguratsua Clark (2015) lanean aurkezten dena izan zen. Aipamen-bikote eredua erabiltzen du, aipamen bikote bakoitzeko ezaugarri ezberdinez osatutako bektore bat sortuz. Honako ezaugarri nagusiak erabiltzen ditu: aipamen bakoitzaren eta testuinguruaren esanahi semantikoa biltzen duten 50 dimentsiotako hitz-embeddingak, aipamenen arteko distantziak eta string-matching bezalako ezaugarriak. Neurona-sarearen konfigurazioari dagokionez bi ezkutuko geruza (*hidden layer*) dituen sare estandarra erabiltzen du, 300 eta 100 neuronatakoa bakoitza. Ondoren, antzeko ezaugarriak erabiliz entitate-aipamen eredua aplikatzen du, multzokatzean akatsak sahiesteko, emaitzak hobetuz (CoNLL *F-measure* 63).

Beste alde batetik, Wiseman et al. (2015) eta Wiseman et al. (2016) artikuluetan aurkeztutako sistemak ditugu. Horietan ere korreferentzia-ebazpena jorratzen da ikasketa sakonean oinarrituz, artearen egoeran hobekuntza txikiak lortuz. Wiseman et al. (2015) lanean ezaugarri linguistiko sorta zabala darabilte eta neurona-sarearen pisuak hasieratzeko bi azpi-atazetan bereizita entrenatzen dute sistema. Azpi-ataza horiek aipamenak korreferentzia erlaziorik duen erabakitzea eta aipamen korreferenteak elkartzea dira (CoNLL *F-measure* 63,4). Wiseman et al. (2016) lanean, berriz, neurona-sare errekurrenteak (RNN) erabiltzen dituzte, aipamenetatik abiatuz entitate multzoak sortzeko (CoNLL *F-measure* 64,2).

Gerora Clark and Manning (2016a) eta Clark and Manning (2016b) artikuluetan, aurreko sistemaren hobekuntzak proposatzen dituzte, artearen egoeran hobekuntzak lortuz. Lehenengoan, sistemak ikasketan erabilitako aipamen bakoitzeko ezaugarri kopurua handitzen dute, eta aipamen-mailakatze, multzo-mailakatze, aipamen-bikote eta entitate-aipamen ereduak konbinatzen dituzte (CoNLL *F-measure* 65,3). Bigarrenean, galera-funtzioa kalkulatzeko (*loss function*) korreferentzia-ebazpenaren ebaluaziorako B^3 metrika optimizatzen dute (CoNLL *F-measure* 65,7).

Lee et al. (2017) lanean aldiz, neurona-sareak erabiltzen dituzte, aipamen-detekzio automatikorik eta ezaugarriak erauzteko tresnarik gabe, korreferentzia-ebazpenerako sis-

tema itxi batekin, orain arteko korreferentzia-ebazpeneko sistema guztiei gailenduz. Hitz bakoitza eta hitz multzoak aipamenak izateko hautagai gisa hartzen ditu, eta neurona-sare konplexu bat erabilita, burua (sintagmaren hitz nagusia) topatu eta aipamenak detektatzea lortzen du. Behin aipamen-detekzioa burututa, korreferentzia-ebazpenerako beste neurona-sare multzo bat erabiltzen da. Sistema honen neurona-sarearen arkitektura oso konplexua izan arren, aurretik sortutako sistema guztiei gailendu zaie (CoNLL *F-measure* 68,8), eta hau aipamen-detekziorako edo ezaugarri linguistikoen erauzketarako tresnen beharrik gabe lortu da.

Orain arte aipatutako sistema gehienek nagusiki ingeleserako korreferentzia-ebazpena zuten helburu. 2010ean eginiko *SemEval-2010 Task 1* konferentzian, korreferentzia-ebazpena hizkuntza ezberdinentzat (ingelesa, nederlandera, alemaniera, italiera, gaztelera eta katalana) landu zen. Bertan hizkuntza ezberdinen arteko aldeak eta amankomunean zituzten ezaugarriak landu ziren. Urtebete beranduago *CoNLL 2011 Shared task*-ean ingelesezko *Ontonotes* corpusa zabaldu zen eta 2012an, *CoNLL 2012 Shared task*-en ingelesaz gain txinera eta arabiera landu ziren (Chen and Ng, 2012; Fernandes et al., 2012). Ordutik, hizkuntza txikiagoetan edo baliabide gutxiagokoetan ere egin dira korreferentzia ebazpenean aurrerapenak, tartean euskararako (Ogrodniczuk and Ng, 2017).

Ingelesa ez beste hizkuntzetarako ere hasi dira neurona-sareetan oinarritutako korreferentzia-ebazpenerako sistemak garatzen. Clark and Manning (2016b) artikuluan txinerarako sistema aurkezten da, Park et al. (2016) artikuluan korearrerakoa, eta Niton et al. (2018) artikuluan polonierarako. Polonierarako sistema hau aukeratu da GrAL honen abiapuntutzat, polonierak eta euskarak antzeko ezaugarriak dituztelako eta euskararen antzera, polonierak ingelesak baino baliabide mugatuagoak dituelako.

3.2 Euskararako Korreferentzia-ebazpena

Euskara eranskaria, buru-azkena, ordena librekoa eta pro-drop hizkuntza da. Euskara eranskaria denez, lemek forma ezberdinak har ditzakete, numeroa edo kasuaren arabera; adibidez, *"etxe"* lemak, *"etxea"*, *"etxeak"* eta *"etxetik"* bezalako formak har ditzake. Horregatik, string-parekaketa arrunta egitea soilik ez da nahikoa korreferentzia ebazteko euskararen kasuan. Ordena librekoa izateak sintaxiaren analisian sortutako anbiguotasunek korreferentzia-ebazpenaren ataza zailtzen dute. Pro-drop hizkuntza izateak ere, subjektu eta objektuen elipsiekin, atazari zailtasuna gehitzen dio. Gainera euskararen sistema nominalak ez du generorik eta izenordainek ez dute bizidun/bizigabe propietaterik

eta ezaugarri hauek lagungarriak izaten dira beste hizkuntza batzuetarako korreferentzia-ebazpenean. Ezaugarri hauek direla eta, korreferentzia-ebazpenerako sistema nagusiek (ingeleserakoak zein eleanitzek) ez dute emaitza onik ematen euskarako eta hauek egokitzeko beharra dago.

Euskarazko testuetan korreferentzia-ebazpen automatikoari dagokionez, Soraluze (2017) lana nabarmendu behar da. Tesi lan honetan daude bilduta euskararako korreferentzia-ebazpenaren bueltan egin diren aurrerapenak. Lehenik aipamen-detektatzaile automatiko bat garatzen dute, erregelatan oinarritzen dena eta euskarazko aipamenen egiturak kontuan hartzen dituena, (Soraluze et al., 2017b). Aipamen-detektatzaileak % 74,57 eta % 80,57 puntuko F-measure balioak lortzen ditu ebaluaziorako *Exact Matching* (urrezko aipamenaren berdina bada) eta *Lenient Matching* (aipamen automatikoaren mugak urre-patroiaren mugen barnean eta burua aipamenaren barnean kokatzen denean) protokoloetarako.

Soraluze et al. (2015) lanean, ingeleserako diseinatutako erregela bidezko Standfordeko sistema (Lee et al., 2013) egokitu da korreferentzia-ebazpena euskarazko testuetan egiteko. Stanford sistema 10 bahez osatua dago, baheak doitasun handienekotik hasi eta txikienekora aplikatzen dira, hasieran hartutako erabakiak ahalik eta ziurrenak izan daitezen, eta erabakirik zailenak amaierarako utziz. Bahe horiek 3 multzotan sailkatzen dira: string-parekatzean oinarritzen direnak, egitura bereziak tratatzen dituztenak eta izenordainen ebazpena egiten dutenak. Standfordeko sistemak dituen 10 baheak euskararen ezaugarrietara moldatu dituzte; euskaraz elipisiari aurre egiteko, bahe bat gehitu zaio sistemari elipsia behar den moduan tratatzeko. Egokitutako sistemak, nahikoa ditu Ixa taldean garatutako analisi-katearen eta euskararako sortutako aipamen-detektatzailearen irteera jasotzea euskarazko testuetako korreferentzia-ebazpena gauzatzeko. *CoNLL* metrikan 55,74ko *F-measure* balioa lortzen du aipamen automatikoekin eta 76,12 puntukoa urrezko aipamenekin.

Gerora, erregelatan oinarritutako sistema horren errore-analisi sakona burutu eta hobekuntzak proposatzen dituzte Soraluze et al. (2017a) lanean. Errore-analisia egitean *"Osasuna"* eta *"Talde Gorritxoa"* bezalako aipamenak lotzeko arazoak azaleratzen dituzte, ezagutza semantikoaren beharra azpimarratuz. Hori konpontzeko helburuarekin, Wikipedia eta Wordnet baliabide semantikoak ustiatuz sistemari bi bahe gehitzen dizkiote. Ondorioz, korreferentzia-ebazpenean *CoNLL* metrikan 55,98ko *F-measure* balioa (0,24 puntuko hobekuntza) lortzen du aipamen automatikoak erabiliz eta 76,51 (0,39 puntuko hobekuntza) urrezko aipamenekin. Hobekuntza txikia lortzea baliabide semantikoen urritasunagatik dela ondorioztatzen da, eta baliabide semantikoen urritasun hori are ageriagokoa da ba-

liabide gutxiko hizkuntzen kasuan (Versley et al., 2016).

Soraluze et al. (2016) lanean, ikasketa automatikoan oinarritutako sistema bat plazaratzen dute, ingeleserako diseinatutako *BART* korreferentzia-ebazpenerako sistema (Versley et al., 2008) euskararen ezaugarrietara egokituz (besteak beste *lemma-parekaketa* eta distantzia ezaugarriak gehitu dira). Horretarako testuen aurreprozesaketa egiteko Ixa taldean garatutako analisi-katea erabiltzen da analisi morfologikoa, analisi sintaktikoa, entitate izendunak eta chunkak eskuratzeko. Bestalde euskararako sortutako aipamen-detektatzaile automatikoak (Soraluze et al., 2017b) itzulitako aipamenak erabiltzen ditu. Sistema euskarara egokitu, eta EPEC corpusaren korrereferentziarako azpi-atalean (45.000 hitz) aplikatzen da 53,72 puntuko *F-measure* balioa lortuz *CoNLL* metrikan aipamen automatikoekin eta 72,42 puntukoa urrezko aipamenekin.

Azkekin, Soraluze et al. (2017b) lanean, Wikipediatik erauzitako ezaugarri semantikoak gehitu zaizkio aurreko paragrafoan aipatutako sistemari. *CoNLL* metrikan 54,21 puntuko *F-measure* balioa lortu da eta urrezko aipamenekin berriz, 73,94 puntuko *F-measure* balioa.

4. KAPITULUA

Garapena

Aurkibidea

Kapitulu honetan euskarazko testuetan korreferentzia-ebazpenerako neurona-sareetan oinarritutako sistema aurkezten da. Lehenik, erabilitako corpusa aipatuko da, ondoren garatutako sistemaren diseinua aurkeztuko da, aurreprozesaketa eta neurona-sarearen konfigurazioa azalduz, eta azkenik, sistema hobetzeko aldaketa batzuk proposatuko dira.

4.1 Corpusa

EPEC corpusa euskarazko 300.000 hitzez osatutako corpusa da, maila ezberdinetan (morfologikoki, sintaktikoki, esaldi mailan, eta abar) eskuz anotatua dago eta erreferentziazko corpusa da hizkuntzaren prozesamenduko tresnak garatzeko garaian. Corpusa *Euskaldunon Egunkaria* euskarazko egunkaritik hartutako albisteek osatzen dute. Berriki, aipamen eta korreferentzia erlazioak ere anotatuak izan dira, 45.000 hitzez osatutako EPEC corpusaren azpimultzo batean (EPEC-KORREF[1], (Ceberio et al., 2018)). Horretarako, lehenik aipamen-detektatzaile automatikoaren (Soraluze et al., 2017b) bitartez lortutako aipamenak eskuz zuzendu ziren, eta gero aipamen horiek korreferentzia klusterretan sailkatu ziren. Corpusa anotatzeko prozesua *MMAX2* anotaziorako tresna (Müller and Strube, 2006) erabiliz egin zen.

GrAL honetan erabiliko den corpusa hiru zatitan banatuta dago: zati bat entrenatzeko (ikasketa), beste bat garapenean probak egiteko (garapena), eta azkenik ebaluaziorako (ebaluazioa). Ondorengo taulan ikus daitezke banaketa horri buruzko datu zehatzak:

	Hitzak	**Aipamenak**	**Klusterrak**	**Singletonak**
Ikasketa	23.520	6.525	1.011	3.401
Garapena	6.914	1.907	302	982
Ebaluazioa	15.949	4.360	621	2.445
GUZTIRA	46.383	12.792	1.934	6.828

4.1 Taula: EPEC-KORREF corpusaren banaketa

Corpusaren hitzen % 50a ikasketarako da, % 15a garapenerako eta % 35a ebaluaziorako, aipamenak ere proportzio berdinean banatuta daudelarik. Aipamenak klusterretan (edo multzoetan) eta *singletonetan* (korreferentzia-erlaziorik ez duten aipamenak) bereizten dira. Gutxi gorabehera aipamenen laurdenek osatzen dute klusterren bat, eta gainontzekoak *singletonak* dira.

[1] http://ixa.si.ehu.es/node/4487

Lan hau egiteko oinarri gisa hartu den polonierarako sisteman erabili den corpusak 540.000 token eta 180.000 aipamen ditu, euskararako eskuragarri dagoen corpusa baino 10 aldiz handiagoa. Ingelesezko sistemek erabiltzen dituzten copusak, berriz, miloi bat hitzetik gorakoak dira.

4.2 Euskararako korreferentzia-ebazpena neurona-sareekin

Neurona-sareetan oinarritutako euskararako korreferentzia-sistema hau diseinatzeko, Polonierarako sortutako sistema (Niton et al., 2018) hartu da oinarritzat. Ingeleserako arlo honetan lan gehiago egon arren, lan hori aukeratu da bi hizkuntzek, polonierak eta euskarak dituzten puntu amankomunengatik. Biak hizkuntza eranskariak dira, ordena librekoak eta polonieraren baliabideak, euskararakoak bezala, ez dira ingelesekoak bezain handiak. Bestalde, polonierarako sortutako eredua ingeleserako erabili diren gehienak baino xumeagoa da, eta errazagoa behar luke euskarara moldatzeko.

Aipamen-detekzioari dagokionez, Soraluze et al. (2017b) lanaren emaitza den euskararako aipamen-detektatzaile automatikoa eta urre-patroia erabiliko dira, aipamen-detekzioaren atazak emaitzetan duen eragina neurtzeko. Euskararako korreferentzia-ebazpenean, erregelatan oinarritutakoan zein ikasketa automatikokoan, aipamen-detektatzaile automatikoa eta urre-patroia erabili dira, eta orain neurona-sareekin probatuko dira, teknika ezberdinekin lortutako emaitzak konparatzeko.

Polonierarako sortu den neurona-sareetan oinarritutako korreferentzia-ebazpenerako sistemak emaitza hobeak eskuratu ditu aipamen-bikote eredua baliatuz, entitate-aipamen ereduarekin baino (Niton et al., 2018). Horregatik, eta korreferentzia-ebazpeneko hainbat sistemaren abiapuntua denez (Niton et al., 2018; Clark, 2015), aipamen-bikote eredua aukeratu da euskararako eraiki den sistemarako.

Testu bat osatzen duten aipamenen artean, askoz gehiago dira korreferentzia erlaziorik ez duten aipamenak elkarren artean korreferente diren aipamenak baino (Wiseman et al., 2015). Egoera honek, aipamenaren buruak, numeroak, eta generoak bat egiten duten korreferentzia-erlaziorik gabeko aipamenenen eta benetan korreferenteak diren aipamenen artean bereizi ahal izatea zailtzen du.

Testu batean dauden aipamen-bikote posible guztietatik gutxiengoak direnez korreferenteak, ikasketa garaian ere antzeko proportzioak mantentzea komeni da (Niton et al.,

2018). Sailkatzailea entrenatzeko garaian bikote negatiboak gutxitze aldera, metodo ezberdinak aplikatu izan dira, nabarmenenak Soon et al. (2001) eta berriki Sapena et al. (2011) izanik. Ikasketa atalean, aipamen bikoteak Soon et al. (2001) lanean proposatutako algoritmoaren arabera sortu dira. Bertan aipamen bakoitza bere aurrekariarekin lotzen da bikote korreferenteak lortzeko, eta aipamen bakoitza eta bere aurrekariaren arteko aipamen guztiak hartzen dira negatiboen sorkuntzarako. Horrela, 24.000 instantzia sortu dira ikasketarako, 2.000 aipamen-bikote korreferente eta 22.000 aipamen-bikote ez-korreferente.

Garapena eta ebaluazioa ataletarako berriz, aurrekariak zein diren jakiteko informaziorik eskuragarri ez dugunez izango, hainbat kopururekin probatu da, eta azkenean Goenaga et al. (2012) lanean proposatutako aipamen bikoteen sorkuntza erabili da, aipamen bakoitzarentzat bere aurreko 8 aipamenak hartuz aurrekari izateko hautagaitzat. Aipamen baten aurrekaria topatzen da 8 aipameneko distantzian kasuen % 97an.

4.2.1 Sarrerako ezaugarriak

Aipamen bikoteak sortuta ditugunean, aipamen horien ezaugarriak erauzi behar dira, neurona-sarea entrenatzeko sarrera osatuko duten zenbakizko balioak sortzeko.

Neurona-sarea entrenatzeko polonierarako sistemaren (Niton et al., 2018) oinarrilerroa hartu da eredutzat, ahal izan den neurrian, bertako sarrerako ezaugarrien parekideak mantenduz. Ezaugarri multzo horietan aipamen bakoitzaren ezaugarriak eta aipamen-bikotearenak bereizi daitezke. Aipamen bakoitzeko ezaugarriak hitz-embeddingez eta ezaugarri bitarrez daude osatuak; aipamen-bikote bakoitzeko ezaugarriak, berriz, distantzia eta bestelako ezaugarri bitarrez.

Hitz-embedding horiek lortzeko, euskarazko Wikipedia (2016/04/07koa) eta Elhuyar Web Corpusa erabili dira (Goikoetxea et al., 2018), guztira, 160 miloi token. Corpus horretako hitzen erroak erauzi eta word2vec-eko skip-gram erabiliz (hitz bat emanik testuingurua lortzeko baliatzen da) kalkulatu dira 50 dimentsiotako hitz-embeddingak (-1 eta 1 arteko zenbakiez osatuak).

Bikotea osatzen duten aipamen bakoitzeko ondorengo ezaugarriak erauzi dira:

- Aipamenaren buruaren 50 dimentsiotako hitz-embeddinga (50 dimentsio).
- Aipamenaren lehen hitzaren hitz-embeddinga (50 dimentsio).

- Aipamenaren azken hitzaren hitz-embeddinga (50 dimentsio).
- Aipamenaren aurreko bi hitzen hitz-embeddingak (2 x 50 dimentsio).
- Aipamenaren ondorengo bi hitzen hitz-embeddingak (2 x 50 dimentsio).
- Aipamenaren aurreko 5 hitzen hitz-embeddingen batazbestekoa (50 dimentsio).
- Aipamenaren ondorengo 5 hitzen hitz-embeddingen batazbestekoa (50 dimentsio).
- Aipamenaren osatzen duten hitzen hitz-embeddingen batazbestekoa (50 dimentsio).
- Aipamena azaltzen den esaldia osatzen duten hitzen hitz-embeddingen batez bestekoa (50 dimentsio)
- Aipamenaren mota zehazteko 4 ezaugarri bitar; aipamena nominala, pronominala, zero-motakoa (*zero-type*, elipsearekin lotuta) edo bestelakoa den adierazteko (4 ezaugarri).

Aipamen-bikoteko honako ezaugarriak:

- Bi aipamenen arteko distantzia hitzetan. Distantzia 11 ezaugarri bitarretan kodetzen da, ondorengo multzoetako batean sartuz [0,1,2,3,4,5-7,8-15,16-31,32-64,64+,ez jarraitua] (11 ezaugarri).
- Bi aipamenen arteko distantzia aipamenetan. Distantzia 11 ezaugarri bitarretan kodetzen da, ondorengo multzoetako batean sartuz [0,1,2,3,4,5-7,8-15,16-31,32-64,64+,ez jarraitua] (11 ezaugarri).
- Aipamenek elkar ebakitzen dute (ezaugarri bitar 1).
- String-parekaketa (ezaugarri bitar 1).
- Lema-parekaketa (ezaugarri bitar 1).
- Buru-parekaketa (ezaugarri bitar 1).
- Aipamenak esaldi berean daude (ezaugarri bitar 1).
- Aipamenak paragrafo berean daude (ezaugarri bitar 1).
- Aipamenenetako bat bestearen akronimoa da. Euskararen kasuan aipamenak elkarren siglak diren soilik aztertu da (ezaugarri bitar 1).

- Aurrekariak aipameneko hitz-arraroena du barnean. Hitz arraroena zein den finkatzeko, hitz-embeddingak kalkulatzeko erabili den stemmerra pasatutako corpusa hartu eta hitzen maiztasunak kalkulatu dira (ezaugarri bitar 1).

- Aipamenek generoan bat egiten duten. 3 ezaugarri bitar adieraziz ea aipamenek genero maskulinoan, genero femeninoan edo neutroan bat egiten duten (3 ezaugarri bitar).

- Aipamenek numeroan bat egiten duten. 3 ezaugarri bitar adieraziz ea aipamenek numero singularrean, pluralean edo mugagabean bat egiten duten (3 ezaugarri bitar).

- Aipamenek pertsonan bat egiten duten. 3 ezaugarri bitar adieraziz ea izenordainak izanik, aipamenek lehenengo, bigarrengo edo hirugarrengo pertsonan bat egiten duten (3 ezaugarri bitar).

Ezaugarri gehienak eskuratu ahal izan dira euskarazko corpusetik, baina honako ezaugarrien kasuan ez da horrela izan:

- Aipamen mota zehazteko poloniarerako lau ezaugarriak erabiltzen dituzten arren, euskararako lehenengo bi ezaugarriak soilik erauzi dira. Zero-motakoa edo bestelakoa den adierazten duten ezaugarrietan 0ko batekin osatu dira, corpusean ez dugulako horrelako informaziorik etiketatuta.

- Corpusa osatzen zuten paragrafoei buruzko informaziorik ez zegoenez eskuragarri, 0ko bat finkatu da paragrafo berean daudela adierazten duen ezaugarrian.

- Euskarak genero gramatikalik ez duenez kasu guztietan genero neutroan bat egiten dutela adierazi da.

Honenbestez, guztira 1.147 dimentsiotako (aipamen bakoitzeko 554, eta bikoteko 39) bektorea sortu da aipamen bikoteko. Neurona-sareari 1.147 dimentsiotako bektorearekin batera, aipamen-bikotea korreferentea den edo ez (1 edo 0) adierazten duen klasea pasatzen zaio sarreran entrenatzeko. Ebaluazio garaian, sarreran aipamen bikote baten 1.147 dimentsiotako bektorea hartzen du, eta 0 eta 1 arteko zenbaki bat itzultzen du (korreferenteak izateko probabilitate gisa har genezake).

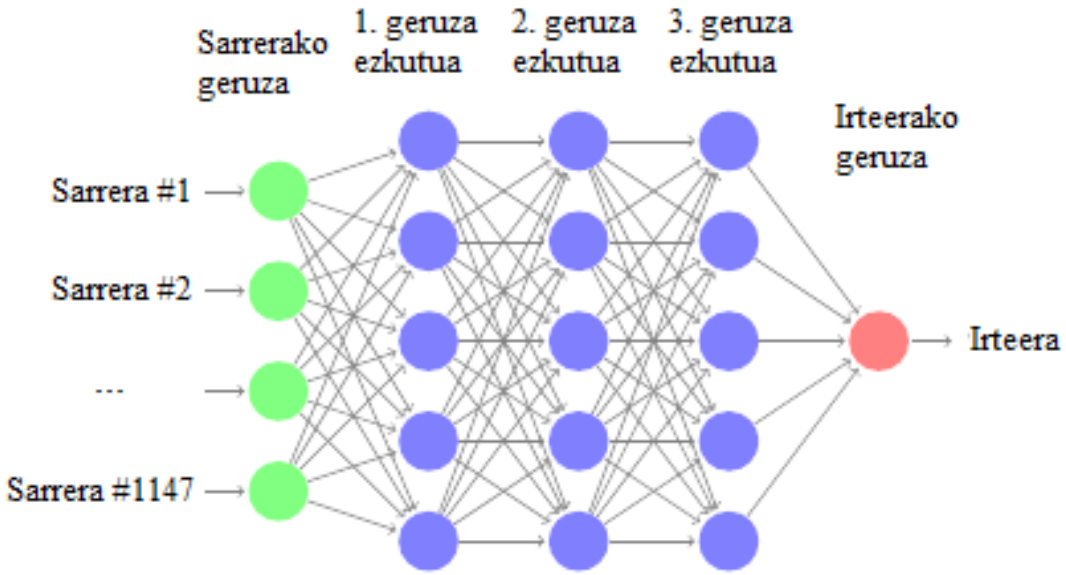
4.2.2 Neurona-sarearen diseinua

Neurona-sare artifizialak (*artificial neural networks*) edo neurona-sareak, garuneko neurona-sare biologikoetan inspiratuta dauden konputazio sistemak dira. Ezagutza gehitzeko erregelarik programatu beharrik gabe, adibideak emanik, ataza ezberdinak burutzen ikasteko gai dira neurona-sare artifizial hauek. Adibidez, gai dira irudi batean objektu edo animaliak detektatzeko, eskuz anotatutako irudietatik ikasiz.

Neurona-sare artifizialak, elkarrekin konektatutako neurona artifizialen (neuronen) multzoez osatuta daude. Neurona artifizial bat gai da jasotzen duen seinalea prozesatu eta konektatuta dagoen beste neurona batzuei seinaleak bidaltzeko. Konektatutako neuronen arteko seinalea zenbaki erreal bat izan ohi da, eta neuronaren sarreren baturaren funtzio ez linealekin kalkulatzen da neuronaren irteera. Neuronen arteko konexioei ertzak deritze, eta hauek pisu bat dute, pisua egokituz neuronen arteko konexioaren indarra finkatzen da. Sareak dagokion ataza ikasteko, unean duen konfigurazioa ikasketarako adibideekin probatu eta pisuak doitzen joaten da.

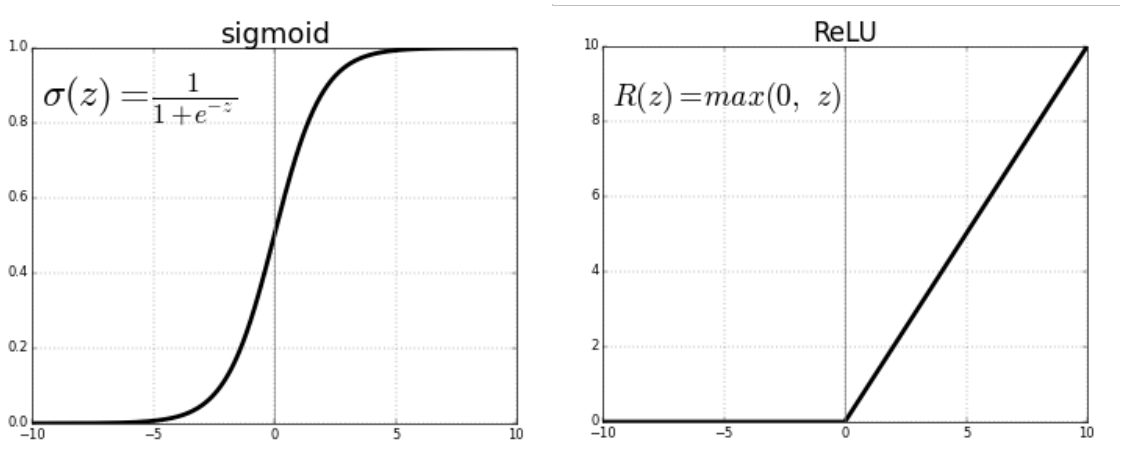
Normalean, neuronak geruzetan antolatzen dira, geruza bakoitzak sarrerari transformazio ezberdin bat egiten diolarik. Geruza hauen artean, ezkutuko geruzak (*hidden layers*) eta irteera geruza (*output layer*) bereizten dira. Ezkutuko geruzak sarrera eta irteeraren artean daudenak dira, kutxa beltz moduan funtzionatzen du, eta geruza kopurua eta geruza bakoitzean dauden neurona kopurua sarearen arabera aldatzen da, baina ohikoena geruza bakoitzetik ondorengora neurona kopurua txikitzea da. Irteerako geruza neurona batek edo gutxi batzuk osatzen dute, eta hauek emaitza itzultzen dute.

Lan honetan, 3 ezkutuko geruza dituen sare trinko bat (*fully connected network*) erabili da. Sareak lehenengo ezkutuko geruzan 500 neurona ditu, bigarrenean 300 eta hirugarrengoan 100. Irteerako geruzan, berriz, neurona bakarra du. Neurona-sareak sarrera geruzan 1.147 dimentsiotako instantziak hartu eta 0 eta 1 arteko zenbaki bat itzultzen du irteerako geruzan. Neurona-sarearen arkitekturaren eskema 4.1 irudian ikus dezakegu.



4.1 Irudia: Neurona-sarearen arkitekturaren eskema

Neurona-sareko neurona bati sarrera multzo bat emanik, neurona horren irteera definitzen du aktibazio-funtzioak. Aktibazio-funtzio ez linealek, jasotako sarreren arabera irteera normalizatu bat itzultzen dute (0 eta 1 artean adibidez), neurona "aktibatu"den edo ez adieraziz. Lan honetan, ezkutuko geruzetan *ReLU* (rectified linear unit) aktibazio-funtzioa erabili da, eta irteerako geruzan *sigmoid* funtzioa (4.2 irudia), 0 eta 1 arteko emaitza itzul dezan.



4.2 Irudia: ReLU eta sigmoid funtzioak

Beraz, neurona-sarearen ekuazioak honakoak dira, geruza bakoitzean dagokion

aktibazio-funtzioa aplikatuz:

$$\text{Sarrera bektorea: x} = [e_i, e_j, e_{ij}]$$

$$\text{1. ezkutuko geruza: h}^1 = \text{RELU}(W^1 x + b^1)$$

$$\text{2. ezkutuko geruza: h}^2 = \text{RELU}(W^2 h^1 + b^2)$$

$$\text{3. ezkutuko geruza: h}^3 = \text{RELU}(W^3 h^2 + b^3)$$

$$\text{Irteerako geruza: p(i,j) = sigmoid}(\text{w}^T h^2)$$

e_i eta e_j aipamen bakoitzaren ezaugarriak, e_{ij} aipamen bikotearen ezaugarriak eta W pisuak izanik.

Neurona-sarea entrenatzeko, hau da, neuronen arteko pisuak bilatzeko, galera-funtzioa (*loss function*) minimizatzen da. Galera-funtziorako entropia gurutzatu bitarra (*binary-cross entropy*) aukeratu da. Entrenamendurako 2 *epoch* eta 128 tamainako *mini-batch*-ak finkatu dira. Galera-funtzio eta balio hauek polonierarako sisteman erabiltzen direnak dira; geroago *epoch* kopurua eta *mini-batch*-aren tamaina parametroak doituko dira.

*Epoch*ak ikasketarako datu multzoan egindako iterazio kopurua adierazten du, *mini-batch*aren tamainak berriz, sareari aldiko pasatuko zaion instantzia kopurua.

Galera minimizatzeko *Adam* optimizazioa aplikatu zaio (Kingma and Ba, 2014). Ikasketa sakonean pisuak eguneratzeko garaian *Adam* algoritmoa ohikoa bilakatu da, ikasketa azkartzen duelako emaitza onetara lehenago iritsiz.

Ezkutuko geruza bakoitzari *batch* normalizazioa aplikatu zaio (Ioffe and Szegedy, 2015). *Batch* normalizazioak neurona bakoitzaren aktibazio-funtzioaren irteera normalizatzen du, muturreko pisuen balioen eragina txikitzeko eta ikasketa azkartuz *mini-batch* bakoitzean.

Neurona-sarea 0,2ko ratioko *dropout*-a erabiliz doitu da (Srivastava et al., 2014). *Dropout*-a aplikatzean neurona batzuk ausaz desaktibatzen dira (0 balioa emanez) ikasketa garaiarako soilik. Hau eginez gehiegizko egokitzea (*overfitting)* saihestea lortzen da.

Neurona-sarearen modeloa *KERAS*ekin (Chollet et al., 2015) inplementatu da, azpitik *TENSORFLOW* (Abadi et al., 2016) darabilerarik. Sarearen Arkitektura eta optimizazio teknikak Niton et al. (2018) artikuluan erabilitako berdinak dira eta parametroak hasieratzeko, polonierarako emaitza onenak eman zizkietenak erabili dira.

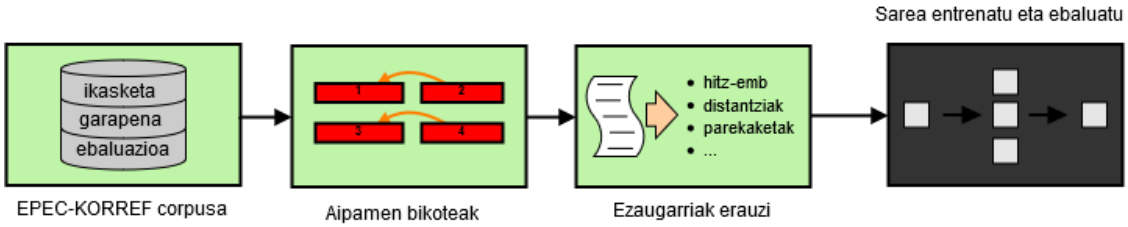
4.2.3 Sistema

Neurona-sareetan oinarrituta euskararako eraiki den korreferentzia-ebazpenerako sistema azalduko da jarraian. Sistema hiru atal nagusitan banatzen da, aurreprozesaketa, neurona-sarearen entrenamendua eta ebaluazioa.

Aurreprozesaketa atazaren abiapuntua, EPEC-KORREF corpusa da. Corpus honek euskarazko testu multzo batean korreferentzia erlazioak ditu anotatuta, korreferentzia-ebazpenerako baliagarriak izan daitezkeen beste ezaugarri askorekin batera. Corpusak ikasketa, garapena eta ebaluazioa azpi-atalak dakartza aurrez finkatuta.

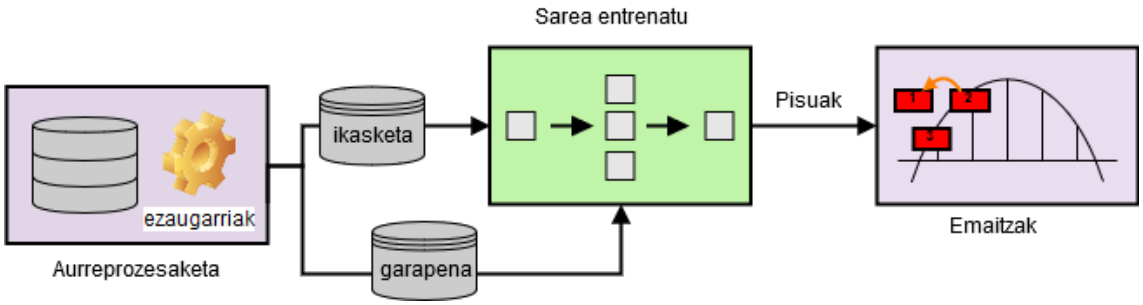
Aurreprozesaketa atalean, corpuseko aipamenak hartu eta bikoteak sortzen dira (aipamen-bikote eredua erabilita), corpusaren azpi-atal bakoitzean negatiboen sorkuntzarako dagokion metodoa aplikatuz. Ikasketa atalean aipamen bakoitza bere aurrekariarekin (bikote positiboa) eta aipamenaren eta aurrekariaren artean daudenekin (bikote negatiboak) parekatzen da. Garapena eta ebaluazioa ataletan, berriz, aipamen bakoitza aurreko zortziekin parekatzen da.

Aipamen bikoteak sortuta ditugunean, korreferentzia-ebazpenerako baliagarriak diren ezaugarriak erauzi dira instantzia bakoitza sortzeko. Ezaugarri horiek (hitz-embeddingak, distantziak eta parekaketak nagusiki) 4.2.1 atalean daude bilduta. Ikus aurreprozesaketa laburbiltzen duen 4.3 irudia.



4.3 Irudia: Aurreprozesaketa atala

Aurreprozesaketa egin ondoren, prest ditugu neurona sareak ikasketarako behar dituen instantziak. Instantzia bakoitza 1147 dimentsiotako bektorea da, eta klaseak (0 edo 1) aipamen bikotea korreferentea den edo ez adierazten du. Neurona-sarea entrenatzeko ikasketa atala erabili da, eta barne balidaziorako (ikasketa prozesuaren parte) garapena atala. Ikasketaren emaitza korreferentzia-ebazpenean entrenatutako neuronen arteko pisuak dira, aipamen-bikote korreferenteak diren erabakitzeko sareak erabiliko dituenak. Ikus 4.4 irudia.

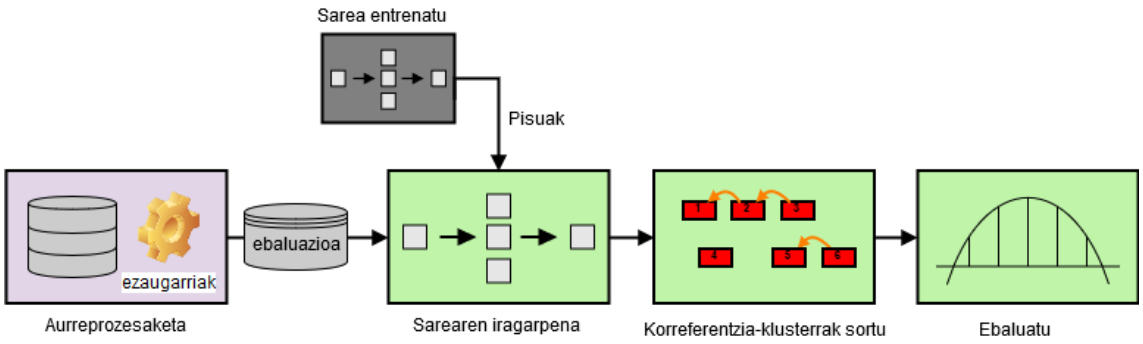


4.4 Irudia: Neurona-sarearen entrenamendua

Neurona-sarea entrenatuta, ebaluazioa ataleko instantzientzat iragarpen bat bueltatzen du, 0 eta 1 arteko balio bat, aipamen bikoteak korreferente diren edo ez erabakitzeko. Behin aipamen bikoteak eta sistemaren iragarpenak izanik, aipamenak korreferentzia klusterretan banatu behar dira. Sistemak korrefente izateko probabilitate altua eman dioten aipamenak multzo berean sartuko dira, betiere atalase (*treshold*) bat gainditzen badute.

Atalasea finkatzeko corpusaren ikasketa atala erabili da entrenatzeko eta balidaziorako (% 90 eta % 10 bakoitzerako) eta garapena atala erabili da sarea ebaluatzeko. Proba ezberdinak egin dira 0,5, 0,75, 0,85, 0,90, 0,95, 0,96, 0,97, 0,98, 0,99, 0,995 eta 0,999 balioentzako, eta emaitzarik onenak 0,99ko atalasearekin lortu dira (0,90ekin ere emaitza onak lortu direlarik). Proben emaitzen arabera, 0,99an finkatu da atalasea.

Atalasea gainditzen duten aipamen bikoteak multzokatuta, korreferentzia-ebazpena bukatutzat ematen da. Ondoren, lortutako emaitzak ebaluatu dira ebaluatzaile ofiziala (Pradhan et al., 2014) baliatuz, korreferentzia-ebazpenean ohikoa den moduan. Ikus 4.5 irudia.



4.5 Irudia: Sistemaren ebaluazioa

Aurretik atalka azaldu den korreferentzia-ebazpenerako sistemaren eskema, bere osotasunean, 4.6 irudian ikus daiteke.

Hemen azaldutako sistemak CoNLL metrikan % 54,47 puntuko *F-measure*a lortu du urrezko aipamenekin eta % 40,91 puntukoa aipamen automatikoekin. Erabilitako metriken azalpena eta lortutako emaitza guztiak 5 kapituluan daude bilduta. Lortutako emaitzak hobetzeko asmoarekin jarraian dauden proba ezberdinak egin dira.

4.3 Hobekuntzak

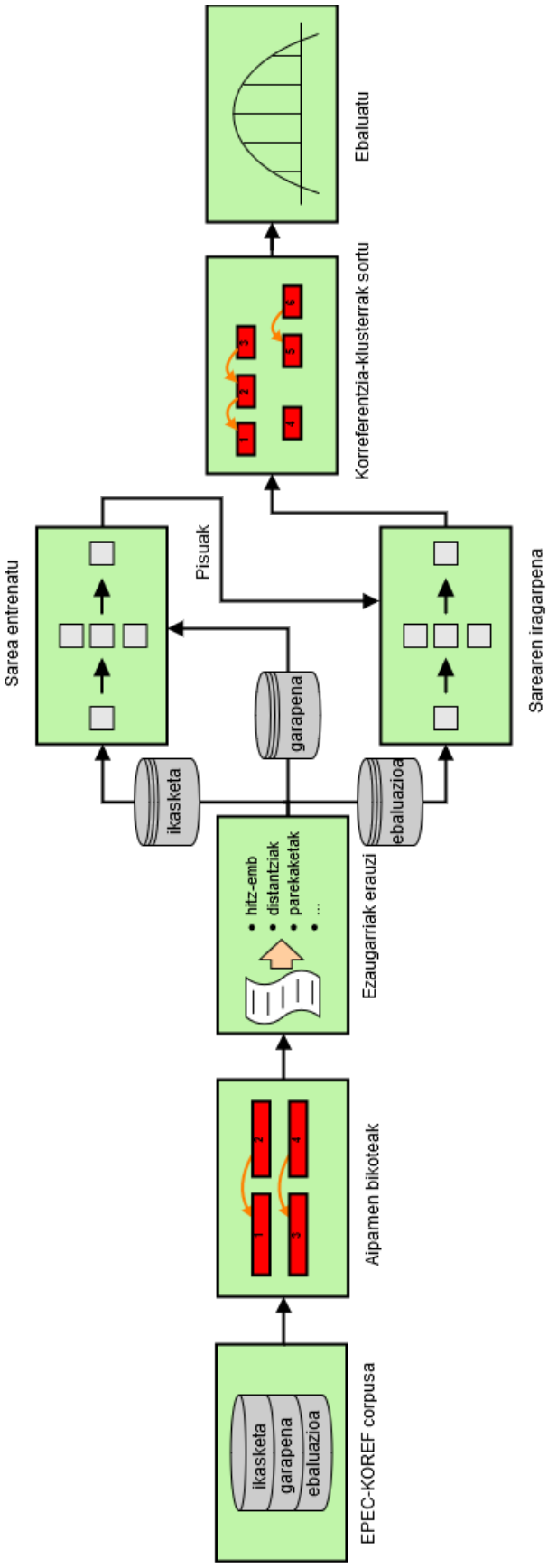
Atal honetan, neurona-sareetan oinarritutako euskararako korreferentzia-ebazpenerako sistema hobetzeko proposamen ezberdinak aurkeztuko dira. Hobekuntzak hiru multzotan bereizi ditzakegu: hitz-embeddingen dimentsioak handitzea, ezaugarri berriak gehitzea eta neurona-sarearen parametroak egokitzea.

4.3.1 Hitz-embeddingak

Lehenik, sistemak ikasketarako erabiltzen dituen hitz-embedding ezberdinak probatu dira, emaitzarik onenak zeinek ematen dituen konparatzeko. Horretarako, dimentsio kopuru ezberdinak probatu dira, oinarrizko sistema 50 dimentsiotako hitz-embeddingekin entrenatu da, polonierarako oinarri-lerroan ala erabiltzen delako, baina ondoren 100, 200 eta 300 dimentsiotako hitz-embeddingekin esperimentatu da. Neurona-sareetan oinarritutako korreferentzia-ebazpenerako sistema gehienek 50 edo 300 dimentsiotakoak erabiltzen dituzte, eta hitzen antzekotasun semantikoa lantzeko atazan 300 dimentsio da estandarra, handik gorako dimentsioetan hobekuntza nabarmenik ikusten ez delako.

Bestalde, lehen diseinuan hitz-embedding-ak ikasteko corpuseko hitzen erroak erabili dira, antzekotasun-semantikoa erauzteko emaitza hobeak ematen baitituzte. Hala ere, euskara hizkuntza eranskaria izanik, corpuseko hitzen erroak erabiltzean atzetik dituzten atzizkiak ezabatzen dira, eta hauek korreferentzia-ebazpenerako informazio baliagarria izan dezakete. Horregatik aurrez zehaztu den dimentsio bakoitzerako (50, 100, 200 eta 300), corpuseko hitzen erroez gain, corpusa bere horretan (hitzen formekin) ere probatu da, emaitzarik onenak emango dituen konbinazioaren bila.

Egindako proba horien emaitza guztiak 5 kapituluan daude ikusgai (Ikus 5.2 eta 5.3 taulak). Orokorrean, hitz-embeddingen dimentsio kopurua handituz ez da lortu emaitzak



4.6 Irudia: Korreferentzia-ebazpenerako sistemaren eskema

hobetzerik, eta antzekotasun semantikorako bezala, korreferentzia ebazpenerako ere, corpuseko erroak erabiliz emaitza hobeak lortu dira, nahiz eta aldea oso txikia izan. Emaitzarik onenak 50 dimentsio eta ikasketarako erroak erabilita lortu dira, CoNLL metrikan % 54,47 puntuko *F-measure*a urrezko aipamenekin eta aipamen automatikoekin % 40,91 puntukoa. Horregatik hemendik aurrera egiten diren probetan 50 dimentsioko eta hitzen erroekin lortutako hitz-embeddingak erabiliko dira.

4.3.2 Ezaugarrien gehitzea

Behin hitz-embeddingak finkatuta, aurreprozesaketan aipamen bikoteen ezaugarri batzuk aldatu dira, emaitzak hobetu ditzaketela iritzita.

Lehenik, polonierarako sisteman erabiltzen diren, baina eskuragarri ez zeuden ezaugarriak, balio finko batean zeudelako, kentzea erabaki da, sistemari ez diotelako informazio gehigarririk gehitzen. Ezaugarri horiek, aipamen bakoitzaren aipamen mota adierazten duten zero-mota (*zero-type*) eta bestelakoa, eta aipamen bikote bakoitzeko generoari lotutako 3 ezaugarriak eta paragrafo berdinean daudela adierazten dutenak dira.

Bestalde korreferentzia-ebazpenerako baliagarriak izan daitezkela dirudien ezaugarri batzuk gehitu dira. Oinarri-lerroan, numeroa eta pertsona ez daude aipamen bakoitzeko adierazita, aipamen bikotearen ezaugarri moduan baizik, ea numeroan eta pertsonan bat datozen adieraziz. Orain berriz, aipamen bakoitzari gehitu zaizkio numeroa eta pertsona ezaugarriak (aipamen bakoitzeko 6 ezaugarri), aipamen bikotearen numero eta pertsonan bat datozen adierazten duten ezaugarriak mantentzeaz gain.

Azkenik, ezagutza semantikoa erabiltzeko garaian hitz-embeddingak soilik hartu ordez, Wikipediatik erauzitako informazioa gehituta probatu da. Horretarako, corpusean Wikipedian ageri diren entitateen aliasak hartu eta aipamen bikoteko 3 ezaugarri erauzi dira:

- Wikipediako aliasak ea berdinak diren.
- Lehen aipamenaren Wikipediako aliasa, bigarren aipamenaren lematizatuaren barne dagoen.
- Bigarren aipamenaren Wikipediako aliasa, lehengo aipamenaren lematizatuaren barne dagoen.

Ezaugarri gehigarriekin probatutako sistemaren emaitzak 5 kapituluan daude ikusgai. Lortutako emaitzak aztertuz, esan genezake oinarri-lerroaren antzeko emaitzak lortu direla. CoNLL metrikan % 54,27 puntuko *F-measure*a lortu da urrezko aipamenekin eta aipamen automatikoekin % 40,75 puntukoa, oinarri-lerrokoaren azpitik, baina beste metrika batzuetan emaitza hobeak eman ditu. Ikus 5.4 taula.

4.3.3 Sarearen parametroen doitzea

Sarearen parametroen artean, *epoch* kopurua eta *mini-batch*en tamaina ezberdinak probatu dira emaitzetan hobekuntzarik dagoen ikusteko. *Mini-batch*aren 3 tamaina probatu dira: 64, 128 eta 256; *mini-batch*aren tamaina bakoitza 1, 2, 4, 8 epoch-ekin probatu da. *Epoch* kopurua sareak ikasketarako instantzietan egiten dituen iterazio kopurua da, eta *mini-batch*aren tamainak sareak aldiko ikasketarako hartuko dituen instantzia kopurua adierazten du. Bi parametroen kasuan, hasieran finkatutako balioak handitu eta txikitu egin dira, parametroen balio egokienak topatu nahian. Ez da bilaketan sakondu, bi parametroen kasuan ez delako hobetzeko joerarik antzeman. Bi *epoch* eta 64 tamainako *mini-batch*arekin CoNLL metrikan % 54,66 puntuko *F-measure*a lortu da urrezko aipamenekin eta % 41,20 puntukoa aipamen automatikoekin. Emaitza guztiak 5.5 eta 5.6 tauletan daude bilduta.

5. KAPITULUA

Emaitzak

Aurkibidea

5.1 Ebaluaziorako metrikak

Korreferentzia-ebazpenean, sistema automatikoak itzulitako erantzunak urrepatroiarekin konparatzen dira. Korreferentzia-ebazpenerako sistemen kalitatea neurtzeko, eta sistema ezberdinen arteko konparaketa egin ahal izateko ebaluazio metrika ezberdinak erabili dira azken hamarkadetan. Berriki metrika berriak proposatu dira, aurretik proposatutakoen gabeziak konpontzeko.

Gaur egun, korreferentzia-ebazpeneko sistemak ebaluatzeko honako metrika hauek erabiltzen dira: MUC (korreferentzia-loturetan oinarritua, Vilain et al. (1995)), B^3 (aipamenetan oinarritua, Bagga and Baldwin (1998)), CEAFe (ϕ antzekotasuna entitateetan, Luo (2005)) eta CEAFm (ϕ antzekotasuna aipamenetan, Luo (2005)), BLANC (korreferentzia-loturak eta ez-loturak erabiliz, Recasens and Hovy (2011)) eta LEA (entitateak eta hauen garrantzia erabiliz, Moosavi and Strube (2016)).

MUC, B^3 eta CEAFe neurrien batezbesteko aritmetikoa den CoNLL neurria ere (Denis and Baldridge, 2009) ohikoa da korreferentzia-ebazpenerako sistemen kalitatea neurtu eta konparatzeko.

Neurona-sareak itzulitako emaitzekin korreferentzia multzoak osatu ondoren, korreferentzia-ebazpenean erabiltzen diren metrika nagusi horiek biltzen dituen errefe-rentziazko tresna (Pradhan et al., 2014) erabili da sistema ebaluatzeko.

5.2 Lortutako emaitzak

Ondorengo tauletan laburbiltzen dira proiektu honetan korreferentzia-ebazpena atazan neurona-sareetan oinarritutako sistemak lortutako emaitzak.

5.1 taulan ikus daitezke euskararako neurona-sareetan oinarritutako lehen korreferentzia-ebazpenerako sistemaren emaitzak. Lehenengo zutabean aipamen detekzioaren ehunekoa ematen da (urre patroia erabiliz % 100 eta aipamen automatikoekin %73,79). CoNLL metrikan % 54,47 puntuko *F-measure*a lortu da urrezko aipamenekin eta % 40,91 puntu aipamen automatikoekin.

5.1 Taula: oinarri-lerroaren emaitzak, urre-patroia eta aipamen-detekzioa automatikoarekin (hitzen erroetatik ikasitako 50 dimentsiotako hitz-embeddingekin).

	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
urre	100	10,87	79,48	67,70	73,07	52,12	48,02	54,47
auto	73,79	8,73	58,33	53,15	55,66	29,30	35,49	40,91

Urrezko aipamenekin B^3 metrikan % 79,48 puntu lortu ditu eta aipamen automatikoekin % 58,33 puntu, bi balioen arteko aldea handia izanik. B^3 metrika aipamenetan oinarritutako metrika dela esaten da, korreferentzia-ebazpena baloratzeko zuzen multzokatutako aipamenak hartzen baiditu kontutan. Aipamen detektatzaile automatikoak aipamenen % 73,79a soilik identifikatzen ditu eta aipamenen laurdena baino gehiago falta izateak eragin handia du B^3 bezalako metrika batean.

5.2 taulan polonierako sistema oinarri hartuz eta hitz-embedding ezberdinekin probatuz lortu diren emaitzak biltzen dira. 50, 100, 200 eta 300 dimentsiotako hitz-embeddingak erabili dira, hitzen erroak (err) eta hitzen formak (for) zituzten corpusetatik ikasiak. Hitz-embeddingen dimentsio zehatz bat gailentzen ez den arren (metrikaren arabera ezberdinak dira), orokorrean, hitzen erroetatik ikasitako eta 50 dimentsiotako hitz-embeddingekin lortu dira emaitzarik onenak, CoNLL metrikan % 54,47 puntuko *F-measure*a lortuz urrezko aipamenekin eta % 40,91 puntu aipamen automatikoekin.

5.2 Taula: Aipamen automatikoak erabiliz, sistemaren emaitzak hitz-embedding ezberdinentzat. Tartean urrezko aipamenekin lortutako emaitza onena sartu da.

auto	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
50 err	100	10,87	79,48	67,70	73,07	52,12	48,02	**54,47**
50 err	73,79	8,73	58,33	53,15	55,66	29,30	35,49	**40,91**
100 err	73,79	**8,82**	58,16	52,84	55,40	29,27	35,08	40,79
200 err	73,79	6,63	59,06	53,63	56,25	**29,73**	37,39	40,64
300 err	73,79	0,15	**60,30**	**53,91**	**56,46**	27,20	**40,73**	38,97
50 for	73,79	6,55	58,90	53,47	56,13	28,69	37,12	40,53
100 for	73,79	6,43	59,07	53,63	56,30	28,68	37,69	40,60
200 for	73,79	7,48	58,61	53,19	55,85	28,92	36,46	40,64
300 for	73,79	0,58	60,07	53,75	56,43	27,28	40,11	39,02

Hitz-embeddingen dimentsioak handitu ahala, CoNLL metrikan emaitza okerragoak lortu dira, joera hori, baina, ez da metrika guztietan ematen. 300 dimentsiotako hitz-

embeddingetatik ikasiz, korreferentzia erlazio gutxiago identifikatzen ditu sistemak, horregatik MUC metrikan oso emaitza baxuak lortu dira. B^3, CEAF eta LEA metriketan berriz, emaitza hobeak lortu dira, korreferenteak ez diren aipamen oso gutxi lotzen dituelako, hau da, estaldura (R) oso txikia lortzen delako. Atalasea txikitzearekin batera estaldura handitzea lortzen da, baina hori eginik doitasuna (P) dezente jeisten da, CoNLL metrikan emaitza baxuagoak lortuz.

5.3 eta 5.4 tauletan, sistemak ikasketarako erabiltzen dituen ezaugarrien gehitzea eginda (ezaugarri+) lortutako emaitzak ikus daitezke, lehenengo ezaugarri multzoarekin lortutakoen (oinarri-l.) alboan. Taulan ikus daitekeenez, ezaugarri gehiagarriak erabiltzen dituen sistemak emaitza hobeak ematen ditu metrika gehienen kasuan urrezko aipamenentzat zein automatikoentzat. B^3, CEAFm, CEAFe eta LEA metriketan emaitza altuagoak lortu dira, MUC, BLANC eta CoNLL metriketan aldiz baxuagoak. CoNLL metrikan emaitza baxuagoak lortu dira, MUC metrikan oinarri-lerroarekiko izandako jeitsiera nabarmenagatik.

5.3 Taula: Ikasketarako ezaugarrien aldaketarekin emaitzen konparaketa (urrezko aipamenekin).

urre	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
oinarri-l.	100	10,87	79,48	67,70	73,07	52,12	48,02	**54,47**
ezaugarri+	100	8,94	80,20	68,29	73,66	51,55	50,06	54,27

5.4 Taula: Ikasketarako ezaugarrien aldaketarekin emaitzen konparaketa (aipamen automatikoekin).

auto	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
oinarri-l.	73,79	8,73	58,33	53,15	55,66	29,30	35,49	**40,91**
ezaugarri+	73,79	7,69	58,73	53,28	55,82	28,93	36,57	40,75

Azkenik, sarearen parametroak doitze aldera *epoch* kopurua eta *mini-batch*en tamaina ezberdinak konbinatuz lortutako emaitzak daude bilduta 5.5 taulan. Oinarri-lerroan erdietsitako emaitzak baino hobeak lortu dira. Urrezko aipamenekin, emaitzarik onena, *epoch* 1 eta 256 tamainako *mini-batch*a erabiliz lortu da (*CoNLL*n % 54,75 puntuko *F-measure* balioa); aipamen automatikoekin, 2 *epoch* eta 64 tamainako *mini-batch*arekin (*CoNLL*n % 41,20 puntuko *F-measure* balioa). Azken helburua aipamen automatikoekin funtzionatuko duen emaitzak lortzea izanik, 2 *epoch* eta 64 tamainako *mini-batch*arekin entrenatutako neurona-sarea hartu da azken sistematzat.

5.5 Taula: Aipamen automatikoak erabiliz, sistemaren emaitzak epoch eta mini-batch ezberdinentzat. Tartean urrezko aipamenekin lortutako emaitza onena sartu da.

a. automatikoak	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
ep=1, mb=64	73,79	8,58	58,50	53,45	56,12	29,21	36,26	41,07
ep=2, mb=64	73,79	9,32	58,40	53,28	55,87	29,41	35,79	**41,20**
ep=4, mb=64	73,79	8,46	58,36	53,08	55,71	29,18	35,68	40,84
ep=8, mb=64	73,79	4,76	**59,51**	**53,84**	**56,44**	28,26	**38,69**	40,24
ep=1, mb=128	73,79	5,68	59,13	53,56	56,21	28,48	37,74	40,34
ep=2, mb=128	73,79	8,73	58,33	53,15	55,66	29,30	35,49	40,91
ep=4, mb=128	73,79	9,24	57,90	52,59	55,16	29,39	34,33	40,77
ep=8, mb=128	73,79	10,28	57,14	52,10	54,63	29,68	32,77	40,68
ep=1, mb=256	73,79	4,76	**59,51**	**53,84**	**56,44**	28,26	**38,69**	40,24
ep=2, mb=256	73,79	9,87	57,82	52,70	55,29	29,57	34,33	40,99
ep=4, mb=256	73,79	9,97	57,55	52,49	55,17	29,57	34,05	40,90
ep=8, mb=256	73,79	**10,87**	56,07	51,12	53,55	**29,87**	30,48	40,16
ep=2, mb=256	100	13,13	78,66	66,98	72,47	52,72	46,16	**54,75**

5.3 Emaitzen konparaketa

Euskararako gainerako korreferentzia sistemekin konparatuz nabarmen ikus daiteke 5.6 eta 5.7 tauletan, neurona-sareetan oinarrituta eraikitako sistemarekin lortutako emaitzak, aurretik sortutako erregelatan eta ikasketa automatikoan oinarritutako sistemen emaitzak baino baxuagoak direla. Urrezko aipamenekin eta aipamen automatikoekin diferentziak antzekoak dira proportzioan.

5.6 Taula: Euskararako korreferentzia-ebazpenerako sistemen konparaketa, urrezko aipamenekin.

Urrezko a.	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
Erregelak	100	58,99	86,99	80,71	83,57	73,00	68,97	76,51
Ikas. auto.	100	55,38	85,30	78,17	81,14	72,07	64,74	73,94
Ikas. sakona	100	11,31	79,37	67,76	73,29	52,22	48,16	54,66

5.7 Taula: Euskararako korreferentzia-ebazpenerako sistemen konparaketa, aipamen automatikoekin.

Aip. auto.	AD	MUC	B^3	CEAFm	CEAFe	BLANC	LEA	CoNLL
Erregelak	73,79	42,95	62,97	61,56	62,04	43,48	49,26	55,98
Ikas. auto.	73,79	40,98	61,66	59,82	60,00	43,48	45,97	54,21
Ikas. sakona	73,79	9,32	58,40	53,28	55,87	29,41	35,79	41,20

Korreferentzia-ebazpenerako erregelatan oinarritutako sistemek corpus txikia erabili arren, emaitza onak eman ditzakete, sistema adimendunak darabilen ezagutza eskuz sortutako erregelekin osatzen baita. Ikasketa automatikoko sistemek, erregeletan oinarritutakoek baino corpus handiagoa eskatzen dute, sistemak datuetatik ikasten duelako, baina corpus ez oso handiekin ere emaitza onak eman ditzakete. Ikasketa sakonean ordea, neurona-sareak entrenatzeko datu gehiago behar da, proiektu honetan erabili den EPEC-KORREF euskarazko corpusa txiki geratu da, eta lortu diren emaitza kaxkarren erantzule nagusia da.

Ez dago zehaztuta neurona-sareak ikasteko behar duen corpusaren tamaina sistema arrakastatsua izan dadin, baina euskararako 45.000 hitzekin motz geratzen da, eta polonierarako 540.000 hitzekin lortutako emaitzak erregelatan oinarritutakoen parekoak dira. Hizkuntza ezberdinak izateak eraginik izango duen arren, neurona-sareetan oinarritutako euskararako korreferentzia-ebazpenerako sistemak emaitza onak eman ditzan, corpusaren tamaina 100.000-500.000 hitzetara gerturatu beharko litzatekela uste da, gutxienez tamaina bikoiztuz.

6. KAPITULUA

Ondorioak

Aurkibidea

6.1 Ondorioak

6.2 Proiektuaren ondorioak

GrAL honetan, neurona-sareetan oinarritutako euskararako lehenengo korreferentzia-ebazpenerako sistema eraiki da. Horretarako, EPEC-KORREF (45.000 hitz), EPEC corpusaren korreferentzia-ebazpenerako anotatutako atala erabili da. Euskararako sortutako sistema polonierarako sortu berri den sisteman oinarritzen da (Niton et al., 2018), oinarri-lerrorako ezaugarriak eta sarearen arkitektura baliatuz. Neurona-sarea sarreran hitz-embeddingekin, distantziekin eta bestelako ezaugarriekin entrenatu da. Sarearen emaitzetatik korreferentzia-multzoak sortu eta erreferentziazko ebaluatzailearekin ebaluatu da sistema.

Oinarri-lerrotzat hartu den lehen sistemak, *CoNLL* metrikan % 54,47 puntuko *F-measure*a eskuratu du urrezko aipamenekin eta % 40,91 puntuko *F-measurea* aipamen-detektatzaile automatikoarekin. Lortutako emaitzak, 5.7 taulan ikus daitekeenez, aurretik euskararako korreferentzia-ebazpenerako eraiki diren sistemenak baino baxuagoak dira.

Lortutako emaitzak hobetze asmoz, esperimentu ezberdinak egin dira. Lehenik oinarri-lerroan erabiltzen ziren hitz-embeddingak aldatu dira, dimentsio ezberdinekin eta ikasketako corpusari stemmerra pasatu gabe probatu dira. Hori eginik ez da emaitzetan hobekuntza nabarmenik lortu; metrika ia guztiak hobetzea lortu den kasuetan, sistemak korreferentzia erlazio gutxi topatu ditu, izan duen estaldura baxuarengatik eta horregatik MUC metrikan jaitsiera nabarmena izan da. MUC metrikan izandako jeitsierak *CoNLL* metrikan oinarri-lerroak lortutako balioak ez hobetzea ekarri du.

Ezaugarrietan egindako aldaketekin, ez da lortu esperotako hobekuntzarik, informazioa gehitzen ez zuten balioak kendu eta sistemari korreferentzia-ebazpenerako informazio baliagarria izan zitekeena gehituz. Aldaketa horiekin, urrezko aipamen zein automatikoen kasuan, B^3 *CEAFm*, *CEAFe*, eta *LEA* metriketan emaitza altuxeagoak lortu dira, *MUC*, *BLANC* eta *CoNLL* metriketan, berriz, baxuagoak.

Entrenamendu garairako neurona-sarearen ikasketa parametroak aldatuz, *epoch* kopurua eta *mini-batch*aren tamaina aldatuz, *CoNLL* metrikan emaitza hobeak lortu dira lehen aldiz oinarri-lerroarekiko. Emaitza onenak eman dituzten kasuak, ordea, sakabanatuta daude, *epoch* kopuruarekiko 1, 2 eta 8en kasuan. *Mini-batch*ari dagokionez, 256ra handituta lortu badira ere emaitzarik onenak metrika gehienetarako, aipamen automatikoekin *CoNLL* metrikan emaitzarik onena 64ko tamainarekin lortu da. Urrezko aipamene-

kin, emaitzarik onena, *epoch* 1 eta 256 tamainako *mini-batch*a erabiliz lortu da (*CoNLL*n % 54,75 puntuko *F-measure* balioa); aipamen automatikoekin, 2 *epoch* eta 64 tamainako *mini-batch*arekin *CoNLL*n % 41,20 puntuko *F-measure* balioa).

Aurrez esan bezala, GRAL honetan eraikitako neurona-sareetan oinarritutako euskararako korreferentzia-ebazpenerako sistemarekin ez da lortu erregelatan oinarritutako eta ikasketa automatikoko sistemek emandako emaitzetara hurbiltzerik. Lortu diren emaitzak, aurreko sistemekin konparatuz, 10 puntu baino gehiagoko aldea dute urrezko aipamen zein aipamen automatikoen kasuan. Hori, neurona-sareak korreferentzia-ebazpenaren ataza zuzen ikasteko datu gehiago behar dituelako gerta daiteke eta nagusiki erabili den corpusaren tamaina txikiagatik izan dela ondorioztatu da. 46.000 tokeneko corpusa erabili da, ikasketarako erdia erabiliz. Kokatzeko, polonierarako corpusak, adibidez, 540.000 hitz ditu eta ingeleserako erabili ohi direnak, milioi bat hitz baino gehiagokoak dira.

6.2.1 Ondorio pertsonalak

Proiektu hau oso aberasgarria izan da maila pertsonalean. Betidanik erakargarria egin zaidan hizkuntzaren prozesamenduaren alorrean nuen ezagutzan sakondu dut, eta pilpilean dagoen neurona-sareen teknologia ezagutu, ikasi eta erabiltzeko aukera izan dut. Lehen aldia da ikerketa alorreko proiektu bat egiten dudana, eta erronka handia izan da, graduan ikasitakoa eta lortutako gaitasunak praktikan jartzeko aukera izan dut. Programatzeko garaian sortutako arazoei konponbidea emateko gai izan naiz eta proiektuan zehar ez aurrera eta ez atzera geratutako uneetan irtenbidea topatu dut, nahiz eta batzuetan ondo kostata izan. Azkenik, artikuluak bilatzen eta irakurtzen trebatu naiz eta proiektua txukun dokumentatzen ikasi dut, etorkizunean, biak ere, baliagarri izango zaizkit.

6.3 Etorkizunerako Lana

- Etorkizunean ikasketa sakoneko euskararako korreferentzia-sistema lantzen jarraitu nahi bada, ezinbestekoa izango da lehenik corpus handiago bat izatea aipamen eta korreferentziak anotatuak dituena. Euskararako korreferentzia-ebazpenean neurona-sareekin ikertzen jarraitu nahi bada, beharrezkoa izango da gutxienez corpusaren tamaina bikoiztea, 100.000 hitz inguruko corpusa izateko. EPEC corpusa osotasunean (300.000 hitz) korreferentziarako anotatuko balitz, ebaluazioa atala-

ren eta garapena atalaren tamaina mantenduz, 10 bat aldiz handiagoa litzatekeen ikasketa atala izango genuke neurona-sarea entrenatzeko.

- Lehenik eraikitako korreferentzia-ebazpenerako sistemaren errore-analisia egitea, erroreak identifikatzeko eta, ostean, horiei erantzun egokia emateko.
- Ikasketarako erabili diren ezaugarriez gain, ezaugarri berriak bilatu eta konbinazio ezberdinak probatzea, ezaugarriak kimatuz (leave one out), edo inkrementalki gehituz, bakoitzaren eragina aztertuz.
- Sarearen arkitektura, tamaina eta optimizazio teknika ezberdinekin esperimentatu.
- Ikasketarako aipamen-bikote instantziak sortzeko beste metodo bat erabiltzea, adibidez, Sapena et al. (2011) artikuluan proposatutakoa. Bertan, aipamen bakoitza finkatutako distantzia baten barruan dauden aurrekari izateko hautagaiekin parekatzen da, instantzia negatiboen kopurua txikitzeko.
- Aipamen-bikote ereduaren ordez, entitate-aipamen edo multzo-mailakatze ereduak inplementatzea, beste hizkuntza batzuetan emaitza onak eman baitituzte.
- Azkenik, Lee et al. (2017) lana eredutzat hartuz, aipamen-detekzioaren eta ezaugarri erauzketaren menpe egongo ez den neurona-sare konplexua eraikitzea interesgarria izango litzateke, ingeleserako artearen egoera finkatzen duen sistema baita. Honetarako ordea ezinbestekoa da eskuragarri den corpusaren tamaina asko handitzea.

Bibliografia

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., Diaz-De-Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. (2006). Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing.

Aduriz, I., Aranzabe, M. J., Arriola, J. M., de Ilarraza, A. D., Gojenola, K., Oronoz, M., and Uria, L. (2004). *A Cascaded Syntactic Analyser for Basque*.

Agirre, E., Alegria, I., Arregi, X., Artola, X., de Ilarraza, A. D., Maritxalar, M., Sarasola, K., and Urkia, M. (1992). Xuxen: A spelling checker/corrector for basque based on two-level morphology. In *Proceedings of the third conference on Applied natural language processing*, pages 119–125. Association for Computational Linguistics.

Arregi, O., Ceberio, K., de Ilarraza, A. D., Goenaga, I., Sierra, B., and Zelaia, A. (2010). A first machine learning approach to pronominal anaphora resolution in basque. In *A. Kuri-Morales, G. R. Simari (Eds). Advances in Artificial Intelligence. Iberamia 2010. LNAI 6433, pp. 234–243. ISBN 978-3-642-16951-9.*

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics.

Carter, D. M. (1986). *A shallow processing approach to anaphor resolution*. PhD thesis, University of Cambridge.

Ceberio, K., Aduriz, I., de Ilarraza, A. D., and Garcia-Azkoaga, I. (2018). Coreferential relations in basque: the annotation process. *Journal of psycholinguistic research*, 47(2):325–342.

Ceberio, K., Aduriz, I., de Ilarraza Sánchez, A. D., and Azkoaga, I. M. G. (2008). Erreferentziakidetasunaren azterketa eta anotazioa euskarazko corpus batean. In *Gramatika jaietan: Patxi Goenagaren omenez*, pages 153–172. Universidad del País Vasco.

Chen, C. and Ng, V. (2012). Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 56–63. Association for Computational Linguistics.

Chollet, F. et al. (2015). Keras.

Clark, K. (2015). Neural coreference resolution.

Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Connolly, D., Burger, J. D., and Day, D. S. (1997). A machine learning approach to anaphoric reference. In *New methods in language processing*, pages 133–144.

Denis, P. and Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42.

Fernandes, E. R., Dos Santos, C.Ñ., and Milidiú, R. L. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 41–48. Association for Computational Linguistics.

Goenaga, I., Arregi, O., Ceberio, K., De Ilarraza, A. D., and Jimeno, A. (2012). Automatic coreference annotation in basque. In *I. HENDRICkX, S. KüBLER & K. SIMOV (arg.), TLT11 Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories. Lisboa: Edições Colibri*, pages 115–126. Citeseer.

Goikoetxea, J., Soroa, A., and Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*.

Grishman, R. and Sundheim, B. (1995). Coreference task definition. version 2.3. In *Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, Maryland, USA*, pages 335–344.

Hirschman, L. (1997). Muc-7 coreference task definition, version 3.0. *Proceedings of MUC-7, 1997*.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 632–642.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Ng, V. (2008). Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 640–649. Association for Computational Linguistics.

Niton, B., Morawiecki, P., and Ogrodniczuk, M. (2018). Deep neural networks for coreference resolution for polish. In *LREC*.

Ogrodniczuk, M. and Ng, V. (2017). Proceedings of the 2nd workshop on coreference resolution beyond ontonotes (corbon 2017). In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*.

Otegi, A., Ezeiza, N., Goenaga, I., and Labaka, G. (2016). A modular chain of nlp tools for basque. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue, TSD 2016, Brno, Czech Republic, Lecture Notes in Computer Science, vol. 9924, pp. 93-100, Springer. ISBN 978-3-319-45509-9. DOI 10.1007/978-3-319-45510-5_11*.

Park, C., Choi, K., Lee, C., and Lim, S. (2016). Korean coreference resolution with guided mention pair model using deep learning. *ETRI Journal*, 38(6):1207–1217.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 968–977. Association for Computational Linguistics.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Sapena, E., Padró, L., and Turmo, J. (2011). Relaxcor participation in conll shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 35–39. Association for Computational Linguistics.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Soraluze, A. (2017). *Korreferentzia-ebazpena euskarazko testuetan.* PhD thesis, University of The Basque Country.

Soraluze, A., Arregi, O., Arregi, X., and de Ilarraza, A. D. (2017a). Enriching basque coreference resolution system using semantic knowledge sources. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 8–16.

Soraluze, A., Arregi, O., Arregi, X., and de Ilarraza, A. D. (2017b). Improving mention detection for basque based on a deep error analysis. *Natural Language Engineering*, 23(3):351–384.

Soraluze, A., Arregi, O., Arregi, X., de Ilarraza, A. D., Kabadjov, M., and Poesio, M. (2016). Coreference resolution for the basque language with bart. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 67–73.

Soraluze, A., Arregi, O., Arregi, X., and Díaz de Ilarraza, A. (2015). Coreference resolution for morphologically rich languages. adaptation of the stanford system to basque. *Procesamiento del Lenguaje Natural*, (55).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Versley, Y., Poesio, M., and Ponzetto, S. (2016). Using lexical and encyclopedic knowledge. In *Anaphora Resolution*, pages 393–429. Springer.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Wiseman, S., Rush, A. M., and Shieber, S. M. (2016). Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.

Wiseman, S. J., Rush, A. M., Shieber, S. M., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. Association for Computational Linguistics.

Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 176–183. Association for Computational Linguistics.

Yangy, X., Su, J., Zhou, G., and Tan, C. L. (2004). An np-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics*, page 226. Association for Computational Linguistics.