Tex2kor: Sekuentziatik Sekuentziarako Euskararako Korreferentzia-Ebazpena

Egilea: Gorka Urbizu Garmendia
Zuzendariak: Olatz Arregi Uriarte eta Ander Soraluze Irureta

HAP/LAP

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2020ko otsailaren 10a

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Korreferentzia-ebazpena testuko bi aipamenek mundu errealeko entitate bera erreferentziatzen dutela identifikatzeari deritzo. Lan honetan, korreferentzia-ebazpena sekuentziatik sekuentziara lantzeko hurbilpen berri bat aurkezten da. Sekuentziatik sekuentziarako ataza burutzeko Transformer arkitektura neuronala erabili da. Transformerrak ikasketarako darabiltzan sekuentzien luzera mugatzeko, dokumentu etiketatuak zatitu eta elkartzeko algoritmo bat sortu da. Euskararako korreferentzia-ebazpena helburu izanik, euskararako emaitzak hobetzeko datu gehikuntzako teknikak eta BPE segmentazioa gehitu zaizkio hurbilpenari eta tex2kor sistema eraiki dugu. Testu hutsetik korreferentzia-kateak eskuratzeko sistemak, CoNLL metrikan 37,14 puntuko F1 balioa lortu du. Honenbestez, euskararako korreferentza-ebazpenerako zeuden emaitzak hobetzerik lortu ez den arren, korreferentzia-ebazpena lantzeko hurbilpen orokor berri bat aurkeztu da.

Abstract

Coreference resolution is the task of identifying the mentions that refer to the same real world entity. In this work, we present a novel sequence to sequence approach for coreference resolution, for which we use a Transformer. To limit the length of the sequences for the training of the Transformer, we create an algorithm to divide and merge the labeled documents. As our aim is the coreference resolution for Basque, we added some data augmentation techniques and BPE segmentation to build our tex2kor system. The system which converts raw text into coreference-chains, gets F1 37.14 points on CoNLL metric. Therefore, although we did not improve the results of the state of the art system for coreference resolution for Basque, we present a new general approach for coreference resolution.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

Glosategia

aipamen
Testuan zehar entitate bati erreferentzia egiten dion espresio testuala.

atentzio mekanismo
Token batek, sekuentzia baten baitako tokenei jarritako arreta.

autoatentzio
Token batek, bere sekuentziaren baitako tokenetan atentzio mekanismoa aplikatzea.

entitate
Mundu errealeko pertsona, objektu edo erakunde bat.

hizkuntzaren prozesamendua
Hizkuntzaren tratamendu automatikoaren inguruko ikerketa-lerroa.

korreferentzia-ebazpen (KE)
Testu bateko bi aipamenek entitate bera erreferentziatzen dutenean, bien artean korreferentzia-erlazio bat dago, horrelakoak ebaztearen atazari korreferentzia-ebazpen deritzo.

korreferentzia-erlazio
Testu bateko bi aipamenek mundu errealeko entitate bera erreferentziatzen dutenean, bien artean korreferentzia-erlazio bat dago.

korreferentzia-kluster
Testu batean entitate berari erreferentzia egiten dioten aipamenek osatzen duten multzoa.

token
Hizkuntzaren prozesamenduaren arloan testu zatituaren unitate bat da. Orokorrean hitzaren baliokidea da, baina hitz-zatia edo karakterea ere adieraz dezake.

sare neuronal / neurona sare
(*Neural Network*, NN), informazioa prozesatzeko adimen artifizialaren arloan erabiltzen den eredu matematiko bat.

sasi-etiketatzea
Corpus bat automatikoki etiketatzea (*pseudo-labeling*).

singleton
Korreferentzia-erlaziorik ez duen aipamena.

Transformer
Autoarretan oinarritutako sekuentziatik sekuentziarako arkitektura neuronala.

1 Sarrera

Adimen artifizialaren barnean kokatutako hizkuntzaren prozesamenduaren arloak, *Natural Language Processing* ingelesez, hainbat problema ezberdin biltzen ditu, tartean, hemen landu den korreferentziarena. Labur esanda, korreferentzia-ebazpena testu batean pertsona edo objektu berbera erreferentziatzen duten espresio testualak identifikatzeri deritzo.

1.1 Motibazioa

Korreferentzia-ebazpena (KE) zer den ulertzeko, atazaren definizioa eta inguruko terminologia azaldu aurretik jo dezagun ataza honek zertarako balio duen eta hizkuntzaren prozesamenduko beste atazetan duen eragina ikustera.

Hizkuntzaren prozesamenduko ataza ezagunenetako bat itzulpen automatikoa dugu. Sarean eskuragarri dagoen itzultzaile automatiko batekin[1] ondorengo esaldia itzuli dugu euskaratik gaztelaniara:

(1) *EU: Irakaslearen senarra ezagutu nuen, eta erizaina zela esan zidan.*
ES: Conocí al esposo de la maestra y él me dijo que era enfermera.

Adibide honetan ikus dezakegunez, euskarazko *erizain* hitza itzultzean, gaztelaniaz beharrezkoa den genero marka erabakitzeko garaian, bi aukeren artean, femeninoa aukeratzen du. Pertsona batek, aldiz, senarrari buruz ari garela jakinda, maskulinoa erabiliko luke.

Pentsa liteke, ikasketa prozesuan barneratutako genero aurreiritziek (*gender bias*) eragindako akatsa dela honakoa, hitz-bektoreetan zein ikasketa sakoneko sistemetan aztertu den moduan, datuetatik ikasten duten sistema horiek, gizakiok idatzitako testuetatik ikastean gure gizarteko hainbat aurreiritzi barneratzen baitituzte (Bolukbasi et al., 2016; Garg et al., 2018). Baina kasu honetan horrelakorik ez dela gertatu ikus genezake esaldia sinplifikatuta:

(2) *EU: Bere senarra erizaina zen.*
ES: Su esposo era enfermero.

Itzulpen automatikorako sistema gai da *erizaina* itzultzerakoan generoa egoki esleitzeko, senarrari dagokion *enfermero*, maskulinoan erabiliz. Beraz, ondoriozta genezake, lehenengo adibidean, itzulpen sistema ez dela gai ulertzeko *erizaina*, aurrez aipatutako *senarra* bera dela. Hau da, ez da gai *senarra* eta *erizaina*ren arteko korreferentzia-erlazioa ulertzeko, pertsona berari egiten diotela erreferentzia, alegia, eta horren arabera itzulpen zuzena egiteko.

Hizkuntzaren prozesamenduaren beste aplikazio orokor eta ezagunenetako bat elkarrizketa sistemak (*chatbot*-ak) dira.

Korreferentzia lantzen ez duten elkarrizketa sistemek, elkarrizketan zehar aurrez aipatutakoak, uneko elkarrizketarekin lotzeko gaitasuna falta dute. Kode irekiko, ingelesezko elkarrizketa sistema batek[2] honako elkarrizketa ematen digu:

[1] https://translate.google.com/
[2] http://www.square-bear.co.uk/mitsuku/chat.htm

(3)
Human: Hi Mitsuku.
Mitsuku: Hi there.
Human: What is your favorite movie?
Mitsuku: My favourite movie is Terminator. Have you seen it?
Human: No.
Mitsuku: I would recommend you check it out.
Human: Who directed it?
Mitsuku: No idea. I will guess at Steven Spielberg.
Human: Who directed Terminator?
Mitsuku: James Cameron.

Adibide honetan ikus dezakegunez, elkarrizketa sistemari filmaren zuzendaria zein den galdetuz, erantzun zuzena emateko gai den arren, galdera hori aurretik aipatutako pelikula bati buruzkoa bada, ez da gai erantzun zuzena emateko. Akats hori, izenordainak (*it*) eta aurrez agertutako filmeak (*Terminator*) duten korreferentzia-erlazioa ez harrapatu izanak eragin du. Hori dela eta, ez da gai *it* agertzen denean *Terminator* filmari buruz galdetzen ari garela ondorioztatzeko eta filmaren zuzendaria zein den esateko.

Informatikaren hastapenetan, 1950ean, Alan Turingek konputagailuen adimena ebaluatzeko plazaratutako Turingen Testa (Turing, 2009) bera hobetzeko proposatutako Winograden eskema probak (Levesque et al., 2012) ere korreferentzia-ebazpena du muinean. Identitate faltsu bat eraikiz, edota galdera zailak erantzutea txantxak eginez saihestuz Turingen testa gaindi daiteke benetako adimenik izan gabe (Levesque et al., 2012). Winograden eskema probak, horrelakoak saihesteko, egitura honetako esaldi pareak ditu ardatz:

(4) *Garaikurra ez da maleta marroian sartzen hura txikiegia delako. Zer da txikiegia?*
A: Garaikurra
B: Maleta

(5) *Garaikurra ez da maleta marroian sartzen hura handiegia delako. Zer da handiegia?*
A: Garaikurra
B: Maleta

Galdera horiek zuzen erantzun ahal izateko, *hura* garaikurrari ala maletari dagokion jakin behar dugu lehenik, eta korreferentzia-erlazio anbiguo hori ebazteko testuinguruaren eta mundu errealaren ezagutza ezinbestekoa da. Honenbestez, azken adibide horrek mahai gainean jartzen du korreferentzia-ebazpenaren zailtasuna.

Korreferentzia-ebazpena oso garrantzitsua da hizkuntzaren ulermen sakona eskatzen duten hizkuntzaren prozesamenduko edozein azken helburuko aplikaziotan. Besteak beste, itzulpen automatikorako (Werlen eta Popescu-Belis, 2017; Ohtani et al., 2019), elkarrizketa-sistemetarako (Agrawal et al., 2017; Zhu et al., 2018), testuen laburpen automatikorako (Steinberger et al., 2016; Kopeć, 2019), sentimenduen analisirako (Krishna et al., 2017) edo informazio erauzketarako (Wang et al., 2018; Singh, 2018) onuragarria da korreferentzia-ebazpena. Erregelatan oinarritutako hizkuntzaren prozesamenduko sistemetan, korrefe-

rentziaren ataza esplizituki landu beharreko alor bat da. Sare neuronaletan oinarritutako hizkuntzaren prozesamenduko sistemek, berriz, testu hutsezko corpus handietatik korreferentzia-ebazpena inplizituki ikasteko gai direla erakutsi dute beste ataza batzuetarako entrenatu bitartean (Clark et al., 2019; Tenney et al., 2019). Guk dakigula, ordea, oraindik ez da aztertu ea itzulpen automatikoa edo elkarrizketa-sistemak bezalako atazetarako sistema neuronaletan, korreferentzia-ebazpenean inplizituki ikasketa gehigarririk egiteak hobekuntzarik ekarriko lukeen.

1.2 Atazaren definizioa

Korreferentzia-ebazpena (KE) terminoaren esanahia *Message Understanding Conference* (MUC-6, 1995) konferentzian zehaztu zen hizkuntzalaritza konputazionalaren ikuspuntutik. Eta orduz geroztik, ataza hori ebazteko sistema automatikoen garapenean aurrerapauso handiak egin dira.

Korreferentzia-ebazpena honela definitzen da (Hirschman eta Chinchor, 1998):

"Testu bateko bi espresio testualek mundu errealeko pertsona, erakunde edo 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:

(6) ***Lujanbiok*** *lortu du Bertsolari Txapelketa Nagusiko txapela, Mendiluzeri gailenduta irabazi du* ***Maialenek*** *eta* ***berak*** *irabazitako txapela etorkizuneko bertsolariei eskaini die.*

Lujanbiok, ***Maialenek*** eta ***berak*** espresio testualek korreferentzia-erlazioa dutela edo korreferenteak direla esan dezakegu, mundu errealeko pertsona berari egiten diotelako erreferentzia. *Mendiluzeri* espresio testualak ere, mundu errealeko pertsona bati egiten dio erreferentzia, baina pertsona berbera ez denez, aurreko hiru aipamenekin korreferentzia-erlaziorik ez du, edota ez da korreferentea.

Ataza honetan sarri erabiltzen diren beste hiru termino *entitatea*, *aipamena* eta *singletona* dira. Entitatea mundu errealeko pertsona, objektu edo erakunde bat litzateke. Aipamena, berriz, testuan zehar entitate bati erreferentzia egiten dion espresio testuala da. Azkenik, korreferentzia-erlaziorik ez duten aipamenei *singleton* deritze.

Termino horiek hobeto ulertzeko, azter dezagun sakonago 6. adibideko esaldia:

(7) [***Lujanbiok***] *lortu du* [[***Bertsolari Txapelketa Nagusiko***] ***txapela***], [***Mendiluzeri***] *gailenduta irabazi du* [***Maialenek***] *eta* [[***berak***] ***irabazitako txapela***] [***etorkizuneko bertsolariei***] *eskaini die.*

7. adibidean, kortxete artean ikus ditzakegunak aipamenak izango lirateke, tartean, erabat ohikoa den moduan, aipamen batek beste aipamen luzeago baten parte izan dai-

tekeela ikus daiteke, [[***Bertsolari Txapelketa Nagusiko***] ***txapela***] kasu. [***Lujanbiok***], [***Maialenek***] eta [***berak***] aipamenek entitate berari egiten diotenez, erreferentzia korreferentzia-erlazio bat dute eta korreferentzia kluster edo multzoa osatzen dute. Bestalde, [***Bertsolari Txapelketa Nagusiko txapela***] eta [***berak irabazitako txapela***] aipamenek ere entitate berari erreferentzia egiten diotenez, korreferentzia-erlazioa dute, beste korreferentzia kluster bat osatuz. Azkenik, testuko gainontzeko aipamenek ([***Bertsolari Txapelketa Nagusiko***], [***Mendiluzeri***] [***etorkizuneko bertsolariei***]) bakoitzak entitate ezberdin bati egiten diote erreferentzia, eta horregatik ez dute gainerako aipamenekin korreferentzia-erlaziorik, beraz, hauek *singletonak* lirateke.

1.3 Proiektuaren helburuak

Korreferentzia-ebazpenaren arloan egindako ikerketaren gehiengoa hizkuntza bakarraren bueltan izan da, ingelesarenean. Atazarako corpus handienak eta anitzenak izatearekin batera, arloko aurrerapen gehienak ere ingeleserako sistementzat egin dira. Kasu gehienetan sistema horiek beraiek, zuzenean erabili edo egokitu daitezke beste hizkuntza batzuetara. Sare neuronalek KEeko atazara ekarritako aurrerapenekin ere joera berdina izan da, hizkuntza bakoitzerako eskuragarri dagoen datu kopuruak are garrantzi handiagoa duelako, baliabide gehien duten hizkuntzetan bildu da ikerketa, nagusiki ingelesean. Gainontzeko hizkuntzak gutxiago landu dira, eta zer esanik ez baliabide gutxiko hizkuntzak, hiztun gutxi dituztenak edo hizkuntza gutxituak.

Master amaierako lan honetan, euskararako korreferentzia-ebazpenean egindako lanari (Soraluze, 2017) jarraipena eman nahi zaio, sare neuronalen bidea jorratuz. Orain arte, euskararako korreferentzia-ebazpena erregela eta ikasketa automatikoaren bidez landu da nagusiki, eta sare neuronalekin egindako saiakerek (Urbizu et al., 2019a,b) ez dute arrakastarik izan, besteak beste euskararako erabilgarri dagoen corpusaren tamaina mugatuarengatik.

Lan honen aurretik landutako euskararako korreferentzia-ebazpen neuronalak, erregelatan oinarritutako aipamen-detektatzailea darabil, eta aipamen-bikote eredu sinplea eta neurona sare arrunt bat konbinatzen ditu.

Haatik, lan honetan, korreferentzia-ebazpenerako muturretik muturrerako (*end2end*) sistema bat eraikiko da, sarreran testu gordina hartu eta, ezaugarri linguistiko gehigarririk gabe, korreferentzia-ebazpena gauzatzeko. Horretarako, sekuentziatik sekuentziarako (*seq2seq)* eredu bat eraiki da, Transformer (Vaswani et al., 2017) arkitektura erabiliz, eta datu eskasiari aurre egiteko datu gehikuntzako hainbat teknika erabili dira.

Honela, euskararako korreferentzia-ebazpenerako sistema ez neuronalen emaitzak hobetzea da helburua. Aurrez eraikitako euskararako sistema neuronaletan ateratako ondorio nagusia, ikasketa sakoneko teknikak erabiltzeko datu kopuru handiagoa behar dela izan da, eta lan honetan, hurbilpena aldatu eta arkitektura konplexuagoak erabiltzeaz gain, euskarazko sasi-etiketatutako eta baliabide handiagoko hizkuntzetako corpusak erabiliz, euskararako korreferentzia-ebazpenerako sistema neuronal bat eraiki da.

Gainera, lan honetan planteatutako hurbilpenak, beste hizkuntza batzuetara, zein korreferentzia-ebazpeneko azpi-atazetara aplikatzeko aukera ematen du, hizkuntza edo ata-

za horretarako etiketatutako corpusa izanez gero. Bestalde, baliabide gutxiko hizkuntzetara ere zabal liteke sistema, eta hau entrenatzeko datu nahikoa izan ezean, hizkuntza arteko transferentzia bidezko ikasketako teknikak jorratu daitezke.

Master amaierako lan hau honela dago egituratua: 2. atalean aurrekariak aurkeztuko dira, eta artearen egoera aztertuko da. 3. atalean erabilitako metodologia azalduko da, erabilitako corpusak (3.1. atalean), sekuentziatik sekuentziarako hurbilpena (3.3. atalean) eta erabilitako Transformer arkitektura (3.4. atalean) azalduko dira besteak beste. 4. atalean egindako esperimentu ezberdinak azalduko dira. 5. atalak, lortutako emaitzak biltzen ditu, eta azkenik, 6. atalean master amaierako lan honen ondorioak eta etorkizunerako lana aurkeztuko dira.

2 Aurrekariak

2.1 Korreferentzia-ebazpenaren hastapenak

Ataza honen hastapena 60-90eko hamarkadetan kokatzen da, garai hartako lan esanguratsuenen artean "*Resolving Pronoun References*"(Hobbs, 1978), "*A Shallow Processing Approach to Anaphor Resolution*"(Carter, 1986) eta "*Algorithm for Pronominal Anaphora Resolution*"(Lappin eta Leass, 1994) artikuluak azpimarra ditzakegu.

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

Kasu askotan, KEaren beharrean, bere azpi-ataza den anafora-ebazpena landu izan da; testuko korreferentzia kate guztiak topatu beharrean, izenordain edo izen-sintagma batek osatutako aipamena, dagokion aurrekari korreferentearekin lotzean datzana. Beste batzuetan, korreferentzia-erlazio mota jakin batzuk bakarrik landu izan dira, adibidez, izenordainen ebazpena, edota entitate mota bakarrari dagokiona (pertsonak, lekuak,...).

2.2 Korreferentzia-ebazpen ez neuronala

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 korreferentzia klusterretan multzokatu behar dira, entitate berberari erreferentzia egiten diotenak elkartuz. Berriki, sare neuronalen teknologiarekin, biak aldi berean lantzea ari da nagusitzen, baina, hala ere, ingeleserako sistema ezberdinak konparatzeko erabiltzen den CoNLL-2012 corpusak ez du *singletonik*, eta beraz korreferentzia-erlazioren bat duten aipamenak detektatu eta hauek egoki multzokatzea da ikasi eta ebaluatzen dena sistemarik onena zein den konparatzeko.

Aipamen-detekziorako, erregelatan oinarritutako sistemak erabili dira nagusiki, aipamenak multzokatzeko erregelatan oinarritutako hurbilpenetan ez ezik, ikasketa automatikoan oinarritutako hurbilpenetan ere bai. Sistema bat nabarmentzearren, Stanforden erregelatako aipamen-detektatzailea (Lee et al., 2011) azpimarratuko genuke. Berriki, ikasketa sakonaren arrakastak, eta sare neuronalak muturretik muturrera entrenatzeak joera aldaketa bat ekarri du aipamen-detekziora.

Korreferentzia klusterrak sortzeko berriz denetariko hurbilpenak aztertu eta baliatu dira. Lehen urteetan, erregelatan oinarritutako sistemak izan ziren nagusi, horiek Mitkov (1999) berrikuspen bibliografiak daude bilduta. Ordutik hona, aurkeztutako sistemen artean bat aipatzearren, Lee et al. (2013) nabarmenduko genuke, ordurarteko emaitzarik onenak lortu baititu. Hala ere, ikasketarako eskuragarri dagoen corpus etiketatuaren tamaina eta ordenagailuen konputazio gaitasuna handitu ahala, erregelatan oinarritutako sistemek beste domeinu eta hizkuntza batzuetara egokitzeko duten zurruntasuna medio,

ikasketa automatikoko eta, berriki, ikasketa sakoneko sistemak gailendu dira arloan.

Ikasketa automatikoko sistemen artean, hurbilpen ezberdinak proposatu dira. Soon et al. (2001) lanean aurkeztutako ikasketa automatikoko lehen sistema arrakastatsuak bidea zabaldu zuenetik, ikasketa automatikoko sistema ugari aurkeztu dira. Gehienek, aipamen-detektatzaile automatikoren bat darabilte, ondoren aipamen horiek sailkatzaile edo multzokatze algoritmoren baten bitartez multzokatzeko. Algoritmo horiek honakoetan sailka genitzake: Aipamen-bikote eredua, entitate-aipamen eredua, aipamen-mailakatze eredua eta multzo-mailakatze eredua.

- 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 Sapena et al. (2011).

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

  Eredu honetan aipamen bat aurretik sortutako aipamen multzo batekin korreferen tea 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 eta Ng, 2009).

Ikasketa automatikoan oinarritutako sistema esanguratsuenak Ng (2010), Ng (2017) eta Sukthanker et al. (2020) berrikusketa bibliografikoetan daude bilduta. Bat aipatzearren, Versley et al. (2008) lanean aurkeztutako BART sistema azpimarratuko genuke, duen arkitektura modularra dela eta, beste hizkuntza batzuetara egokitu delako: italierara (Poesio et al., 2010), alemaniarrera (Broscheit et al., 2010), polonierara (Kopec eta Ogrodniczuk, 2012), arabierara eta txinerara (Uryupina et al., 2012), euskarara (Soraluze et al., 2017a) eta Indiako zenbait hizkuntzatara (Sikdar et al., 2016).

2.3 Korreferentzia-ebazpen neuronala

Azken urteeetan, adimen artifizialeko gainontzeko alorretan bezala, datu kopuru eta konputazio gaitasun handiak eskura izateak, ikasketa sakonaren iraultza ekarri du hizkuntzaren prozesamendura. Edozein atazatan, aurrez finkatutako emaitza gehienak hobetzea lortu da, kasu askotan, modu nabarmen batean gainera. Ikasketa automatikoko sistema klasikoek erabiltzen dituzten ezaugarri linguistikoen beharrik gabe, testu hutsetik abiatuz, neurona sareetan oinarritutako sistemak ari dira gailentzen ataza gehienetan.

2.3.1 Ingeleserako korreferentzia-ebazpen neuronala

Korreferentzia ebazpenerako sare neuronalak arrakastaz erabiltzen lehenek, ikasketa automatikoko hurbilpenak baliatu dituzte, aipamenak multzokatzeko erabiltzen diren sailkatzaile eta multzokatze algoritmoak neurona sare trinko batez ordezkatuz (Wiseman et al., 2015, 2016; Clark eta Manning, 2016a,b). Sistema horiek, erregela bidezko aipamen automatikoak eta aurreprozesatutako ezaugarri linguistikoak baliatzen dituzte oraindik.

Aurreprozesatutako ezaugarri linguistikoak erabiltzea, ordea, Lee et al. (2017) eta Lee et al. (2018) lanetan aurkeztutako muturretik muturrerako sistemekin amaituko da. Sistema horiek aipamen eta ezaugarri linguistikoen beharrik gabe, testu hutsetik abiatuta, gai dira korreferentzia-ebazpena burutzeko, eta emaitzetan hobekuntza handiak lortzen dituzte gainera. Horrela, aurreprozesaketan eskuratutako ezaugarrietan dagoen erroreen propagazioa saihestuz. Horretarako, lehenik, Lee et al. (2017) lanean aipamen-ranking ereduaren antzeko bat darabilte, hitz-bektoreak, BiLSTM geruza bat eta atentzio-mekanismoak konbinatuz aipamenak detektatzeko eta multzokatzeko. Lee et al. (2018) lanean, hobekuntzak proposatzen dituzte, ELMo hitz-bektore testuingurudunak (Peters et al., 2018) erabiliz, eta multzokatze ereduan hobekuntzak eginda. Gerora, sistema horri hobekuntzak proposatu zaizkio, aipagarriena errefortzu bidezko ikasketa gehitzen diona (Fei et al., 2019).

Berriki, hizkuntzaren prozesamenduko ia ataza oro asaldatu duen teknika berri bat, testu gordinaren gainean aurre-ikasketa ez gainbegiratua egin eta ondoren edozein atazetara ikasitakoa transferitzen duena (Devlin et al., 2019), aplikatu da KEaren atazan. Kantor eta Globerson (2019) eta Joshi et al. (2019b) lanetan, aurre-ikasketako BERT (Devlin et al., 2019) eredua Lee et al. (2018) arkitekturaren sarreran txertatuz, hobekuntza nabarmenak lortzen dira. Joshi et al. (2019a) lanean ere, aurre-ikasketako eredu bat gehitu zaio Lee et al. (2018) arkitekturari, kasu honetan ordea, BERT eredua erabili beharrean, SpanBERT eredua proposatzen dute. SpanBERT, KEean zein hizkuntzaren prozesamenduko beste ataza batzuetan emaitza hobeak ematen dituen BERTen aldaera bat da, ikasketako garaian ondoz-ondoko hitzak maskaratzen (ezkutatzen) dituena. Horrela, aurreko sistemek baino emaitza hobeak lortzen dituzte, alde nabarmenarekin gainera. Berriki, Hourali et al. (2020) lanean, gaur egungo emaitzarik onenak plazaratu dituzte, RoBERTa (Liu et al., 2019) eredu aurre-ikasiaren hitz-bektoreak erabiliz, eta MCDM neuronal batekin korreferentzia-ebazpena egiteko. Ingeleserako CoNLL-2012 ingeleseko corpusean sistema ezberdinek lortutako emaitzak CoNLL metrikan, lehenengo taulan daude ikusgai.

Sistema	F1
(Versley et al., 2008)[1]	56,1
(Lee et al., 2011)	58,3
Wiseman et al. (2016)	64,2
Clark eta Manning (2016b)	65,7
Lee et al. (2017)	68,8
Lee et al. (2018)	73,0
Fei et al. (2019)	73.8
Kantor eta Globerson (2019)	76,6
Joshi et al. (2019b)	77,1
Joshi et al. (2019a)	79,6
Hourali et al. (2020)	80,0

[1] Uryupina et al. (2012)k aurkeztutako emaitzak

Taula 1: Ingeleserako korreferentzia-ebazpenaren artearen egoera: CoNLL-2012 corpusean lortutako CoNLL metrikaren F1 balioak.

2.3.2 Korreferentzia-ebazpen neuronala ingelesetik at

Korreferentzia-ebazpen neuronalaren arloan, ingelesa ez beste hizkuntza batzuk ere landu dira, besteak beste, poloniera (Nitoń et al., 2018), koreera (Park et al., 2016), txinera (Park et al., 2016), japoniera (Shibata eta Kurohashi, 2018), frantsesa (Grobol, 2019), telugua (Annam et al., 2019), nederlandera (Allein et al., 2020), errusiera (Sboev et al., 2020) eta euskara (Urbizu et al., 2019a,b). Guk dakigula, gaur gaurkoz, euskaraz gain, Indiako telegu hizkuntza da KEerako sistema neuronala duen baliabide gutxiko hizkuntza bakarra.

Sistema elebakar horietaz gain, elearteko transferentzia bidezko ikasketa ere landu da KE neuronalerako, hizkuntzaren prozesamenduko beste ataza batzuetan, itzulpen automatikoan eta hizkuntza ereduan adibidez (Lample eta Conneau, 2019), emaitza onak eman ondoren. Cruz et al. (2018) lanean, portugeserako KE neuronalerako sistema bat garatu dute, elearteko hitz-bektoreak erabilita espainierako corpusetik ikasiz, bi hizkuntzak gertukoak direla baliatuz. Kundu et al. (2018) lanean, antzeko hurbilpena erabiltzen dute, ingeleseko corpus handietatik, gaztelaniako eta txinerarako KEa ikasteko.

2.4 Euskararako korreferentzia-ebazpena

Euskarazko testuetan KE automatikoan egindako lana, Soraluze (2017) tesi lanean dago bildua. Euskararako, gainontzeko hizkuntzetan bezala, hainbat teknika ezberdin ikertu dira ataza lantzeko. Lehenik aipamen-detektatzaile automatiko bat garatzen dute, erregelatan oinarritzen dena eta euskarazko aipamenen egiturak kontuan hartzen dituena, (Soraluze et al., 2017b). Aipamen-detektatzaile hori darabilte gaur gaurkoz, lan honen aurretik euskararako sortutako KEerako sistema guztiek. Soraluze et al. (2015) lanean, ingeleserako diseinatutako erregela bidezko Stanfordeko sistema (Lee et al., 2013) egokitu da KEa eus-

karazko testuetan egiteko. Stanfordeko 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. Stanfordeko sistemak dituen 10 baheak euskararen ezaugarrietara moldatu dituzte; euskaraz elipsiari aurre egiteko, bahe bat gehitu zaio sistemari elipsia behar den moduan tratatzeko. Egokitutako sistemak, nahikoa ditu Ixa taldean garatutako analisi-katearen (Aduriz et al., 2006; Otegi et al., 2016) eta euskararako sortutako aipamen-detektatzailearen irteera jasotzea euskarazko testuetako korreferentzia-ebazpena gauzatzeko. CoNLL metrikan 55,74ko F1 balioa lortzen dute horrela.

Gerora, erregelatan oinarritutako sistema horren errore-analisi sakona burutu eta hobekuntzak proposatzen dituzte (Soraluze et al., 2017a, 2019). 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 F1 balioa (0,24 puntuko hobekuntza) lortzen dute aipamen automatikoak erabiliz.

Soraluze et al. (2016) lanean, ikasketa automatikoan oinarritutako sistema bat plazaratzen da, ingeleserako diseinatutako BART korreferentzia-ebazpenerako sistema (Versley et al., 2008) euskararen ezaugarrietara egokituz (besteak beste lema-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 eta aipamen-detektatzaile automatikoaren aipamenak. Sistema euskarara egokitu, eta EPEC-KORREF corpusean ebaluatzen da 53,72 puntuko F1 balioa lortuz CoNLL metrikan. Azkenik, (Soraluze et al., 2017a) lanean, Wikipediatik erauzitako ezaugarri semantikoak gehitu zaizkio ikasketa automatikoko sistemari. CoNLL metrikan 54,21 puntuko F1 balioa lortu da.

Azken urteetan KEean hizkuntza gehienetan hartutako bideari segika, euskararako KEean ere sare neuronalak ikertzeari ekin zaio. Urbizu et al. (2019b) lanean, polonierarako eraikitako aipamen-bikote eredua darabilen sare neuronala (Nitoń et al., 2018) egokitu da euskararako. Sistemak, erregelatan oinarritutako aipamen-detektatzailearekin (Soraluze et al., 2015) detektatutako aipamenak korreferentzia klusterretan biltzen ditu, ezaugarri linguistiko gehigarrien beharrik gabe. CoNLL metrikan 41,20 puntuko F1 balioa lortzen du, sistema ez neuronalek baino baxuagoa. Euskararako eskuragarri dagoen corpusaren tamaina txikiagatik sistemak korreferentzia-erlazio gutxi sortzen dituela ondorioztatu da.

Urbizu et al. (2019a) lanean, euskararako eskuragarri dagoen KEerako corpusaren tamaina ikusita, elearteko sistema batek baliabide gutxiko hizkuntzetarako KEa gauzatzen lagun dezaketen aztertu da. Bertan, Urbizu et al. (2019b) lanean erabilitako arkitektura mantendu da, eta ingeleserako KEerako corpus handiago batetik ikasi da euskararako ataza burutzen, elearteko hitz-bektoreak erabiliz (Artetxe et al., 2017, 2018).

Euskararako eraiki diren sistemen emaitzak 2. taulan daude ikusgai (metrikei buruzko xehetasun gehiago 5.1. atalean). Orain arteko emaitzarik onenak erregelatan oinarritu-

Sistema	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
Soraluze et al. (2019)[1]	73,79	42,95	62,97	61,56	62,04	43,48	49,26	55,98
Soraluze et al. (2017a)[2]	73,79	40,98	61,66	59,82	60,00	43,48	45,97	54,21
Urbizu et al. (2019b)[3]	73,79	9,32	58,40	53,28	55,87	29,41	35,79	41,20

[1]erregelatan oinarritua [2]ikasketa automatikoan oinarritua [3]ikasketa sakonean oinarritua

Taula 2: Euskararako korreferentzia-ebazpenerako sistemen konparaketa. EPEC-KORREF corpusean lortutako CoNLL metrikaren F1 balioak.

tako sistemak (Soraluze et al., 2019) lortu ditu, CoNLL metrikan 55,98ko F1 balioarekin. Ikasketa automatikoko sistema (Soraluze et al., 2017a), erregelatatan oinarritutakoen emaitzetatik gertu gertatzen da, eta ikasketa sakoneko teknikekin eraikitako sistemak (Urbizu et al., 2019b), beste bi hurbilpenen emaitzetatik urrun geratu da.

3 Metodologia

3.1 Erabilitako corpora

Jarraian, egin diren esperimentuetan erabili diren corpora azaltzen dira. Batetik, euskararako eskuz etiketatutako EPEC-KORREF corpusa, eta bestetik automatikoki etiketatutako Elhuyar Web eta Elhuyar QTleap corpora, eta ingeleserako eskuz etiketatutako PreCo corpusa.

3.1.1 EPEC-KORREF corpusa

EPEC-KORREF[3] (Ceberio et al., 2018) corpusa, EPEC (Aduriz et al., 2006) corpusaren azpi-corpusa da eta euskarazko *Egunkaria* egunkariko albisteez osatuta dago. Corpus horrek eskuz etiketatutako aipamenak eta korreferentzia-erlazioak ditu, korreferentzia-ebazpenerako baliagarriak izan daitezkeen beste ezaugarri askorekin batera. Corpusak 61.281 hitz ditu eta 16.881 aipamen dauzka etiketatuta, *singletonak* barne. 3. Taulan ikus daitekeen moduan, ikasketa, garapena eta ebaluazioa ataletan banatua dator.

	Dokumentuak	**Hitzak**	**Aipamenak**	**Klusterrak**	**Singletonak**
Ikasketa	97	27.611	7.613	1.151	4.062
Garapena	41	11.948	3.282	496	1.759
Ebaluazioa	75	21.722	5.986	904	3.194
Guztira	213	61.281	16.881	2.551	9.015

Taula 3: EPEC-KORREF corpusa

EPEC-KORREF corpusaren ikasketa atala neurona-sarea entrenatzeko erabili da, garapeneko atala, garapenean zehar arkitekturaren doiketa egin eta esperimentu ezberdinen emaitzak konparatzeko eta corpusaren ebaluazio atala bukaerako sistemaren ebaluazioa egiteko utzi da.

3.1.2 Elhuyar Web corpus sasi-etiketatua

Elhuyar Web Corpusa, sarean euskaraz idatzita dagoen testu guztia biltzen duen corpusa da (189 miloi hitz). Corpusaren zati bat euskararako korreferentzia-ebazpenerako sistema onena (Soraluze et al., 2019) erabilita automatikoki etiketatu dugu. Testu gordin horretatik, 50 esaldiko dokumentu artifizialak sortu dira, eta erregelatan oinarritutako sistema pasa zaio. Sistemak eskatzen duen denbora dela eta, ez da corpus osoa etiketatu, eta guztira, 360.000 hitz dituen korreferentzia automatikoki etiketatutako corpusa lortu dugu.

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

3.1.3 Elhuyar QTleap corpus sasi-etiketatua

Elhuyar QTleap corpusa, itzulpen memorietako testu elebakarrek osatzen dute, guztira 10 miloi hitz, eta hauek dagoeneko korreferentzia-ebazpena automatikoki eginda dute, Ixa-Kat tresnarekin (Aduriz et al., 2004; Otegi et al., 2016). Ixa-Katek euskararako korreferentzia-ebazpena egiteko darabilen sistema, Soraluze et al. (2019) lanean plazaratutakoan oinarrituta dago, baina azkarragoa izan dadin, ez ditu bahe guztiak integratuta, eta ondorioz emaitza kaxkarragoak ematen ditu. Hori dela eta, aipamen-detekzioa nahiko ondo egin arren, korreferentzia-erlazio gutxi ditu sortuak corpus horrek.

3.1.4 PreCo corpusa

PreCo (Chen et al., 2018) korreferentzia-ebazpenerako ingelesezko corpus handi bat da. 38K dokumentuk eta 12,5M hitzek osatzen dute corpus erraldoi hau, eta hiztegi aldetik, nagusiki, lehen hizkuntza ingelesa duten haurren lexikoarekin osatua dago, beste corpus batzuk baino hitz arraroen proportzio txikiagoa duelarik. Korreferentzia-ebazpenean gehien erabiltzen den CoNLL-2012 (Pradhan et al., 2011) corpusean ez bezala, *singletonak* etiketatu dira, horiek ez etiketatzeak, aipamen-detekzioa eta korreferentzia-kateak lotzearen arteko eragina ikertzea zailtzen du eta bi atazak corpus berdinean aldi berean ikastea galarazten du. Corpusa ikasketa eta garapena ataletan dago banatuta, garapenerako 500 dokumenturekin, eta gainontzekoa ikasketarako. Ingeleserako KEerako egindako esperimentuetan, neurona-sarea ikasketa atalean entrenatu da, eta garapena atalean doitu eta ebaluatu.

3.2 Corpusaren formatua

Korreferentzia-ebazpena, sarritan, hizkuntzaren prozesamenduko hainbat atazatan erabiltzen den zutabekako CoNLL formatuan adierazi izan da, zutabeetako batean, korreferentzia-ebazpenaren etiketa parentesi bidezko notazioarekin adierazita.

Formatu horretan, testu gordina tokenizatuta agertzen da ezkerreko zutabean, errenkada bakoitzeko hitz edo token bat eta honi dagozkion ezaugarri linguistikoak ditugularik. Ezaugarri horien artean, ataza eta beharraren arabera, token bakoitzari dagokion kategoria gramatikala (POS), lema, esaldia eta paragrafoa bezalako informazioa gehi dakizkioke. Lan honetan, bi zutabe soilik erabiliko dira: ezkerreko zutabean tokenak, eta eskuinekoan korreferentzia kateei dagokien informazioa parentesi bidezko egitura batean kodetuta:

Parentesiek aipamena noiz hasi eta noiz amaitzen den adierazten dute; parentesi arteko zenbakiak aipamena zein korreferentzia klusterraren parte den. Testuko token bat aipamen bati baino gehiagori badagokio (adibidez, [*bere*] tokena), korreferentzia-kluster ezberdinei dagokien etiketa ezberdinak " | " ikurraz bereizten dira: (1)|(2 . Aipamen baten parte ez diren tokenak azpimarra batekin adieraziko dira. *singletonen* kasuan, testu osoan aipamen bakarrak izango du kluster horri dagokion zenbakia etiketan. Korreferentzia-erlazioa dutenek, zenbaki bera izango dute etiketan. Bestalde, EPEC-KORREF euskarazko corpusean, "Euskal Herria" edo "hori dela eta" bezalako hitz anitzeko unitate lexikaleen (HAUL) tra-

Testua:	Korreferentzia:
Jonek	(1)
bere	(1)|(2
ama	2)
maite	_
du	_
.	_

tamendua egiten da, "Euskal_Herria" eta "hori_dela_eta" bezalako tokenak sortuz. Lan honetarako, eta neurona-sareetan oinarritutako hurbilpena dela eta, tratamendu hori desegitea erabaki da, HAULaren hitz bakoitza bere horretan utziz.

HAULak tratatuz:		HAULak tratatu gabe:	
Hori_dela_eta	_	Hori	_
,	_	dela	_
Euskal_Herria	(5)	eta	_
		,	_
		Euskal	(5
		Herria	5)

3.3 Sekuentziatik sekuentziarako hurbilpena

Korreferentzia-ebazpena, token bakoitzari dagokion korreferentzia-katearen etiketa esleitze moduan uler genezake, sekuentzia-etiketatzea (sequence labeling) hain zuzen. KEaren atazak ordea, POS edo NER etiketatzearen aldean, arazo bat du, etiketa kopurua ez dela mugatua. Gainera, aurrez aipatutako errepresentazioa erabiltzen bada, testu bateko korreferentzia-kate bera, hainbat etiketa baliokideren bidez adieraz daiteke.

Murriztapen horiek, sekuentzia-etiketak bikote moduan tratatu beharrean, korreferentzia kateei dagozkien etiketak domeinu ireki bat izango den sekuentziatik sekuentziara modura tratatzeko beharra eskatzen du.

Sekuentziatik sekuentziarako (*sequence to sequence*, *seq2seq*) hurbilpena, hizkuntzaren prozesamenduko hainbat atazatan arrakastaz erabiltzen den hurbilpena da, guk dakigunez, gaurdaino, inork ez du korreferentzia-ebazpena lantzeko erabili. Berriki, Raffel et al. (2019) artikuluan proposatu bezala, hizkuntzaren prozesamenduko edozein ataza (sailkatzea, etiketatzea, edota testu-sorkuntza) testutik-testurako problema batean bilakatu daiteke, eta sekuentziatik sekuentziarako edozein ataza balitz moduan landu. *Seq2seq* hurbilpena darabilen atazarik ezagunena itzulpen automatikoa da, baina testuen laburpen automatikorako eta elkarrizketa eta galdera-erantzun sistemetarako ere erabili izan da. Itzulpen automatikoko adibide bat 1. irudian dugu ikusgai.

Irudia 1: Itzulpen automatikorako sekuentziatik sekuentziarako hurbilpena.

Beraz, master amaierako lan honetan proposatzen den hurbilpena honako hau da, aipamenak detektatu eta korreferentzia klusterrak sortzeko korreferentzia-erlazioak dituzten aipamenak sailkatzaile edo multzokatze algoritmo tradizionalekin elkartzen ikasi beharrean, testu hutseko sekuentziatik, korreferentzia kateak kodetuta dituzten sekuentziak ikastea zuzenean. Beste era batera esanda, testu hutsezko jatorri-sekuentziatik, dagozkion korreferentzia kateak kodetuta dituen helburu-sekuentziara bihurtzea edo "itzultzea" izango da sare neuronalak ikasi beharreko ataza:

Jatorri-sekuentzia:	Jonek	bere	ama	maite	du	.
Helburu-sekuentzia:	(1)	(1)|(2	2)	_	_	_

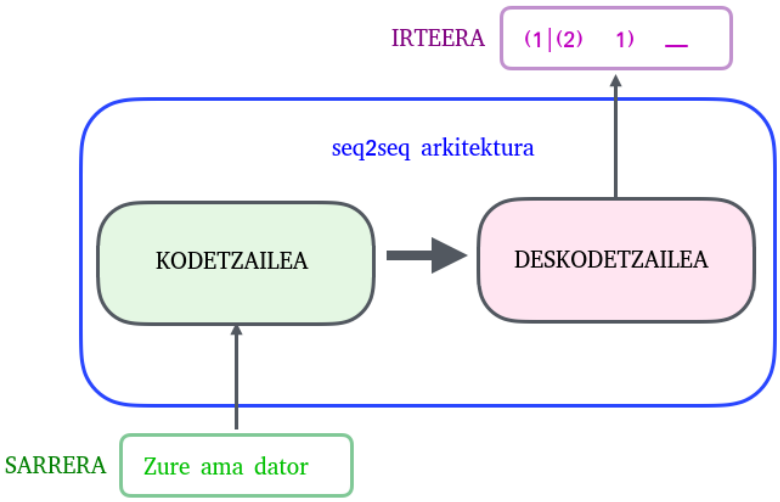
Sekuentziatik sekuentziarako hurbilpenean, sarreran sekuentzia bat hartzen du neurona sareak, eta irteeran beste sekuentzia bat itzuli. Itzulpen automatikoaren kasuan, sarrerako sekuentzia jatorri-hizkuntzako testua izaten da, eta irteeran duguna, itzulpenari dagokion helburu-hizkuntzako testua. Korreferentzia-ebazpenaren kasuan, sarreran testu hutsa izango dugu, eta irteeran parentesi bidezko egiturarekin kodetutako korreferentzia kateen sekuentzia (Ikus 2. irudia). Gainera, jatorri-sekuentziak eta helburu-sekuentziak luzera berbera izango dute eta lerrokatuta egongo dira, sarrerako hitz bakoitzaren parean, dagokion korreferentzia etiketa izango da.

Irudia 2: Korreferentzia-ebazpenerako sekuentziatik sekuentziarako hurbilpena.

3.3.1 Kodetzaile-deskodetzaile arkitektura

Hizkuntzaren prozesamenduko sekuentziatik sekuentziarako hurbilpenerako, sare neuronalak hainbat arkitektura ezberdin eduki ditzake, baina, gehienek, kodetzaile-deskodetzaile (*encoder-decoder*) bana izaten dute, 3. irudian ikus daitekeen bezala.

Kodetzaileak sarrera-sekuentziako hitzak (x_1, ..., x_n) barne errepresentazio jarrai batean (z_1, ..., z_n) bilakatzen ditu. Deskodetzaileak, z hartuta, irteerako sekuentziako etiketak (y_1, ..., y_n) sortzen ditu, aldiko elementu bat sortuz. Sekuentziako etiketa bakoitza (y_n) sortzeko pausu bakoitza, aurrez sortutako irteera-sekuentzia (y_1, ..., y_{n-1}) kontuan hartuta egiten da.



Irudia 3: Kodetzaile-deskodetzaile arkitektura.

Sekuentziatik sekuentziarako arkitektura neuronal esanguratsuenak hiru motatakoak dira: neurona-sare errepikakorretan (*Recurrent Neural Network*, RNN) oinarritutakoak (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), neurona-sare konbuluzionaletan (*Convolutional neural network*, CNN) oinarritutakoak (Gehring et al., 2017) eta Transformerra (Vaswani et al., 2017). Hiru arkitekturak, kodetzaile-deskodetzaile arkitekturan oinarritzen dira.

Guk, *"Attention is all you need"* artikuluan (Vaswani et al., 2017) plazaratutako atentzioan oinarritutako Transformer arkitektura erabiliko dugu, aurreko bi arkitekturen emaitzak hobetzea lortu duelako sekuentziatik sekuentziarako atazetan, oro har.

3.4 Transformer arkitektura

Transformerra (Vaswani et al., 2017), sekuentziatik sekuentziarako ikasketa burutzeko neurona-sareen arkitektura bat da. Horretarako, Transformerrak, auto-atentzioan (*self-attention*) oinarritutako kodetzaile-deskodetzaileak darabiltza.

3.4.1 Atentzioa eta auto-atentzioa

Hizkuntzaren prozesamenduaren arloan, atentzio-mekanismoa (*attention-mechanism*), izenak iradokitzen duen moduan, sekuentzia batean token jakinetan arreta jartzeari deritzo. Atentzio-mekanismo hori, sare neuronalen bitartez ikasten den atentzio-funtzio bat da, sekuentzia jakin batean, garrantzi handieneko hitzei pisu handiagoa eskaintzen diena.

Auto-atentzioa (*self-attention*), edo barne-atentzioa, sekuentzia batek bere buruaren gainean arreta jartzeari deritzo. Sekuentzia baten errepresentazioa kalkulatzeko, token bakoitzarentzat sekuentziaren posizio ezberdinen gainean aplikatutako atentzio-mekanismo bat da.

3.4.2 Buru-anitzeko atentzioa

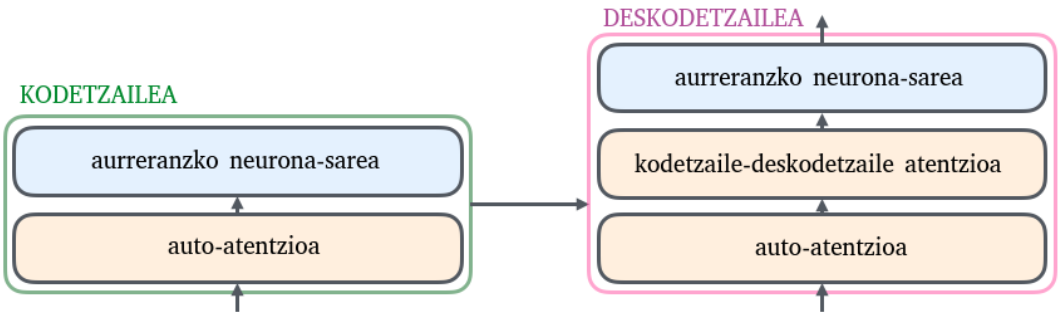
Atentzio-mekanismo gisa, atentzio-funtzio bakarra ikasi beharrean, hainbat atentzio-funtzio edo buru aldi-berean ikasteak, emaitza hobeak ematen ditu. Paraleloan ikasitako atentzio-funtzio ezberdinak uztartu egiten dira gero. Atentzio buru bat baino gehiago izateak, buru bakoitzak posizio ezberdinetako errepresentazio ezberdinetan arreta jartzea ahalbidetzen du. Buru-anitzeko atentzioa erabiltzean, atentzioaren buru bakoitzaren dimentsioak txikitu ohi dira, kostu konputazional totala handitu ez dadin.

Transformer arkitekturan, buru-anitzeko atentzio hau, kodetzaileko eta deskodetzaileko auto-atentzioan eta kodetzailetik deskodetzailera doan atentzioan aplikatzen da.

3.4.3 Kodetzailea

Transformerraren kodetzaile geruza (*encoder layer*), bi azpi-geruzek osatzen dute. Lehenengoa, buru-anitzeko auto-atentzio (*multi-head self-attention*) mekanismo bat dugu, eta bigarrena aurreranzko neurona-sare (NS) trinko (*fully-connected feed-forward network*) sinple bat da (4. Irudia).

Geruza horien artean, sarrerako informazioa bidean galdu ez dadin, hondar-konexioak (*residual connections*) daude, eta jarraian geruza-normalizazioa (*layer-normalization*) egiten da. Hori posible izateko, sarrerako hitz-bektoreen, geruzen arteko errepresentazioen eta irteerako bektoreeen dimentsioek berdinak izan behar dute.



Irudia 4: Kodetzaile-deskodetzaile geruzen barne arkitektura.

3.4.4 Deskodetzailea

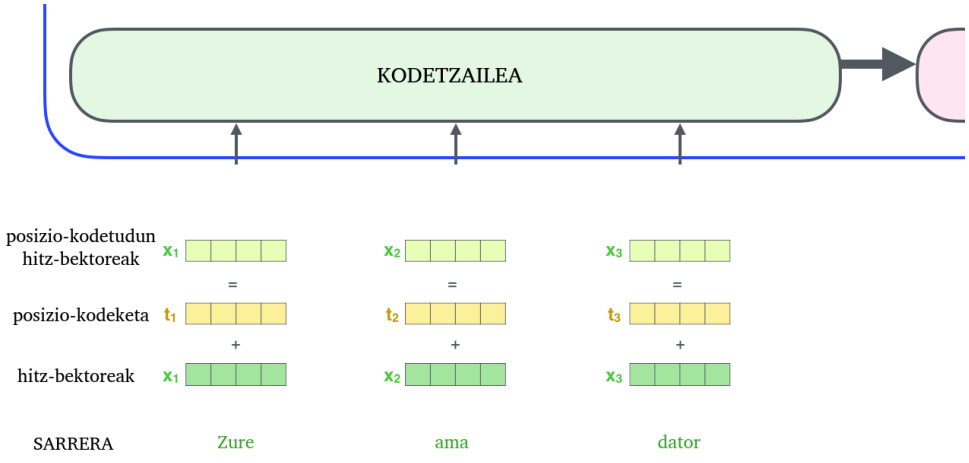
Transformerraren deskodetzaile geruzak (*decoder layer*), kodetzailearen antzeko egitura du. Buru-anitzeko auto-atentzio mekanismoa eta aurreranzko neurona-sare trinkoaren geruzak ditu, baina bi azpi-geruza horien artean, kodetzaileko irteeraren gaineko buru-anitzeko atentzio-mekanismoa du (4. irudia). Deskodetzaile geruzan ere, kodetzailean bezala, hondar-konexioak erabiltzen dira azpi-geruzen artean, jarraian geruza-normalizazioa eginez.

Gainera, deskodetzailearen auto-atentzio azpi-geruzak uneko posizioko tokena edo aurrerago (edo eskuinerago) dauden irteerako tokenen berri izan ez dezan, sekuentzian oraindik ezagunak ez diren tokenak maskaratzen (*masking*) edo ezkutatzen dira. Auto-atentzioko maskaratze horrekin, posizio bakoitzeko irteerako tokenak, kodetzaileko sarrerako tokenen, eta dagoeneko kalkulatu diren irteerako tokenen berri soilik izan dezan lortzen da.

3.4.5 Hitz-bektoreak eta posizio-kodeketa

Transformerrak sarrera-irteeran hitz-bektoreak ikasten ditu corpusetik ataza orokorra ikasteko prozesuan. Jatorri-sekuentziako sarrerako tokenak eta helburu-sekuentziako irteerako tokenak d dimentsiotako errepresentazio bektorialetan bilakatzen dira.

Sarrerarako aurre-ikasitako hitz-bektoreak erabil zitezkeen arren, hitz-bektoreak atazan bertan ikastea erabaki da. Proba esploratorioetan euskarazko *fasttext* hitz-bektoreak (Bojanowski et al., 2017) probatu dira eta emaitzetan hobekuntza nabarmenik ez da ikusi. Gainera, gerora lan hau BERT aurre-ikasitako hizkuntza ereduarekin (Devlin et al., 2019), edota honen aldaeretakoren batekin, konbinatzeko asmoa dagoenez, bide horretan sakontzeko beharrik ez da ikusi.



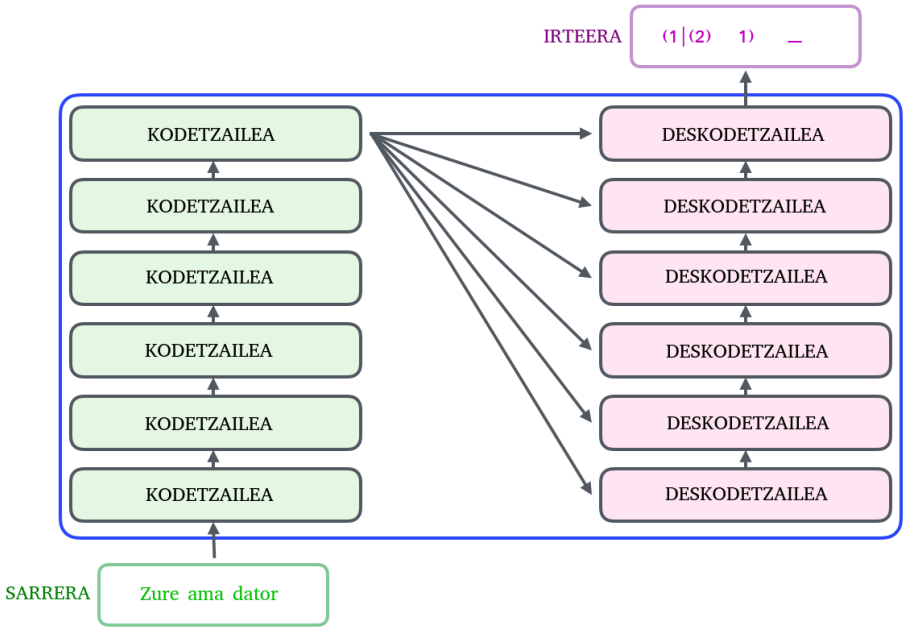
Irudia 5: Transformerraren hitz-bektoreak eta posizio-kodeketa.

Neurona-sare errepikakorretan (RNN) eta konbuluzionaletan (CNN) ez bezala, auto-atentzioa erabiltzean, sekuentziako tokenen ordena adierazteko posizio erlatibo edota ab-

solutuari buruzko informazio gehigarria esplizituki eman behar zaio ereduari. Horretarako, dagokion posizioaren kodeketa gehitzen zaio sarrerako hitz-bektore bakoitzari (5. irudia). Posizioa modu ezberdinetan kodetu daiteke, alde batetik, hitz-bektoreekin egiten den moduan, posizioari dagozkien bektoreak ikas daitezke; beste aldetik, posizio kodeketa finkoak erabil daitezke, adibidez, sinuaren eta kosinuaren funtzioen arteko konbinaziotik ateratzen diren funtzioak. Lan honetarako eraikitako transformerrean, posizioa kodetzeko bektoreak ikasi egin dira hitz-bektoreekin batera. Ondoren, posizio-kodeketa hau dimentsio berekoa izan behar duen hitz-bektoreari gehitzen zaio, tokenaren eta posizioaren informazioa bektore bakarrean biltzeko.

3.4.6 Kodetzaile-deskodetzaile geruzen pila

Aurrez aipatu bezala, kodetzaile eta deskodetzaileen sarrera eta irteerako bektoreak (hitz-bektoreak edo barne errepesentazioak) dimentsio berdinekoak direnez, honek kodetzaile-deskodetzaile geruzak bata bestearen gainean pilatzea ahalbidetzen du (Ikus 6. irudia). Geruzak pilatze hori praktika ohikoa da, Transformerraren eredu normalak 6na kodetzaile-deskodetzaile geruza pilatzen (*encoder-decoder stacking*) dituzte, eta BERT eredu handiak, adibidez, 24na geruza.



Irudia 6: Transformerreko kodetzaile-deskodetzaile geruzen pila.

3.5 Dokumentuaren eta Transformerreko sekuentzien luzera

Transformer arkitekturak aipatu gabeko arazo bat dakarkigu ordea, darabilen atentzioak paralelizatzeko aukera ematen duen arren, testuko token guztien gaineko eragiketa iza-

nik, prozesatu beharreko testuaren luzera handitzearekin batera, prozesatzeko kostua, eta eredua gordetzeko GPUaren memoria beharra handitu egingo da. Hori dela eta, Transformerraren sarreran dokumentuak osorik sartu beharrean, zatika egingo da.

3.5.1 Zatitze algoritmoa

Sarrerako dokumentuak zatitzeko, dokumentua esaldi baten erditik banatuz gero, aipamen bat bi sekuentzia ezberdinetan zati genezake, eta horrek, helburu-sekuentzian korreferentzia-kateak ondo ireki edo itxi gabe geratzea ekarriko luke. Honenbestez, dokumentuak esaldika bereizi dira, baina sarrera-irteerako sekuentziak zenbat esaldiz osatuko diren erabakitzea geratzen da oraindik. Horren aurretik azter dezagun zatitze prozesu hori nola egiten den.

Har dezagun honako testu sintetiko hau:

Ni	nator	.	Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.
(40)	_	_	(12)	_	_	(8)	_	_	(40)	(6|(8)	6)	_	_	_

Demagun lau esaldi laburrez osatutako testu hori sekuentzia laburragoetan zatitu nahi dugula eta esaldiak hirunaka prozesatzea erabakitzen dugula. Dokumentua banatzeko garaian, esaldi bakoitza, ondorengo bi esaldirekin elkartuko genuke (Ikus 4. taula): e_1-e_3, e_2-e_4, e_3-e_5, ..., e_{n-2}-e_n.

Ni	nator	.	Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.
(40)	_	_	(12)	_	_	(8)	_	_	(40)	(6|(8)	6)	_	_	_
Ni	nator	.	Hark	daki	.	Zuk	daukazu	.						
(40)	_	_	(12)	_	_	(8)	_	_						
			Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.
			(12)	_	_	(8)	_	_	(40)	(6|(8)	6)	_	_	_

Taula 4: Zatitze algoritmoa.

Jarraian, esaldi hirukote bakoitzeko korreferentzia-etiketak berridazten dira, etiketaren zenbakizko balioak hasieratuz, beti ere, hirukote barneko korreferentzia-erlazioak mantenduta. (Ikus 5. taula).

Horretarako, sekuentziaren ezkerretik hasita, korreferentzia-etiketako zenbakiak zenbaki txikiagoekin ordezkatuko dira (1etik hasita), beti ere, sortu dugun sekuentzian dauden korreferentzia-erlazioak mantenduz. Horrela, sekuentzia guztiak 1etik sekuentziako kluster kopururainoko zenbakiak dituzten korreferentzia-etiketez osatuko dira, eta hauek, posizioarekiko gorakorrak izango dira, aurrez agertutako aipamen batekin korreferentzia-erlazioa duten aipamenei dagokienak salbuespen izanik.

Berridazketa honek, neurona-sarearen ikasketa prozesua erraztuko du, sekuentzia guztietan patroi gorakor berdina dagoelako. A eranskinean benetako dokumentu baten zati bati dagokion adibide osatuagoa dago.

Ni	nator	.	Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.
(40)	_	_	(12)	_	_	(8)	_	_	(40)	(6|(8)	6)	_	_	_
Ni	nator	.	Hark	daki	.	Zuk	daukazu	.						
(1)	_	_	(2)	_	_	(3)	_	_						
			Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.
			(1)	_	_	(2)	_	_	(3)	(4|(2)	4)	_	_	_

Taula 5: Etiketen berridazketa zatitze algoritmoan.

3.5.2 Elkartze algoritmoa

Transformer ereduak N esaldiko luzerako sekuentzietatik ikasiko du korreferentzia-ebazpena egiten, eta sare neuronalak sarrerako sekuentziari dagokion korreferentzia-kateen inferentzia (I) itzuli ondoren, N esalditako zati horiek elkartu beharko dira azkeneko emaitzan (E), non dokumentu osoaren korreferentzia-ebazpena gordeko den.

Elkartze algoritmoa hobeto ulertzeko itzul gaitezen zatitze algoritmoa azaltzean aipatu dugun adibidera. Dokumentua banatzeko garaian, esaldiak hirunaka hartu ditugu, aldiko esaldi berri bat hartuz, e_1-e_3, e_2-e_4, e_3-e_5,... Behin neurona sarean ikasketa eta inferentzia (I) eginda, esaldi horiek berriro ere elkartu egin beharko dira. Horretarako, lehenik, lehen sekuentziako esaldiak (e_1-e_3) bere horretan hartuko genituzke eta azken emaitzan (E) idatzi (Ikus 6. taula).

	Ni	nator	.	Hark	daki	.	Zuk	daukazu	.
I_1:	(1)	_	_	(2)	_	_	(3)	_	_
E:	**(1)**	_	_	**(2)**	_	_	**(3)**	_	_

Taula 6: Elkartze algoritmoaren hasiera.

Hortik aurrera, sekuentzia berri bat daukagunean, (e_2 - e_4), azken esaldia (e_4) soilik erantsiko diogu aurrez prozesatutako emaitzari (7. taulan letra lodiz), eta beste 2 esaldiak (e_2 - e_3), korreferentzia-erlazioak lotzeko bakarrik erabiliko dira. Sekuentzien arteko korreferentzia-erlazioen elkartze hori nola egiten den hobeto ulertzeko, azter dezagun 7. taulako adibidea.

e_2 - e_4 sekuentziaren inferentzia (I_2) egin ostean, lehen bi (N-1) esaldietan (e_2 - e_3), aurrez emaitzan (E) idatzitako aipamen berdinik dagoen aztertu behar da lehenik. Aipamen berdinak izateko, aipamen bakoitzaren muga diren tokenak sekuentziako posizio berdinetan egon behar dute lerrokatuta. Horien kasuan, inferentzian dauden korreferentzia klusterrak, emaitzan idatzitakoekin lotzen dira ([Hark: $(1)_{I_2} - (2)_E$] ; [Zuk: $(2)_{I_2} - (3)_E$]), eta ordezkapen hori inferentziako sekuentziako azken esaldiko korreferentzia-etiketa guztietan egiten da [zure: $(2)_{I_2} - (3)_E$]. Korreferenteak ez diren klusterren kasuan, kluster zenbaki berri bat esleitzen zaio, azken esaldiko "Nik" hitzaren kasuan $(3)_{I_2} - (4)_E$ aldaketan bezala.

Lortzen den emaitza (E), 8. taulako adibidearen beheko aldean ikus dezakegun urre-

	Ni	nator	.	Hark	daki	.	Zuk	daukazu	.							
I_1:	(1)	_	_	(2)	_	_	(3)	_	_							
E:	(1)	_	_	(2)	_	_	(3)	_	_							
				Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.	
I_2:				(1)	_	_	(2)	_	_	(3)	(4|(2)	4)	_	_	_	
E:	(1)	_	_	(2)	_	_	(3)	_	_	**(4)**	**(5|(3)**	**5)**	_	_	_	

Taula 7: Elkartze algoritmoa.

patroiaren (U) oso antzekoa da. Akats bat dago, lehen esaldiko *Ni* eta azken esaldiko *Nik* aipamenak, biak korreferenteak izan beharko lukete (urre-patroian (40) etiketarekin). Baina, algoritmoak etiketa eta, beraz, korreferentzia kluster ezberdinak, esleitu dizkie, (1) eta (4), sekuentziaren luzeratik at geratzen delako haien arteko erlazioa. Salbuespen horrekin, gainontzekoan emaitza zuzena izango litzateke, korreferentzia kluster berak adierazteko korreferentzia-etiketa ezberdinak (baina baliokideak) erabiltzen diren arren.

	Ni	nator	.	Hark	daki	.	Zuk	daukazu	.							
I_1:	(1)	_	_	(2)	_	_	(3)	_	_							
E:	(1)	_	_	(2)	_	_	(3)	_	_							
				Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.	
I_2:				(1)	_	_	(2)	_	_	(3)	(4|(2)	4)	_	_	_	
E:	**(1)**	_	_	(2)	_	_	(3)	_	_	**(4)**	(5|(3)	5)	_	_	_	
	Ni	nator	.	Hark	daki	.	Zuk	daukazu	.	Nik	zure	anaia	jo	dut	.	
U:	**(40)**	_	_	(12)	_	_	(8)	_	_	**(40)**	(6|(8)	6)	_	_	_	

Taula 8: Elkartze algoritmoaren emaitza eta urre-patroiaren konparaketa.

3.5.3 Sekuentzien luzera (N) finkatzea

Eredua ikasteko dokumentuak N esaldika banatu aurretik, Nren balioa erabaki behar dugu. N txikia hartuta, demagun 2, transformerrean ez dugu memoria arazorik izango, baina esaldiak elkartzeko garaian, ondoz ondoko esaldietan dauden korreferentzia-erlazioak soilik hartuko dira kontuan, eta hori baino distantzia luzeagoko korreferentzia-erlazioak galdu egingo dira, tartean zubi-lanak egingo dituen korreferentzia-kluster bereko aipamenik ez badago.

N oso handia hartuko bagenu, dokumentuak osorik, edo ia osorik tratatu ahalko genituzke, korreferentzia-erlazioen distantzia posiblea asko handituz, baina segituan memoria arazoak izango genituzke transformer eredu osoa sekuentzia luzeekin GPU bakarrean ikasteko.

Nren balioa erabakitzeko, EPEC-KORREF euskarazko corpusean, garapenerako atalarekin, proba batzuk egitea erabaki da. Dokumentua N esalditan banatu da, eta eredua

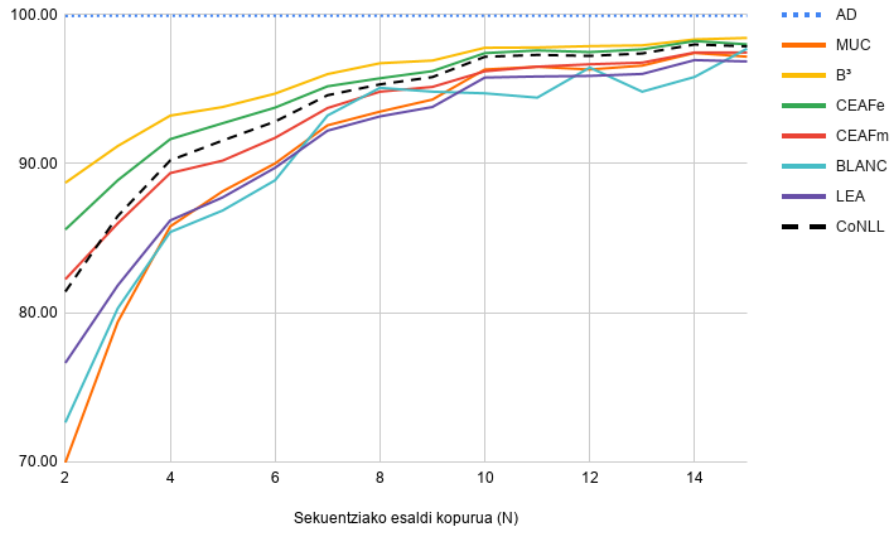
ikasi eta inferentzia egin beharrean, urre-patroiko korreferentzia kateak erabili dira berriro elkarketa egiteko. Horrela, banatze-elkartze algoritmo honekin zenbateko galera dugun aztertu da.

N 2 eta 14 zenbakien artean probatu da: 2 zenbakia, esaldien artean korreferentzia-kateak lotzeko, gutxienez bi esaldi behar direlako, eta 10 esaldiko langa igaro ondoren, aldiko ia dokumentu osoak diren sekuentziak prozesatzen ari garenez gero, F1 balioa konbergitzen hasten da. 14.era iristean esplorazioa moztea erabaki da.

Lortutako emaitzak, 9. taulan daude. Aipamen-detekzioan (AD) %100 zuzen dago, aipamenak urre patroitik hartutakoak direlako, eta elkartze algoritmo honetan aipamenen galerarik ez dagoelako. CoNLL metrika orokorrean, berriz, banatze-elkartze algoritmoko sekuentziaren luzera handitu ahala, F1 balioak gorantz egiten du.

N	AD	CoNLL
2	100	81,41
3	100	86,49
4	100	**90,22**
5	100	91,56
6	100	92,83
7	100	**94,60**
8	100	95,33
9	100	95,82
10	100	**97,18**
11	100	97,31
12	100	97,24
13	100	97,41
14	100	98,01

Taula 9: Nren eragina zatitze-elkartzean.



Irudia 7: Esaldi kopuruaren eragina zatitze-elkartzean grafikoki.

9. taulan, zein 7. irudian ikus dezakegun bezala, esaldiak binaka (N = 2) soilik banatuta CoNLL metrikan (grafikoan lerro eten beltzez) 81 F1 puntutik gora lor ditzakegu (hau urre-patroia izanik, benetako eredu bat ikastean goi-sabaitzat har genezake). N handitu ahala, CoNLL metrikak gorantz nabarmen egiten duela ikus daiteke, distantzia luzeagoko korreferentzia-erlazio gehiago lotzen direlako. N = 4, N = 7 eta N = 10 esaldi hartzean aurrekoekiko CoNLL metrikan hobekuntza nabarmena lortzen da, beraz Transformer ereduaren memoria mugak baimentzen badigu hautagai aproposak izan litezke azken sistemarako, beti ere emaitza hauek garapenerako corpusari lotutakoak direla jakinda.

3.6 Inplementazioa eta baliabide konputazionalak

Egin diren esperimentu ezberdinetarako, 3.4. atalean azaldu den Transformer arkitektura, 2 esalditako luzerako sekuentziekin entrenatu da. Transformer arkitektura, python pro-

gramazio lengoaiako ikasketa sakonerako Pytorch (Paszke et al., 2019) ingurunean kodetu da.

Adam optimizatzailea erabili da (0,0005 ikasketa-erritmoarekin) Transformerretan erabili ohi den Noam (Adam-en bertsio hobetua baina ezegonkorragoa, ikasketa-erritmo aldakorra duena) erabili beharrean, hiper-parametroien doiketa errazteko. Bestalde funtzio sinusoidaletan oinarritutako posizio-kodeketa finkoa erabili beharrean, posizio-kodeketa ikasi egin da. Oinarrizko Transformerraren arkitekturari bi aldaketa hauek egin zaizkio, BERTen Devlin et al. (2019) familiako Transformerrek optimizatzaile eta posizio kodeketa hori erabiltzen baitute.

Transformerra bezalako eredu neuronal handiak datu-multzo handietatik ikasteak, kostu konputazional handia dakar. Sare neuronalak entrenatzeko, grafikoak prozesatzeko unitateak (GPUak) erabiltzen dira, ikasketa prozesua azkartuz. Lan honetarako Transformer ereduaren garapenerako eta ikasteko 12GBeko memoria duen TITAN Xp GPU bat erabili da. Garapena azkartzeko, corpusen garapen atalak erabili dira ikasketa noiz moztu (*early stopping*) erabakitzeko.

Sekuentziaren esalditako luzera handitu ahala, emaitza hobeak lortzen dira, baina GPUen memoria handiagoa eskatzen du. Esaldiak binaka bakarrik elkartuta ere KEaren ataza neurri batean ikas daitekeela ikusita (CoNLL metrikan 81 puntuko goi-sabaiarekin), hasiera batean N = 2 finkatu da. Hori, KEa sekuentziatik sekuentziara ikasteko Transformer eredu bat erabiltzeak zentzurik baduen ikusteko, eta arkitektura honen bideragarritasuna aztertzeko nahikoa izango da. Horrela, denbora eta memoria aldetik ikasketa eta esperimentazioa erraztuko dira. Gerora hau handitzeko muga bakarra eredua GPUen memorian sartzea izango litzateke.

4 Esperimentazioa

Atal honetan, egin diren esperimentu ezberdinak azalduko dira. Lehenik, Transformer arkitekturak KEean zer nolako portaera eta joera duen aztertzeko, ingeleserako KEa PreCo corpusa erabili da (esp0). Hurrengo esperimentua, euskararako egin da eta horretarako EPEC-KORREF corpusa erabili da (esp1). Jarraian datu gehikuntzako teknika ezberdinak aztertu dira euskararako: ausazko esaldi parekaketa (esp2), sasi-etiketatzea (esp3-6) eta hizkuntza arteko ikasketa (esp7) landu dira. Azkenik, euskararako emaitza onenak eman dituen sistemari, BPE segmentazioa aplikatu zaio (esp8).

4.1 0 esperimentua: ingeleserako korreferentzia-ebazpena

Aurrez euskararako korreferentzia-ebazpena hurbilpen neuronalekin aztertu denean (Urbizu et al., 2019b,a), ez dira emaitza onak lortu, eta hori atazan ikasketarako eskuragarri dagoen corpusaren tamaina txikia delako gertatzen dela ondorioztatu da. Beraz, hurbilpen honetan ere, gauza bera gerta daitekeela aurreikusita, lehenik eta behin, 3. atalean azaldutako sekuentziatik sekuentziarako hurbilpena KEerako erabilgarria den edo ez aztertzeko, baliabide askoko hizkuntza batetara jo da, ingelesera, lortutako emaitzek, datu-eskasiaren eraginik izan ez dezaten. Horretarako, ingelesezko corpus handi bat, PreCo corpusa erabili da ikasketarako.

Hiper-parametroen esplorazioa eginda, 128 dimentsiotako hitz-bektoreak, 512 dimentsiotako aurreranzko azpi-geruza, 2 kodetzaile-deskodetzaile geruza, 4 buruko atentzioa eta 256ko batch tamaina finkatu dira. Transformer eredua GPUaren memorian sar zedin, sarrerako hiztegia mugatu egin behar izan da (64K token), irteerakoa aldiz osorik mantendu da (37K token). Guztira 19 miloi parametroko neurona-sarea entrenatu da.

Ingeleserako KEeko atazan entrenatutako ereduak inferentzian itzultzen dituen sekuentziak ikusita, Transformer arkitektura KEaren ataza ikasteko gai dela ikus dezakegu, aipamen ia denak egoki identifikatuz, eta sekuentzia barneko korreferentzia-erlazio gehienak egoki lotuz. Hona garapen atalerako inferentzian lortutako adibide bat:

```
jatorria   = Divorced from  my  mother when I was 11 ,   my   dad
helburua   =     _     _  (1|(2)   1)    _ (2) _ (3) _ (2)|(4  4)
iragarpena =     _     _  (1)|(2   2)    _ (1) _ (3) _ (1)|(4  4)

could n't be around  his  kids as often as he would have liked .
  _    _  _     _  (5|(4)  5)  _    _   _  (4)  _    _     _   _
  _    _  _     _  (4)|(5  5)  _    _   _  (4)  _    _     _   _

Money was also tight ; even weekend visits were rare .
 (6)   _    _    _   _  _   (7|(8)    7)    _    _   _
 (6)   _    _    _   _  _   (7|(8)    7)    _    _   _
```

PreCo corpusaren garapenerako azpi-atal osoko inferentzia egin eta sekuentziak dokumentuka elkartu ondoren, 10. taulako emaitzak lortu dira (metrikak 5.1. atalean azaltzen dira). Ikus dezakegunez, emaitza nahiko onak lortu dira, CoNLL metrikan 59 puntuko

F1 balioa lortuz. Emaitza hori, bi esaldiko sekuentziak banatu eta elkartzetik lor genezakeen emaitza hoberenetik (urre-patroitik) 13 puntutara geratzen da. Honenbestez, korreferentzia-ebazpenerako sekuentziatik sekuentziarako hurbilpen hau bideragarria dela, eta Transformer arkitektura erabiliz, KEa ikertzen jarraitzea zentzuzkoa dela ondorioztatu dugu.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp0 (EN)	87,52	45,13	66,50	59,04	65,61	44,17	51,52	59,08
goi-sabaia	97,27	58,92	79,01	70,69	79,55	56,98	66.62	72,49

Taula 10: Ingeleserako sekuentziatik sekuentziarako KE.

4.2 1. esperimentua: euskararako korreferentzia-ebazpena

Proba batzuk egin ondoren, ingeleserako erabilitako hiper-parametro berberak erabili dira euskararako, sarrerako hiztegia (8.000 token) eta irteerakoa (400 token) osorik erabiliz, mugatu beharrik gabe. Guztira 2,4 miloi parametroko neurona-sarea entrenatu da.

Neurona-sarea EPEC-KORREF corpusetik ateratako sekuentzietan (1.600 sekuentzia) entrenatu ondoren, garapena atalean inferentzia egin da. Oraingoan, ordea, euskarazko corpusean lortutako emaitzak oso bestelakoak izan dira, ondorengo corpuseko adibidean ikus daitekeen bezala:

```
jatorria   = Finalerdietan , berriz , Arroxela kanporatu zuen , 14-12 irabazita .
helburua   =     (1)       _    _   _    (2)       _      _   _  (3)      _     _
iragarpena =     (1        1)  _   _    _         _      _   _   _      (2     2)

Asko kosta zitzaien lapurtarrei garaipena lortzea , eta 6-6 heldu ziren atsedenaldira .
 _     _      _         (4)        (5)       _    _  _  (6)   _     _         (7)     _
 _     _      _          _          _        _    _  _   _    _     _          _      _
```

Lehen begiradan ikus dezakegunez, euskararako KEerako entrenatutako Transformerra ez da gai KEaren ataza egoki egiteko. Ereduak itzulitako sekuentziak elkartu ondoren metrika ezberdinetan lortutako F1 balioak, 11. taulan daude bilduta. CoNLL metrikan, 2,23ko F1 balioa lortu da, banatze-elkartze algoritmoak 2 esaldiko sekuentzientzat urre-patroiarekin lortutako goi-sabaitik oso urrun.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp1 (EU)	3,64	0	2,75	3,59	3,92	0,04	2,11	2,23
goi-sabaia	100	69,93	88,72	85,58	82,24	72,62	76,62	81,41

Taula 11: Euskararako sekuentziatik sekuentziarako KE.

Laburbilduz, euskararako KEerako ereduarekin lortutako emaitzak oso kaxkarrak dira, Transformerra, ez da gai ataza ikasteko. Emaitza horiek, euskararako KEerako dauden

baliabide eskasiak eraginda dela ondorioztatu da, 12,5 miloi hitzeko corpusarekin emaitza onak ematen dituen hurbilpen honek, ez duelako funtzionatzen 61.000 hitzeko corpus txikiarekin.

4.3 2-7 esperimentuak: datu gehikuntza

Euskararako eta ingeleserako lortutako emaitzen arteko aldea ikusita, esan bezala, hurbilpen honek ikasketarako datu gehiago behar dituela ondorioztatu da. Datu eskasiari aurre egiteko, datu gehikuntzako (*data augmentation*) hainbat teknika proposatuko dira atal honetan.

4.3.1 2. esperimentua: ausazko esaldi-parekaketa (AEP)

Lehenik eta behin, euskararako EPEC-KORREF corpusa hartu, eta dokumentu bakoitzean, esaldiak binaka ondoz ondokoarekin elkartzeaz gain, dokumentuko gainontzeko esaldi guztiekin ere parekatu dira, bien arteko ordena mantenduz. Horrela, ikasketarako 1.600 sekuentzia izan beharrean, 18.000 sekuentzia ditugu. Hori eginik, sarrerako testu eta hiztegi berririk ikasiko ez duen arren, korreferentzia etiketak zuzenago osatzen ikas dezake sistemak. Horrela lortutako emaitzak (12. taulan ikusgai), zer edo zer hobeak diren arren, (aipamen-detekzioan 3 puntu, eta CoNLL metrikan puntu bat), gure helburutik oso urrun daude.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp1 (EU)	3,64	0	2,75	3,59	3,92	0,04	2,11	2,23
esp2 (EU+AEP)	6,27	0	4,56	6,00	5,90	0,18	3,09	3,49

Taula 12: Datu gehikuntza: ausazko esaldi-parekaketaren (AEP) emaitzak.

Ausazko esaldi parekaketarekin ikasketarako sekuentzia kopurua handituz lortutako emaitzak, oraindik ere kaxkarrak diren arren, zertxobait hobeak dira, aipamen-detekzioan 3 puntuko igoerarekin eta CoNLL metrikan puntu batekoarekin.

4.3.2 3-6 esperimentuak: sasi-etiketatzea (SE)

Datu gehikuntzarako aztertu den bigarren hurbilpena, sasi-etiketatzea (*pseudo-labeling*) deritzon erdi-gainbegiratutako teknika izan da. Sasi-etiketatzearen teknika, dauzkagun datuekin ikasketa automatikoko (sakonekoak barne) eredu bat atazan entrenatzean datza, gero eredu horrekin automatikoki etiketatutako instantzia berriak ikasketan erabiltzeko. Prozesu hori iteratiboki aplika daiteke, *bootstraping* deituriko teknika aplikatuz. Gaur egun, itzulpen automatikoko sistemak entrenatzeko, antzeko zerbait egiten da, *back-translation* izeneko teknikarekin (Sennrich et al., 2016a). KEaren atazan sasi-etiketatzea aplikatzeko orduan, daukagun datuetatik ikasitako Transformer eredua erabili beharrean corpus berria sortzeko (*bootstrapping* teknika), erregelatan oinarritutako euskararako KEerako sistema

(Soraluze et al., 2019) erabili da. Alde batetik, emaitzarik onenak ematen dituen sistema delako, eta bestetik, urre-patroia oinarri izanda datu berdinen gainean behin eta berriz ikasita baino, giza ezagutzan oinarritutako erregelatan ikasitako sistemarekin lortutako corpus sasi-etiketatua, eta EPEC-KORREF corpusa konbinatuz emaitza hobeak espero direlako.

Modu horretan, KEa automatikoki etiketatuta duten bi corpus sortu dira: Elhuyar Web corpusa (ikus 3.1.2. atala) eta Elhuyar QTleap corpusa (3.1.3. atala), SE1 eta SE2 hurrenez hurren.

Sasi-etiketatzearekin egin den lehen esperimenturako (esp3), ausazko esaldi-parekaketa eginda duen EPEC-KORREF eta sasi-etiketatutako Elhuyar Web corpusa elkartu eta ikasketarako erabili dira (39K sekuentzia). Aurreko esperimentuetako Transformer arkitektura eta hiper-parametro berdinak mantendu dira. Datu horietatik ikasita, 1. esperimentuko (EU) emaitzak hobetu diren arren, EPEC-KORREF corpusarekin ausazko esaldi-parekaketa egindakoaren (esp2, +AEP) antzekoak dira lortutako emaitzak (13. taula).

Hurrengo esperimenturako (esp4), EPEC-KORREF eta sasi-etiketatutako Elhuyar Web corpusa elkartu dira ikasketarako, bi corpusetan ausazko esaldi-parekaketa egin ondoren. Hori eginda, ikasketarako instantzia kopurua asko handitu da (591K sekuentzia), eta emaitzak askoz hobeak dira, 13. taulan ikus daitekeen moduan. Aipamen-detekzioan 42 puntuko hobekuntza, eta CoNLL metrikan 22 puntukoa lortu dira. Hala ere, Transformerrak metrika gehienetan hobekuntza handiak izan dituen arren, MUC metrikak adierazten duenez, korreferentzia-erlaziorik ez du sortzen. Hori, datu-gehikuntzako hurbilpen ezberdinekin sortutako sekuentzietan, korreferentzia-erlazio oso gutxi daudelako izan daiteke.

Jarraian, esperimentu berri batean (esp5), EPEC-KORREF (+AEP) eta sasi-etiketatutako Elhuyar QTleap corpusa elkartu dira (521K sekuentzia). Corpus handiagoa erabiltzearekin batera, jatorri-sekuentzien hiztegia ere handitu egin da, eta hiztegi hori mugatu den arren (70K), batch tamaina erdira murriztu behar izan da, ikasketa mantsotuz, eredua GPUaren memorian sartzeko. Hori eginik, laugarren esperimentuko antzeko emaitzak lortu dira, puntu t'erdiko irabaziarekin (ikus 13. taula).

Sasi-etiketatzearen atal honetan egin den azken esperimentuan (esp6), EPEC-KORREF (+AEP), sasi-etiketatutako Elhuyar Web corpusa (+AEP) eta sasi-etiketatutako Elhuyar QTleap corpusa elkartu dira (1,1M sekuentzia). Euskararako eskura ditugun corpus guztiak elkartuta, aurreko esperimentuetan baino emaitza hobeak lortu dira, CoNLL metrikan 4 puntuko hobekuntzarekin.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp1 (EU)	3,64	0	2,75	3,59	3,92	0,04	2,11	2,23
esp2 (EU+AEP)	6,27	0	4,56	6,00	5,90	0,18	3,09	3,49
esp3 (SE1)	6,54	0	5,07	6,41	6,54	0,24	3,72	3,87
esp4 (SE1+AEP)	48,83	0	38,74	37,82	39,03	12,36	25,45	25,92
esp5 (SE2)	52,88	0,36	41,60	39,38	40,32	14,55	26,49	27,43
esp6 (SE)	60,75	0,53	47,84	44,02	45,70	18,83	30,41	31,35

Taula 13: Datu gehikuntza: sasi-etiketatzearen (SE) emaitzak.

Datu gehikuntzaren bitartez, eredua ikasteko erabilitako sekuentzia kopurua handitzearekin batera emaitzetan hobekuntzak lortu dira. EPEC-KORREF corpusean ausazko esaldi parekaketa eginda (esp2) eta Elhuyar Web corpusa sasi-etiketatua gehituta (esp3), emaitzetan hobekuntza txikia lortu den arren, datu-gehikuntzarako bi teknikak konbinatzean (esp4), ikasketarako datu-kopurua handitzearekin batera, lehen aldiz, emaitzetan igoera handiak lortu dira (aipamen-detekzioan 45 puntuko igoera, eta CoNLL metrikan 23koa). Antzeko emaitzak lortu dira EPEC-KORREF corpusaz gain sasi-etiketatutako Elhuyar QTleap corpusa erabiltzean (esp5). 4. eta 5. esperimentuetako ikasketarako datuak elkartuz, igoera nabarmen bat lortu da, aipamen-detekzioan 60,75 puntuko F1 balioa lortu da eta CoNLL metrikan 31,35 puntuko F1 balioa.

4.3.3 7. esperimentua: hizkuntza-arteko ikasketa

Sasi-etiketatutako corpusarekin, datu kopurua handitu dugu. Baina, sarrerako jatorri-sekuentzien kalitatea ona izan arren, helburu-sekuentziari dagokion sekuentziak kalitate eskasa dute, automatikoki lortutakoak baitira. Transformerrak helburu-sekuentziak sortzeko duen gaitasuna hobetzeko asmoz, baliabide gehiagoko beste hizkuntza bateko corpusak ere ikasketarako erabiltzea aztertu da.

Ingelesezko corpusak askoz ere handiagoak izanik, PreCo corpusa aukeratu da, corpus handia delako, *singleton*ak ere etiketatuak dituelako (CoNLL-2012k ez bezala), eta corpusa eskuraerraza delako.

Beste hizkuntza bateko corpusetik euskararako korreferentzia-ebazpena ikasi ahal izateko, lehenik erdarazko corpusa itzulpen automatikoa erabiliz euskaratzea aztertu da. Hurbilpen hori, ordea, ez da bideragarria, esaldiak osorik itzuliz gero, testuak korreferentzia-ebazpenari dagokion etiketekin duen lerrokatzea hautsiko litzatekeelako. Esaldien itzulpena aipamenen mugak errespetatuz zatika egitea, berriz, ez da lan erraza.

Gaur egun, neurona-sareetan elearteko hurbilpenetan, elearteko hitz-bektoreak, edota elearteko hizkuntza ereduak erabiltzen dira. Lan honetarako, halaber, etorkizunean bide hori jorratzeko asmoa dagoen arren, master amaierako lan honetan ingelesezko corpusa bere horretan erabili da, elearteko aurre-ikasitako hitz-bektore edo hizkuntza ereduen azterketa etorkizunerako lanerako utziz.

Ingelesezko corpusa, 6. esperimentuan erabilitako sasi-etiketatutako euskarazko corpusekin (SE1 eta SE2) eta EPEC-KORREF corpusarekin (+AEP) konbinatu da (1,2M sekuentzia). Bi hizkuntzetako hiztegiak konbinatu ahal izateko, neurona-sarearen hiztegiaren tamaina handitu egin da, eta eredua memorian sartu zedin batch tamaina 64ra txikitu. Horrela, 14. taulan ageri diren emaitzak lortu dira. Euskarazko datu gehikuntzarekin soilik baino emaitza hobeak lortu dira metrika guztietan, hala ere, elearteko sistema, oraindik ez da gai korreferentzia-erlazioak zuzen lotzeko.

4.4 8. esperimentua: BPE segmentazioa

Sasi-etiketatzearen atalean dagoeneko Transformerraren sarrerako testuaren hiztegia mugatu behar izan da, eredua GPUaren memorian sartu ahal izateko. Irteerako sekuentzien

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp6 (SE)	60,75	0,53	47,84	44,02	45,70	18,83	30,41	31,35
esp7 (SE+EN)	64,41	1,61	50,92	46,46	48,25	21.68	32,47	33,59

Taula 14: Datu gehikuntza: sasi-etiketatuari (SE), ingelesezkoa (EN) gehituta lortutako emaitzak.

etiketen hiztegiarekin ez da arazorik egon, kopuruz mugatuagoak dira eta. Bi hizkuntzetako corpusetatik aldi berean ikastean, hiztegia handitu da, bi hizkuntzetako hiztegia ikasi ahal izateko. Hori egiteak memoria gehiago eskatzen du, baina hizkuntzaren prozesamenduko beste ataza batzuetan egin ohi den moduan, hiztegi zabalagoa barne hartzeko, BPE segmentazioa erabil dugu.

BPE (*Byte Pair Encoding*) teknikak (Sennrich et al., 2016b), testuko hitzak segmentuetan banatzen ditu, testuko hiztegi osoa ikasteko aukera emanez. BPE segmentazioak, euskara bezalako hizkuntza eranskarien kasuan, emaitza hobeak ematen ditu, hitz barruko lema eta atzizkiak bereizteko gai delako (nahiz eta maiztasun handieneko konbinazioak osorik azalduko diren).

BPE segmentazioa, itzulpen automatikoa bezalako atazetan sarrerako zein irteerako hiztegiei aplikatzen zaie. KEerako sekuentziatik sekuentziarako hurbilpen honetan, ordea, irteerako hiztegia (korreferentzia-etiketak) handiegia ez denez, sarrerako testuan soilik aplikatzea erabaki da. Hona BPEk egiten duen hitzen segmentazioaren adibide bat:

```
Espainiako estatuan , garapen txikiena duten herrialdeen artean
Andaluz@@ ia , Asturias , Kan@@ ariar irl@@ ak , Kantabr@@ ia ,
Gaztela , Valentzia , Extrem@@ adura , Galizia , Murtz@@ ia ,
Ce@@ uta eta Mel@@ illa daude .
```

Hasiera batean jatorri-sekuentzian BPE aplikatu eta helburu-sekuentzia bere horretan utzi da, sekuentziatik sekuentziarako atazetan ohikoa da eta luzera ezberdinetako sekuentzia bikoteak lantzea. Egin diren esperimentuetan ordea, BPE aplikatu gabekoetan baino emaitza nabarmen okerragoak lortu dira. Hori, jatorri-sekuentziako hitzak banatzean, eta helburu-sekuentziak bere horretan uztean, haien arteko lerrokatzea galdu delako gertatu dela ondorioztatu da. Horregatik, helburu-sekuentziei, jatorri-sekuentziei dagokien banaketa berdina aplikatu zaie:

```
BPE gabe:
    ... Kanariar irlak , Kantabria ...
          (1       1)  _    (2)
BPErekin:
    ... Kan@@ ariar irl@@ ak , Kantabr@@ ia ...
         (1     _     _   1) _   (2      2)
```

Helburu-sekuentziei jatorri sekuentzien banaketa berdina aplikatzean, lerrokatuta geratzen dira, eta Transformerrak hobeto ikasten du sekuentziatik sekuentziarako ataza. Datu

gehikuntzako teknika guztiak (ausazko esaldi parekaketa, sasi-etiketatzearena eta elearteekoa) konbinatzen dituen esperimentuari BPE segmentazioa aplikatu zaio.

BPE segmentazioa ikasteko, euskarazko eta ingelesezko corpusak nahastu dira, baina euskarazkoari pisu handiagoa eman zaio. Euskara hizkuntza eranskaria izanik, ingelesak baino hiztegi zabalagoa du, eta corpus ez orekatuaren gainean BPE segmentazioa ikasita, hiztegi komunean bi hizkuntzak neurri berean ikastea bultzatzen da horrela. Guztira 50.000 segmentuko hiztegia ikasi da. BPE segmentazioa hori zazpigarren esperimentuari aplikatuz lortutako emaitzak 15. taulan daude.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp7	64,41	1,61	50,92	46,46	48,25	21.68	32,47	33,59
esp8 (+BPE)	71,89	2,76	56,69	50,75	53,39	26,55	36,11	37,61

Taula 15: BPE segmentazioarekin lortutako emaitzak.

Lehenengo esperimentuan euskararako lortutako emaitzetatik 35 puntuko igoera lortu da datu gehikuntzako teknikak eta BPE segmentazioa aplikatuta. Hala ere, ikasitako eredua korreferentzia-kateak zuzen lotzeko gai ez dela ikusita, ez da sekuentzien esaldi kopuru handituta probarik egin. Esaldi bikoteen barnean, korreferentzia-ebazpena egiten ez denean, esaldiz kanpoko loturak lotzerik ez baitago.

5 Emaitzak

5.1 Metrikak

Korreferentzia-ebazpenean, sistema automatikoak itzulitako erantzunak urre-patroiarekin konparatzen dira ebaluatzeko. Korreferentzia-ebazpenerako sistemen kalitatea neurtzeko, eta sistema ezberdinen arteko konparaketa egin ahal izateko ebaluazio metrika ezberdinak erabili dira azken hamarkadetan. Proposatutako metrika berriek, aurretik proposatutakoen gabeziak konpontzeko helburua dute.

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

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

Neurona-sareak itzulitako sekuentziak dokumentu mailan elkartu ondoren, KEeko metrika nagusi horiek biltzen dituen erreferentziazko tresna (Pradhan et al., 2014) erabili da sistema ebaluatzeko.

5.2 Garapenean lortutako emaitzak

Korreferentzia-ebazpenerako eraikitako Transformerra euskarazko eta ingelesezko corpusekin entrenatu da lehen bi esperimentuetan (esp0 eta esp1). Ingeleserako PreCo corpusa (12,5M hitz) erabili da ikasketarako, euskararako, berriz, tamaina txikiagoko EPEC-KORREF corpusa (27K hitz).

Bi hizkuntzen kasuan emaitzak garapeneko atalean kalkulatu dira, 16. taulan daude ikusgai. Ingelesez, lortutako emaitzak nahiko onak dira, Transformer ereduak aipamenak detektatu eta korreferentzia-ebazpena kasu gehienetan zuzen egiten du. Korreferentzia-ebazpena egiteko gai da sistema, baina oraindik, goi-sabaitzat hartu den N=2rekin lortu zitekeen emaitza hoberenetik 13 puntura dago CoNLL metrikan. Gainera, distantzia luzeko korreferentzia-kateak lotu gabe geratu dira, N=2 finkatu delako eta ez delako sekuentzien esaldi kopurua handitu. Beraz, ingeleserako baliabide askorekin hurbilpen hori noraino irits daitekeen ikertu eta artearen egoerako gainontzeko lanekin konparatuz non dabilen ikusteko beharra ikusten da.

Euskaraz, aldiz, oso emaitza txarrak lortu dira, eta Transformer ereduak ez du ia ezer ikasten. Sistema ez da gai aipamen-detekzioa ere egiteko, eta metrika bakar batean ere, ez da lau puntura iristen.

Esperimentu horietatik ateratako ondorio garbiena da, euskararako daukagun corpusa baino askoz ere handiagoa behar dela hurbilpen hau KEean erabiltzeko. Euskararako KEerako horrelako corpus handi bat etiketatzeak, ordea, kostu handia izango luke. Horre-

Sistema	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp 0 (EN)	87,52	45,13	66,50	59,04	65,61	44,17	51,52	59,08
esp 1 (EU)	3,64	0	2,75	3,59	3,92	0,04	2,11	2,23

Taula 16: KEerako ingeleserako (esp0) eta euskararako (esp1) lortutako emaitzak.

gatik, lan honetan datu-gehikuntzarako hainbat ideia proposatu dira, batzuk etorkizuneko lanerako utziz.

Datu-gehikuntzaren atal horretan, lehenik EPEC-KORREFeko esaldiak dokumentuko gainontzeko guztiekin elkartu dira. Horrela corpus sintetiko handiagoa lortu da. Honek emaitzak hobetu dituen arren, oraindik oso baxuak dira (ikus 17. taula, esp2). Aipamen-detekzioan 3 puntutik 6 puntura igo da, eta honek metrika guztien igoera ekarri du, MUC metrikaren salbuespenarekin. MUCek 0 F1 balioa izateak, korreferentzia-erlaziorik ez dela lotu adierazten du.

Jarraian, sasi-etiketatzearen bidea jorratu da. Horretarako, automatikoki etiketatutako Elhuyar Web Corpusa (360K hitz) eta Elhuyar QTleap corpusa (10M hitz) erabili dira. EPEC-KORREF corpusa eta sasi-etiketatutako Elhuyar Web Corpusa corpusaren (SE1) gainean ikasketa egin ondoren, ausazko esaldi parekaketarekin lortutako antzeko emaitzak lortzen dira 17. taulan ikus dezakegunez (esp3).

Laugarren eta bosgarren esperimentuetan, SE1 ausazko esaldi parekaketaren (AEP) konbinatuta edo, Elhuyar QTleap corpusa (SE2) erabiliz, ikasketarako sekuentzia kopurua miloi erdira handituz, emaitza hobeak lortzen hasi gara. Sistemek aipamen-detekzioa egiten ikasi dute, 48,83 eta 52,88 puntu lortuz, eta honek eraginda, metrika gehienetan, hobekuntza handiak lortu dira, CoNLL metrikan barne: 25,92 puntuko F1 balioa (esp4) eta 27,43 puntukoa (esp5) lortu dira, hurrenez hurren. Metrika gehienetan aurreko esperimentuekin konparatuz hobekuntza nabarmena lortu den arren, MUC metrikan oso balio baxua lortzen da, eta, beraz, sare neuronala korreferentzia-erlazioak lotzeko gai ez dela esan genezake. Hori, sasi-etiketatutako eta artifizialki handitutako corpusak aipamen erlazio oso gutxi dituelako da. 4. eta 5. esperimentuetan erabilitako corpusak konbinatuz (SE1+AEP + SE2), milioitik gora ikasketarako sekuentziekin entrenatu da eredua 6. esperimentuan (esp6). Horrela, 17. taulan ikus dezakegunez, datu gehikuntzaren ataleko emaitzarik onenak lortu dira aurreko esperimentuetan erabilitako corpusak konbinatuz (EU+AEP + SE1+AEP + SE2). Aipamen-detekzioan 60,75 puntu lortu dira, eta CoNLL metrikan 31,35 puntu. Hala ere, aurreko esperimentuen emaitzetan bezalaxe, MUC metrikak korreferentzia-loturarik ia ez dela sortzen adierazten digu.

Euskararako korreferentzia-ebazpena helburu izanik ere, sistemak ingeleseko corpusetik helburu-sekuentzietan etiketak zuzen sortzen ikas zezakeelakoan behintzat, euskarazko sasi-etiketatutako corpusari, ingelesekoa gehitu zaio. Elearteko ikasketa gehituta, metrika guztietan hobekuntzak lortu dira (ikus 17. taula), CoNLL metrikan pare bat puntukoak. Gainera, MUC metrikan gora egin du 0,5etik 1,6ra, askorik ez izan arren, ingeleseko corpusean aipamenak korreferentzia-kate ugari daudenez, euskarazkoan ere aipamenen arteko korreferentzia-lotura gehiago sortzen dira.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp1 (EU)	3,64	0	2,75	3,59	3,92	0,04	2,11	2,23
esp2 (+AEP)	6,27	0	4,56	6,00	5,90	0,18	3,09	3,49
esp3 (SE1)	6,54	0	5,07	6,41	6,54	0,24	3,72	3,87
esp4 (SE1+AEP)	48,83	0	38,74	37,82	39,03	12,36	25,45	25,92
esp5 (SE2)	52,88	0,36	41,60	39,38	40,32	14,55	26,49	27,43
esp6 (SE)	60,75	0,53	47,84	44,02	45,70	18,83	30,41	31,35
esp7 (SE+EN)	64,41	1,61	50,92	46,46	48,25	21.68	32,47	33,59

Taula 17: Datu gehikuntzako esperimentuekin lortutako emaitzak.

Azkenik, datu gehikuntzan sasi-etiketatutako corpusa erabiltzeak, eta batez ere ikasketarako bi hizkuntza erabiltzeak sarrerako hiztegia asko handitu du. Horri aurre egiteko, hizkuntzaren prozesamenduko hainbat atazatan erabiltzen den BPE segmentazioa erabili da, testuko hitzen segmentazioa eginez, segmentu-hiztegi mugatu batekin testuko hiztegi osoa barne hartzeko. BPE teknikak gainera oraindik eta onuragarriagoak dira euskara bezalako hizkuntza eranskarietan, lemak eta morfemak hobeto tratatzen laguntzen duelako. Datu gehikuntzako konbinazio onenari (esp7) BPE segmentazioa gehituta, 18. taulan ikusgai dauden emaitzak lortu dira. Metrika guztietan hobekuntzak lortu dira, aipamen-detekzioan 7 puntuko igoera egon da, MUC metrikan puntu batekoa, eta gainontzekoetan 4-6 puntukoa. Horrela aipamen-lotura gehiago sortzen diren arren, oraindik ere MUC metrika ez da 3 puntuko F1 baliora iristen, eta hori, aurrez datu gehikuntzan aipatu bezala, datu gehikuntzarako erabilitako teknikek, helburu-sekuentzian korreferentzia-kate gutxiko sekuentziak sortu dituztelako, eta bertatik ikasi dugulako.

esperimentua	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
esp7	64,41	1,61	50,92	46,46	48,25	21.68	32,47	33,59
esp8 (+BPE)	71,89	2,76	56,69	50,75	53,39	26,55	36,11	37,61

Taula 18: BPE segmentazioa eginda lortutako emaitzak.

5.3 Azken sistemaren ebaluazioa

Euskararako egindako esperimentu guztietan EPEC-KORREF corpusaren garapen atala erabili da emaitzak lortu eta hurbilpen ezberdinak konparatzeko. Gainera, ikasketa noiz gelditu erabakitzeko (*early stopping*) ere garapenen atal hori erabili da. Esperimentu guztien ondoren, ondoen dabilen corpus eta tekniken konbinazioa aukeratu da (esp 8), EPEC-KORREF corpuseko ebaluazio atalarekin probatzeko. Sistema horri tex2kor izena jarri diogu eta horrekin lortutako emaitzak 19. taulan ikus daiteke. Garapenean eta ebaluazioan lortutako emaitzak antzekoak dira; aipamen-detekzioan 2 puntu gutxiago lortzen

ditugun arren, CoNLL metrikan puntu erdiko aldea dago soilik. Beraz, lan honetan euskararako korreferentzia-ebazpenerako sekuentziatik sekuentziarako tex2kor hurbilpenean CoNLL metrikan 37,14 puntuko F1 balioa lortu da.

azpi-corpusa	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
garapena (esp8)	71,89	2,76	56,69	50,75	53,39	26,55	36,11	37,61
ebaluazioa (tex2kor)	69,53	2,67	55,60	50,59	53,16	24,33	37,07	37,14

Taula 19: Tex2kor sistemarekin lortutako emaitzak EPEC-KORREF corpuseko garapen eta ebaluazio azpi-ataletan.

Jarraian, inferentzian Transformer ereduak itzulitako sekuentzia bat dugu ikusgai, tex2kor sistemak KEaren atazan zer nolako emaitzak lortzen dituen metriketatik at aztertzeko.

```
jatorria   =  BAI Marokoko Gobernuak bai Fronte Pol@@ is@@ arioak
helburua   =   _     (1        1)     _    (2    _     _     2)
iragarpena =   _     (1        1)     _    (2    _     _     2)

Mendebaldeko Saharako gatazka autodetermin@@ azioari buruzko erreferendu@@ m batekin
    (3         _         3)       (4|(5        5)       _          _       _    4)
     _         _        (3          _          3)       _          _       _   (4)

konpontzea hitzartu zuten hamarkada hasieran .  Nazio Batuen Erakundea izan zen
     _        _       _      (6        6)    _   (7     _        7)      _   _
     _        _       _      (5        5)    _ (6|(7)   _        6)      _   _

bitartekari eta hark hartu zuen prozesua zaintzeko ardura .
    (7)      _  (7)    _    _    (8|(9)     _         8)  _
    (8       _   8)    _    _     (9)       _         _   _
```

Adibidea aztertuta, tex2kor sistemak lehen lerroan korreferentzia-ebazpena zuzen egiten duela ikus dezakegu. Bigarrengoan aldiz, aipamenak egoki detektatzeko ere ez da gai izan, aipamenen mugak non daudenaren nozioa baduen arren. Hirugarren lerroan behar baino aipamen bat gehiago (*Nazio*) sortzen du, baina gainontzekoa zuzen egiten du. Laugarren lerroan berriz, aipamen bat zuzen hartzen du, beste bi elkarrekin elkartzen ditu, eta azkena berriz ez du detektatzen. Horretaz gain, hirugarren eta laugarren lerroen arteko korreferentzia-erlazioa (*Nazio Batuen Erakundea*, *bitartekari* eta *hark*) ez du lotzen.

Emaitzen azterketa sakonagoa behar izan arren ondorio zuzenak ateratzeko, esan dezakegu, eraiki den tex2kor sistemak, korreferentzia-ebazpena, erdizka bada ere, egiteko gai dela. Aipamen-detekzioan nahiko ondo dabil, aurreko sistemetatik ez oso urrun. Korreferentzia-erlazioak lotzeari dagokionez, ordea, irteerako adibideek eta MUC metrikak adierazten duten moduan, sistema ez da gai korreferentzia-kateak zuzen sortzeko.

5.4 Emaitzen konparaketa

Transformer arkitektura erabilita sekuentziatik sekuentziarako eraiki den sistemak (tex2kor), artearen egoerako gainontzeko sistemekin konparatuz emaitza eskasak lortu ditu (ikus 20. taula). Erregelatan (Soraluze et al., 2019), ikasketa automatikoan (Soraluze et al., 2017a) eta ikasketa sakonean (Urbizu et al., 2019b) oinarritutako euskararako KEerako orain arteko sistemek, erregelatan oinarritutako aipamen-detektatzailea darabilte (F1 73,79). Lan honetan, berriz, aipamen-detekzioa burutzen ere ikasi da; aurreko sistemarekin konparatuz erregelatako aipamen-detektatzaileak baino 4 puntu gutxiago lortzen ditu (F1 69,53). Korreferentzia-ebazpenari dagokionez, CoNLL metrikan, aurreko hiru sistemek baino emaitza baxuagoak lortu dira (F1 37,14), sistema neuronalarekin konparatuz, 4 puntu gutxiago lortu dira (LEA metrikan 2 puntuko hobekuntza lortu den arren) eta erregelatan oinarritutako sistemarekin konparatuz, tex2kor sistemak 19 puntu gutxiago lortu ditu.

Sistema	AD	MUC	B^3	CEAF_m	CEAF_e	BLANC	LEA	CoNLL
Soraluze et al. (2019)[1]	73,79	42,95	62,97	61,56	62,04	43,48	49,26	55,98
Soraluze et al. (2017a)[2]	73,79	40,98	61,66	59,82	60,00	43,48	45,97	54,21
Urbizu et al. (2019b)[3]	73,79	9,32	58,40	53,28	55,87	29,41	35,79	41,20
tex2kor	69,53	2,67	55,60	50,59	53,16	24,33	37,07	37,14

[1]erregelatan oinarritua [2]ikasketa automatikoan oinarritua [3]ikasketa sakonean oinarritua

Taula 20: Euskararako korreferentzia-ebazpenerako sistemen konparaketa. EPEC-KORREF corpusean lortutako CoNLL metrikaren F1 balioak.

Lan honetan aurkeztutako euskararako KEerako sekuentziatik sekuentziarako hurbilpenarekin, tex2kor sistemarekin, espero baino emaitza baxuagoak lortu dira. Hasiera batean, EPEC-KORREF corpusean soilik ikasiz emaitza txarrak lortzen zirelako, datu gehikuntzako hainbat teknika erabili dira.

Sortutako corpus artifizialetatik testutik korreferentziarako eredua ikasita, 2,23tik 37,61 punturako igoera lortu da CoNLL F1 metrikan. Aplikatutako datu gehikuntzarako teknika bakoitzarekin eta BPE segmentazioarekin hobekuntza nabarmenak lortu dira. Aipamen-detekzioan erregelatan oinarritutako sistematik gertu geratu da sistema, 4 puntura, baina korreferentzia-kateak egoki lotzeko ez da gai tex2kor sistema, eta aipamenak *singleton* bezala uzten ditu, MUC metrikan lortutako 2,67 puntuko F1 balioak adierazten duen bezala. Sasi-etiketaren bitartez sortutako corpus artifizialak berak korreferentzia-kate oso gutxi dituela etiketatuta izan daiteke emaitza horien arrazoia.

6 Ondorioak eta etorkizuneko lana

6.1 Ondorioak

Master amaierako lan honetan, korreferentzia-ebazpenaren ataza lantzeko hurbilpen neuronal berri bat aurkeztu da. Orain arteko korreferentzia-ebazpenerako hurbilpenak multzokatze algoritmo ezberdinetatik eratorriak dira nagusiki, lan honetan, berriz, lehen aldiz, korreferentzia-ebazpena sekuentziatik sekuentziarako ataza modura lantzea planteatzen da. Horretarako hizkuntzaren prozesamenduko sekuentziatik sekuentziarako hainbat atazatan arrakastaz erabili den Transformer arkitektura erabili da.

Transformer arkitektura erabiltzeak dokumentu osoak sekuentzia bakar modura lantzeko memoria arazoak dakartzanez, korreferentzia-ebazpenerako dokumentuak sekuentziatan zatitzeko eta inferentziaren ostean berriro elkartzeko algoritmo bat aurkeztu da. Algoritmo horretan, dokumentua zer luzeratako sekuentzietan banatzeak emaitzetan duen eragina aztertu da.

Arkitekturaren bideragarritasuna aztertu asmoz, korreferentzia-ebazpenerako diseinatutako sekuentziatik sekuentziarako hurbilpen hori, ingeleserako probatu da. PreCo corpusean lortutako emaitzak nahiko onak dira, CoNLL metrikan 59,08 puntuko F1 balioa lortu da, eta eredua gai dela korreferentzia-ebazpena egiteko ondorioztatzera eraman gaitu. Tratatu diren sekuentzien luzera (2 esaldi) handitu ahala eta BPE segmentazioa aplikatuz edota aurre-ikasitako beste ereduren batekin konbinatuz, emaitza hori hobetzeko tartea dagoela aurreikusten da.

Hurbilpena bera euskararako EPEC-KORREF corpusean soilik ikasketa egiteko erabilita, berriz, oso emaitza txarrak lortu dira, CoNLL metrikan 2,23 puntuko F1 balioa. Entrenatutako Transformer eredua ez da gai ataza ikasteko. Lortutako emaitzak corpusaren tamaina txikiak eragindakoak direla ondorioztatu da, bi hizkuntzen arteko ezberdintasunak tamaina berdineko corpusarekin entrenatu balira, ez luke-eta handiegia izan behar.

Lan honek Euskararako korreferentzia-ebazpena duenez helburu, euskararako ikasketarako corpusa handitzeko datu gehikuntzako hainbat teknika aplikatu dira. Ikasketarako jatorri-helburu sekuentzia bikote gehiago izateko, ausazko esaldi parekaketa, sasi-etiketatutako corpusak, eta hizkuntza arteko ikasketa landu dira. Ondoren, landutako hurbilpen bakoitzak emaitzen hobekuntza nabarmena ekarri du; datu gehikuntzako hiru teknikak konbinatuz, emaitzak hobetzea lortu da, CoNLL metrikan 33,59 puntuko F1 balioa lortuz. Transformer ereduak aipamen-detekzioa egiten ikasi du, (64,41 F1), baina MUC metrikan lortutako balio baxuak sistemak korreferentzia-kateak sortzeko arazoak daudela adierazten du.

Datu gehikuntzako konbinazio onena ematen duen sistemari BPE hitzen segmentaziorako teknika aplikatuz, emaitzak hobetzea lortu da, aipamen-detekzioan 71,89 puntu eta CoNLL metrikan 37,61 puntuko F1 balioa lortuz. Sistemak aipamenen arteko korreferentzia-kateak behar bezala sortu gabe jarraitzen duen arren, aipamen-detekzioan emaitza onak lortzen ditu.

Garapenean ikasketarako sasi-etiketatutako eta ingelesezko corpusak gehitzeak eta BPE segmentazioa aplikatzeak, 2 puntutik 37 punturainoko hobekuntza ekarri dute CoNLL

metrikan.

Korreferentzia-ebazpenerako sekuentziatik sekuentziarako Transformerra darabilen euskararako tex2kor sistemak, EPEC-KORREFen ebaluazio ataleko corpusean CoNLL metrikan 37,14 puntu lortzen ditu, garapenerako atalarekin lortutakoaren antzekoak. Lortutako emaitzak aurreko sistemekin lortutakoak baino nabarmen baxuagoak dira. Tex2kor sistemak aipamen-detekzioa nahiko ondo egin arren, MUC metrikak adierazten duen moduan, sistema ez da gai korreferentzia-kateak zuzen lotzeko, eta aipamen gehienak *singleton* moduan uzten ditu. Hori, Transformer ereduaren ikasketan erabilitako corpusak etiketatuta korreferentzia-erlazio gutxi dituelako gertatu dela ondorioztatu da, datu gehikuntzarako erabilitako corpusen kalitateak eraginda.

6.2 Etorkizuneko lana

Etorkizunerako lanen artean, lehenik eta behin, korreferentzia-ebazpenerako hurbilpena ingelesa bezalako baliabide askoko hizkuntza batetan sakonki ikertu nahi da. Horretarako ARRAU (Uryupina et al., 2020) eta CoNLL-2012 corpusak erabiltzeko asmoa dago, lortutako emaitzak artearen egoerako gainontzeko sistemekin konparatu ahal izateko.

Horretaz gain, hemen planteatzen den sekuentziatik sekuentziarako hurbilpena, aurre-ikasketa ez-gainbegiratuko ereduekin konbinatu nahi da (BERT (Devlin et al., 2019), ...), eredu hauek korreferentzia-ebazpena modu ez-gainbegiratuan ikasteko gai direla jakina da eta (Clark et al., 2019; Tenney et al., 2019).

Euskarari dagokionez, lortutako emaitzetan, aipamen-detekzioaren azpi-atazan nahiko emaitza onak lortu dira, sistemarik onenetik 4 puntutara geratuz. Hemen aurkeztutako hurbilpen bera korreferentzia-ebazpenerako beharrean aipamen-detekziorako landuz emaitzak hobetu litezkeen aztertu nahi da. Sasi-etiketatutako corpusaren aipamen-detekzioaren etiketatzea kalitate onekoa da; korreferentzia-etiketetatik klusterrei dagokien zenbakiak kenduta eta parentesiak soilik utziz, helburu-hiztegi txikiagoa izango genuke, eta honek aipamen-detekzioa ikastea erraztuko luke.

Euskararako korreferentzia-ebazpenari dagokionez, sasi-etiketatzearen bidetik, beste saiakera bat egin nahi da, erregelatan oinarritutako euskararako korreferentzia-ebazpenerako sistemarekin (Soraluze et al., 2019). Lan honetan tresna horrekin sasi-etiketatu den corpusa baino corpus handiago bat eraiki nahi da. Testu garbiagoa, domeinu berekoa eta dokumentuka antolatua dagoena oinarri hartuz, kalitate altuagoz sasi-etiketatutako corpus batetatik korreferentzia-ebazpena ikastea errazagoa izango delakoan gaude. Bestalde, ereduak aipamen-loturak sortzera bultzatzeko, esaldiak norbere buruarekin elkartuta (tartean ausazko esaldiak sartuz, agian, zarata sartzearren) sortutako sekuentziak erabiltzea aztertu nahi da. Teknika sinplea izan arren, ikasketarako sekuentzietan aipamen-lotura gehiago egotea ekarriko luke.

Bestalde, ingeleserako bezala, euskararako ere aurre-ikasitako ereduekin konbinatzeak (izan Transformerraren kodetzailean txertatuta, edota hitz-bektore modura) emaitzak hobetuko lituzkeela espero da. Gainera, aurre-ikasitako ereduak jatorri-sekuentziak kodetzeko erabiltzeak, elearteko eredu bat ikasteko aukera emango luke, ingelesaz gain gaztelania, frantsesa edo beste edozein hizkuntzetatik ikasiz, deskodetzaileak ikasi beharreko ataza

komuna delako hizkuntza guztietan. Elearteko hurbilpen honek euskara bezalako baliabide gutxiko hizkuntzetan korreferentzia-ebazpena sare neuronalen bitartez lantzeko bidea emango lukeela uste da.

Aurre-ikasitako eredu ez-gainbegiratu hauek ingeleserako korreferentzia-ebazpenerako sistemetan txertatuak daude dagoeneko, eta lan honetan aurkeztutako hurbilpena ere, BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2019a), RoBERTa (Liu et al., 2019), aurre-ikasitako transformer eleanitza (Liu et al., 2020), eta antzekoekin elkartzeko aproposa dela uste da, guzti horiek, auto-kodetzaile, edo kodetzaile-deskodetzaile arkitekturak dituztelako.

Horretaz gain, hurbilpen honek izan dezakeen arazo nagusiari, sekuentzien luzera mugatu beharrari, aurre egiteko, *sparce-attention* (Child et al., 2019), transformer-XL (Dai et al., 2019) eta Reformer (Kitaev et al., 2020) bezalako arkitekturak aztertzeko asmoa dago. Horiekin, sekuentzien luzera maximoak luzatu eta dokumentuak zatitzea saihestu edo horren eragina murriztuko genuke, zatitze-elkartze algoritmoak dakarren galera txikituz.

Erreferentziak

Itziar Aduriz, Maxux J Aranzabe, Jose Maria Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, eta Larraitz Uria. A cascaded syntactic analyser for basque. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 124–134. Springer, 2004.

Itziar Aduriz, Maxux Aranzabe, J Arriola, Aitziber Atutxa, Arantza Diaz-De-Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, eta Ruben Urizar. Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing, 2006.

Samarth Agrawal, Aditya Joshi, Joe Cheri Ross, Pushpak Bhattacharyya, eta Harshawardhan M Wabgaonkar. Are word embedding and dialogue act class-based features useful for coreference resolution in dialogue. In *Proceedings of PACLING*, 2017.

Liesbeth Allein, Artuur Leeuwenberg, eta Marie-Francine Moens. Dutch anaphora resolution: A neural network approach towards automatic die/dat prediction. In *The 30th Meeting of Computational Linguistics in The Netherlands (CLIN 30), Location: Utrecht, The Netherlands*, 2020.

Vinay Annam, Nikhil Koditala, eta Radhika Mamidi. Anaphora resolution in dialogue systems for south asian languages. *arXiv preprint arXiv:1911.09994*, 2019.

Mikel Artetxe, Gorka Labaka, eta Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

Mikel Artetxe, Gorka Labaka, eta Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, 2018.

Amit Bagga eta Breck Baldwin. 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, 1998.

Dzmitry Bahdanau, Kyunghyun Cho, eta Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Piotr Bojanowski, Edouard Grave, Armand Joulin, eta Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, eta Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Samuel Broscheit, Simone Paolo Ponzetto, Yannick Versley, eta Massimo Poesio. Extending bart to provide a coreference resolution system for german. In *LREC*. Citeseer, 2010.

David Maclean Carter. *A shallow processing approach to anaphor resolution*. PhD thesis, University of Cambridge, 1986.

Klara Ceberio, Itziar Aduriz, Arantza Díaz de Ilarraza, eta Ines Garcia-Azkoaga. Coreferential relations in Basque: the annotation process. *Journal of psycholinguistic research*, 47(2):325–342, 2018.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, eta Shu Rong. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, 2018.

Rewon Child, Scott Gray, Alec Radford, eta Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Kevin Clark eta Christopher D Manning. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, 2016a.

Kevin Clark eta Christopher D Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 643–653, 2016b.

Kevin Clark, Urvashi Khandelwal, Omer Levy, eta Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019.

Dennis Connolly, John D Burger, eta David S Day. A machine learning approach to anaphoric reference. In *New methods in language processing*, pages 133–144, 1997.

André Ferreira Cruz, Gil Rocha, eta Henrique Lopes Cardoso. Exploring Spanish corpora for Portuguese coreference resolution. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 290–295. IEEE, 2018.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, eta Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Pascal Denis eta Jason Baldridge. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, eta Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Hongliang Fei, Xu Li, Dingcheng Li, eta Ping Li. End-to-end deep reinforcement learning based coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 660–665, 2019.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, eta James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, eta YannÑ Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org, 2017.

Loïc Grobol. Neural coreference resolution with limited lexical context and explicit mention detection for oral french. In *Second Workshop on Computational Models of Reference, Anaphora and Coreference*, page 8, 2019.

Lynette Hirschman eta Nancy Chinchor. Appendix f: Muc-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

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

Samira Hourali, Morteza Zahedi, eta Mansour Fateh. Coreference resolution using neural mcdm and fuzzy weighting technique. *International Journal of Computational Intelligence Systems*, 2020.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, eta Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019a.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, eta Daniel S Weld. Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5807–5812, 2019b.

Ben Kantor eta Amir Globerson. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, 2019.

Nikita Kitaev, Lukasz Kaiser, eta Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkgNKkHtvB`.

Mateusz Kopeć. Three-step coreference-based summarizer for polish news texts. *Poznan Studies in Contemporary Linguistics*, 55(2):397–443, 2019.

Mateusz Kopec eta Maciej Ogrodniczuk. Creating a coreference resolution system for polish. In *LREC*, pages 192–195, 2012.

M Hari Krishna, K Rahamathulla, eta Ali Akbar. A feature based approach for sentiment analysis using svm and coreference resolution. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 397–399. IEEE, 2017.

Gourab Kundu, Avi Sil, Radu Florian, eta Wael Hamza. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 395–400, 2018.

Guillaume Lample eta Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Shalom Lappin eta Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, eta Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics, 2011.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, eta Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013.

Kenton Lee, Luheng He, Mike Lewis, eta Luke Zettlemoyer. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, 2017.

Kenton Lee, Luheng He, eta Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, 2018.

Hector Levesque, Ernest Davis, eta Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, eta Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, eta Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

Xiaoqiang Luo. 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, 2005.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, eta Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 135. Association for Computational Linguistics, 2004.

Minh-Thang Luong, Hieu Pham, eta Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

Ruslan Mitkov. *Anaphora resolution: the state of the art*. Citeseer, 1999.

Nafise Sadat Moosavi eta Michael Strube. 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, 2016.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411. Association for Computational Linguistics, 2010.

Vincent Ng. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Bartłomiej Nitoń, Paweł Morawiecki, eta Maciej Ogrodniczuk. Deep neural networks for coreference resolution for Polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 395–400, 2018.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, eta Manabu Okumura. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, 2019.

Arantxa Otegi, Nerea Ezeiza, Iakes Goenaga, eta Gorka Labaka. A modular chain of nlp tools for basque. In *International Conference on Text, Speech, and Dialogue*, pages 93–100. Springer, 2016.

Cheoneum Park, KyoungHo Choi, Changki Lee, eta Soojong Lim. Korean Coreference Resolution with Guided Mention Pair Model Using Deep Learning. *ETRI Journal*, 38 (6):1207–1217, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, eta Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, eta R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, eta Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Massimo Poesio, Olga Uryupina, eta Yannick Versley. Creating a coreference resolution system for italian. In *LREC*, 2010.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, eta Nianwen Xue. 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, 2011.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, eta Michael Strube. 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, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-2006`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, eta Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Altaf Rahman eta Vincent Ng. 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, 2009.

Marta Recasens eta Eduard Hovy. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, 2011.

Emili Sapena, Lluís Padró, eta Jordi Turmo. 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, 2011.

A Sboev, R Rybka, eta A Gryaznov. Deep neural networks ensemble with word vector representation models to resolve coreference resolution in russian. In *Advanced Technologies in Robotics and Intelligent Systems*, pages 35–44. Springer, 2020.

Rico Sennrich, Barry Haddow, eta Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 2016a.

Rico Sennrich, Barry Haddow, eta Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016b.

Tomohide Shibata eta Sadao Kurohashi. Entity-centric joint modeling of japanese coreference resolution and predicate argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 579–589, 2018.

Utpal Kumar Sikdar, Asif Ekbal, eta Sriparna Saha. A generalized framework for anaphora resolution in indian languages. *Knowledge-Based Systems*, 109:147–159, 2016.

Sonit Singh. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*, 2018.

Wee Meng Soon, Hwee Tou Ng, eta Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.

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

Ander Soraluze, Olatz Arregi, Xabier Arregi, eta Arantza Diaz de Ilarraza. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. In *Procesamiento del Lenguaje Natural*, volume 55, pages 23–30, 2015.

Ander Soraluze, Olatz Arregi, Xabier Arregi, Arantza Diaz de Ilarraza, Mijail Kabadjov, eta Massimo Poesio. Coreference Resolution for the Basque Language with BART. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 67–73, 2016.

Ander Soraluze, Olatz Arregi, Xabier Arregi, eta Arantza Díaz de Ilarraza. 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, 2017a.

Ander Soraluze, Olatz Arregi, Xabier Arregi, eta Arantza Díaz de Ilarraza. Improving mention detection for Basque based on a deep error analysis. *Natural Language Engineering*, 23(3):351–384, 2017b.

Ander Soraluze, Olatz Arregi, Xabier Arregi, eta Arantza Diaz de Ilarraza. Euskor: End-to-end coreference resolution system for basque. *PloS one*, 14(9), 2019.

Josef Steinberger, Mijail Kabadjov, eta Massimo Poesio. Coreference applications to summarization. In *Anaphora Resolution*, pages 433–456. Springer, 2016.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, eta Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 2020.

I Sutskever, O Vinyals, eta QV Le. Sequence to sequence learning with neural networks. *Advances in NIPS*, 2014.

Ian Tenney, Dipanjan Das, eta Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

Gorka Urbizu, Ander Soraluze, eta Olatz Arregi. Deep cross-lingual coreference resolution for less-resourced languages: The case of basque. In *Proceedings of the 2nd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2019), co-located with NAACL 2019*, 2019a.

Gorka Urbizu, Ander Soraluze, eta Olatz Arregi. Neurona-sareetan oinarritutako euskararako korreferentzia-ebazpena. In *III. Ikergazte: Nazioarteko ikerketa euskaraz. pp. 141-147, Baiona. ISBN 978-84-8438-686-5*, 2019b.

Olga Uryupina, Alessandro Moschitti, eta Massimo Poesio. Bart goes multilingual: the unitn/essex submission to the conll-2012 shared task. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 122–128. Association for Computational Linguistics, 2012.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, eta Massimo Poesio. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1): 95–128, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ Gomez, Łukasz Kaiser, eta Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, eta Alessandro Moschitti. 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, 2008.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, eta Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52. Association for Computational Linguistics, 1995.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49, 2018.

Lesly Miculicich Werlen eta Andrei Popescu-Belis. Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, 2017.

Sam Wiseman, Alexander M Rush, Stuart Shieber, eta Jason Weston. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1416–1426, 2015.

Sam Wiseman, Alexander M Rush, eta Stuart M Shieber. Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, 2016.

Xiaofeng Yang, Guodong Zhou, Jian Su, eta Chew Lim Tan. 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, 2003.

Xiaofeng Yangy, Jian Su, Guodong Zhou, eta Chew Lim Tan. An np-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics*, page 226. Association for Computational Linguistics, 2004.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, eta Hai Zhao. Lingke: a fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 108–112, 2018.

Eranskinak

A Zatitze algoritmoaren adibidea corpusean

```
begin document egun.06-2-p3903.2000-06-02.kirola

Frantziako ibilbide gogorretan gogotsu ariko da Euskaltel .
(121 _ 121) _ _ _ (129) _
Alpeetako Klasikoan eta Dauphine Liberen ariko dira hurrena .
(86|(146 146) _ (104 104)|86) _ _ _ _
EUSKALTEL ez da Tourrean izango , baina horrek ez die moralik kendu .
(129) _ _ (133) _ _ _ (31) _ _ (62) _ _
Datozen asteotan , Alpeetan , Dauphinen eta Kataluinian parte hartuko dute .
(88 88) _ (3|(146) _ (104) _ (84)|3) _ _ _ _
Maila handiko 3 proba hauetan , Tourrean lekua badutela erakusten saiatuko dira .
(3 _ _ _ 3) _ (133) (80) _ _ _ _ _
Euskal Bizikletako erakustaldiaren ondoren , Euskaltelek emaitza onak lortu ditzake .
(53|(13 13) 53) _ _ (129) (28 28) _ _ _
Haimar Zubeldia , Mikel Artetxe eta Mikel Pradera sasoi betean daude .
(14|85 85) _ (50 50) _ (48 48)|14) (140 140) _ _
.
.
.
```

Frantziako ibilbide gogorretan gogotsu ariko da Euskaltel . Alpeetako Klasikoan eta Dauphine Liberen ariko dira hurrena .
(1 _ 1) _ _ _ (2) _ (3|(4 4) _ (5 5)|3) _ _ _ _
Alpeetako Klasikoan eta Dauphine Liberen ariko dira hurrena . EUSKALTEL ez da Tourrean izango , baina horrek ez die moralik kendu .
(1|(2 2) _ (3 3)|1) _ _ _ _ (4) _ _ (5) _ _ _ (6) _ _ (7) _ _
EUSKALTEL ez da Tourrean izango , baina horrek ez die moralik kendu . Datozen asteotan , Alpeetan , Dauphinen eta Kataluinian parte hartuko dute .
(1) _ _ (2) _ _ _ (3) _ _ (4) _ _ (5 5) _ (6|(7) _ (8) _ (9)|6) _ _ _ _
Datozen asteotan , Alpeetan , Dauphinen eta Kataluinian parte hartuko dute . Maila handiko 3 proba hauetan , Tourrean lekua badutela erakusten saiatuko dira .
(1 1) _ (2|(3) _ (4) _ (5)|2) _ _ _ _ (2 _ _ _ 2) _ (6) (7) _ _ _ _ _
Maila handiko 3 proba hauetan , Tourrean lekua badutela erakusten saiatuko dira . Euskal Bizikletako erakustaldiaren ondoren , Euskaltelek emaitza onak lortu ditzake.
(1 _ _ _ 1) _ (2) (3) _ _ _ _ _ (4|(5 5) 4) _ _ (6) (7 7) _ _ _
Euskal Bizikletako erakustaldiaren ondoren , Euskaltelek emaitza onak lortu ditzake . Haimar Zubeldia , Mikel Artetxe eta Mikel Pradera sasoi betean daude .
(4|(5 5) 4) _ _ (6) (7 7) _ _ _ (8|9 9) _ (10 10) _ (11 11)|8) (12 12) _ _