Konputazio Ingeniaritza eta
Sistema Adimentsuak Unibertsitate Masterra

Máster Universitario en Ingeniería Computacional
y Sistemas Inteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila

Departamento de Ciencias de la Computación e Inteligencia Artificial

Master Tesia

Tesis de Máster

Diskurtsoko koherentzia erlazioen iragarpenak euskararako BERT sare neuronal aurreentrenatua erabiliz

Egilea / Autor
Erik Angulo Arnaiz

Zuzendaritza
Dirección

Ander Soraluze Irureta

HiTZ Basque Center for Language Technologies - Ixa NLP Group,
University of the Basque Country UPV/EHU

Mikel Iruskieta Quintian

HiTZ Basque Center for Language Technologies - Ixa NLP Group,
University of the Basque Country UPV/EHU

Universidad del País Vasco Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA
Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

Laburpena

Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] teoriak testuak deskribatzeaz arduratzen da, koherentzia azalduz. Deskribapena zuhaitz eran antolatuta adierazten da, testua segmentuetan banatuta egonik. Segmentuak erlazioen bidez lotzen dira, eta erlazio bakoitzak segmentuen arteko efektu bat adierazten du. Proiektu honetan, erlazio hauek iragarriko dira ikasketa automatikoko teknikak erabiliz, zehazki euskararako BERT sare neuronal aurreentrenatua. Oinarri bezala DisCoDisCo [Gessler et al., 2021] sistema erabiliko dugu gure KEIEBA sistema sortzeko, non hobekuntza proposamenak aurkeztuko ditugun. Azkenik, emaitzak azalduko dira eta etorkizuneko lanak proposatuko dira.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Testu bat idatzita dagoen eta irakur daitekeen hitzen multzoa da. Horretarako, hitzek esanahi bat izan behar dute. Gainera, hitzek osatzen dituzten esaldiek hizkuntza baten gramatika arauak jarraitu behar dituzte. Bestalde, esaldiek beraien artean koherentzia mantendu behar dute testua ulergarria izateko. Adibidez, *"Nora zoaz?"-"Komunera"* koherentea da, baina *"Nora zoaz?"-"Sagarrak dakartzat"* ez da koherentea. Azken finean, bi esaldiek ematen dituzten ideiak esanahi osotasun bat izan behar dute eta beraien artean kontraesanhik ez sortu. Badago testuen koherentzia deskribatzen duen teoria, ***R**hetorical **S**tructure **T**heory* [Mann and Thompson, 1988].

RST teoriak testu baten segmentuen artean koherentzia deskribatzen du. Horretarako, testua segmentuetan banatzen da, eta beste segmentu ala segmentu-multzoekin elkartzen da. Multzokatze hori diskurtsoko koherentzia erlazioen bitartez gauzatzen da, segmentuen arteko koherentzia azalduz eta efektu bat gauzatuz.

Proiektu honetan RST erlazioak iragarri nahi dira, BERT sare neuronalen bitartez. Horretarako, DisCoDisCo [Gessler et al., 2021] sistema oinarri bezala hartuko dugu gure sistema eraikitzeko, KEIEBA. Proiektu honen helburua KEIEBA sisteman proposamenak egitea eta emaitzak aztertzea izango da, DisCoDisCo sistemaren erlazioen zehaztasuna hobetzeko asmoarekin.

Lehenik eta behin, 2. kapituluan aurrekariak aztertuko ditugu. Hasteko, *Rhetorical Structure Theory* deskribatuko da 2.1. atalean. Honekin jarraituz, 2.2. atalean RSTren artearen egoera aztertuko da, baita RST erlazioak iragartzeko sistemen artearen egoera ere. Ondoren, iragarpenak egiteko erabiliko diren baliabideak deskribatuko dira 2.3. atalean, horien

artean, sare neuronalak, BERT, erabiliko den corpusa eta erabiliko den DisCoDisCo sistema. Erlazioen iragarpena hobetzeko sisteman proposatutako esperimentuak 3. kapituluan aurki ditzakegu. Esperimentuekin lortutako emaitzak 4. kapituluan azter daitezke. Bukatzeko, proiektuaren ondorioak eta etorkizuneko lanak 5. kapituluan komentatuko ditugu.

Proiektu hau Ixa[1] taldearekin garatu da. Ixa hizkuntzaren prozesamenduaren arloan aritzen den ikerketa-taldea da, UPV/EHUkoa. Ikerlerroen artean, galdera-erantzun sistemak, itzulpen automatikoa, hizkuntza-ereduak, ikaskuntza-automatikoa, informazio erauzketa, morfosintaxia, semantika, eta abar aurkitzen dira. Ikerlerro bakoitzeko datu, corpus, tresna eta aplikazio berriak sortzen dituzte.

[1] Ixa taldearen webgunea: http://ixa.si.ehu.es/

2. KAPITULUA

Aurrekariak

2.1 Rhetorical Structure Theory eta koherentzia erlazioak

Idazkera asmatu zenetik gizakiok testuz inguratuta bizi gara. Batez ere, garrantzitsuak izan daitezkeen ideiak erregistratuta uzteko. Adibidez, gure historia, legeen arauketa, pertsona baten osasun erregistroa, errezetak, eta abar. Dena den, idazkeraren helburu nagusia komunikazioa da, horien artean, egunkariak, emailak, mezuak eta iragarkiak, adibidez.

Testu baten azterketa egin nahi izanez gero hizkuntzaren prozesamenduko teknikak erabili ahal dira. Aurreko adibideak aztertuta, testuak etiketak erabiliz sailka daitezkeela ikus daiteke. Adibidez, narratiboa, zientifikoa eta argudiokoa. Hau jakinik testuaren ideia orokor bat izan genezake testua irakurri baino lehen bere titulua irakurrita. Testuak sailkatzeko beste modu bat testua positibo ala negatibo bezala klasifikatzea da. Adibidez, argudioko testu batean idazlea pozik ala haserre egotearen arabera.

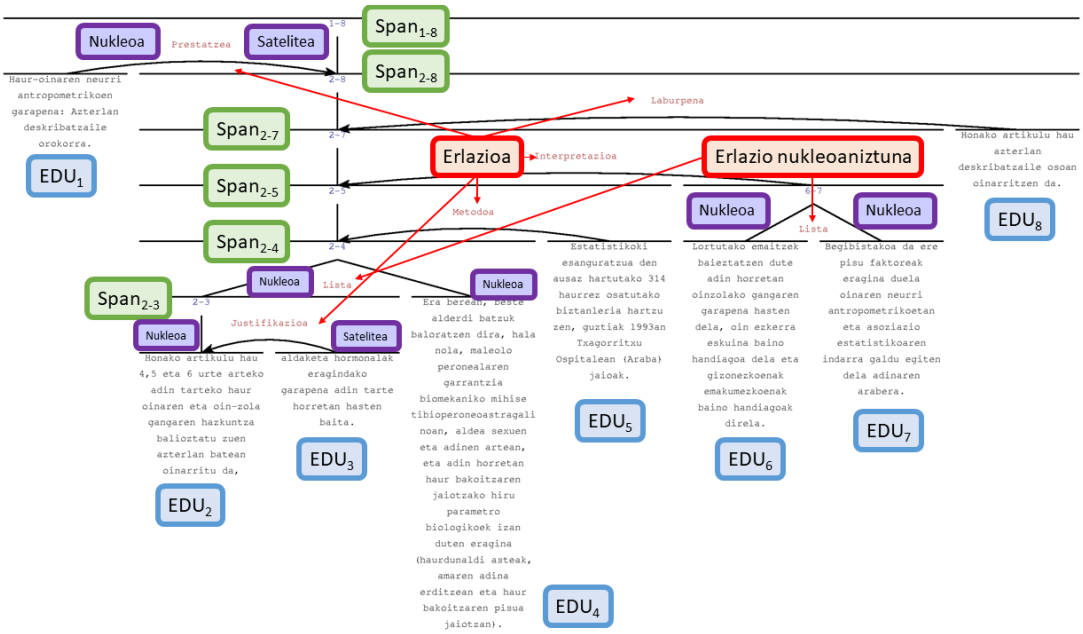
Hala ere, etiketazio metodo horiek nahiko orokorrak dira. Testuen deskribapen sakonagoak egiteko aukera badago beste hizkuntzaren prozesamenduko teknikak erabiliz. Horien artean Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] dago. Teoria honek testu batean dauden parte ezberdinen artean egon daitezkeen erlazioak deskribatzen ditu. Gainera, era hierarkikoan egiten da deskribapena, testuko zatiak erlazionatuz. Beste era batera esanda, testuaren koherentzia azaltzen du, hots, haien artean kontraesanik edo inkoherentziarik sortzen ez duten bi osagaien arteko erlazio logikoa azaltzen du.

RSTren jatorria testuak era automatikoki sortzeko helburutik dator [Marcu, 2000], baina

teoria aldatuz joan izan da jatorrizko teoria hobetzen duten hainbat moldaketekin [Taboada and Mann, 2006b]. Batez ere, deskribapenen erregelak berrituz eta finkatuz.

Testuaren deskribapena era hierarkikoan aurkezten da, zuhaitz itxurarekin. Hierarkian testua zatitan ala segmentuetan agertzen da, segmentazio prozesua jarraituz. Segmentaziotik lortzen diren osagaiek, segmentuek alegia, osotasun funtzional independente izan behar dute, [Mann and Thompson, 1988]ek proposatutako segmentu definizioaren arabera.

Segmentuak *Elemmentary Discourse Unit* (EDU) ere deitu daitezke. EDUak oinarrizko diskurtso unitateak dira, eta hauek multzokatuz *span* bat lortuko da, hau da, unitate-multzoa. Era berean, *span*ak beste EDU edo *span*ekin elkar daitezke. EDUak zenbakizko identifikadore batekin izendatzen dira, agerpen ordena jarraituz. *Span*ak izendatzeko, honek barne dituen EDUen zenbaki tartea izango du identifikadore gisa.



2.1. irudia: RST zuhaitz-diagramaren egitura.

2.1. irudian testu baten RST zuhaitz egitura ikus daiteke. Bertan, testuko segmentuak agertzen dira (urdinez) eta beraien artean lotzen dira, unitate-multzoak osatuz (berdez) goreneko mailan adierazita egonik. Adibidez, EDU_2 eta EDU_3 erlazionatzen dira goreneko mailan $span_{2-3}$ osatuz. Gero, *span* hau EDU_4arekin lotzen da $span_{2-4}$a osatuz.

Gerta daiteke lotutako EDU edo *span* batek ideia garrantzitsuena izatea. Kasu horretan unitate hori nukleoa izango da. Besteak, sateliteak, nukleoaren esanahia aldatuko du. Nukleoek garrantzi handiagoa daukate testuan sateliteak baino, ondorioz batzutan sateliteak

ematen duen informazioa ulergaitza da nukleoko informazioa izan ezean. Hau ez bada gertatzen biak izango dira nukleoak, eta lortutako unitate-multzoa nukleoaniztuna (multinuklearra) izango da. Zuhaitz-diagrametan (ikusi 2.1. irudiko adibide moreak) multzoko nukleoa eta satelitea gezi batekin adierazten dira. Nukleoa geziaren helmugarekin adierazita dago eta satelitea geziaren abiapuntuarekin. Nukleoaniztasunen kasuan, bi lerro zuzenekin adierazten dira multzoko nukleoak.

Unitateen multzokatzea koherentzia erlazioekin gauzatzen da (irudian gorriz), nukleoaniztunetan eta nukleo-satelite motetan sailkatzen direnak. Erlazio multinuklearren kasuan, biak lotzeak sortzen duen erlazioko izenaren efektua da. Erlazioa nukleo-satelite bada, sateliteak nukleoari esanahia aldatzen dionaren araberakoa izango da. Hauek bi azpikategoriatan banatzen dira: edukizkoak eta aurkezpenezkoak. Edukizkoek erlazioko izenaren efektua eragiten dute, izaera semantikoa izanik eta aurkezpenezkoek, ordea, unitateen arteko loturak irakurlearengan efektu jakina eragiten dute, izaera erretorikoa izanik.

RST teoria sortu zenetik, hasierako erlazioez gain beste erlazio-zerrenda berri batzuk proposatu dira. Batez ere, hizkuntzalariek faltan botatzen zituzten fenomenoak deskribatzeko [Taboada and Mann, 2006b]. Horiek euskara hizkuntzarako ere egokituak izan dira [Iruskieta, 2014]. Euskaran erabiltzen diren koherentzia erlazio guztiak A. eranskinean ikus daitezke.

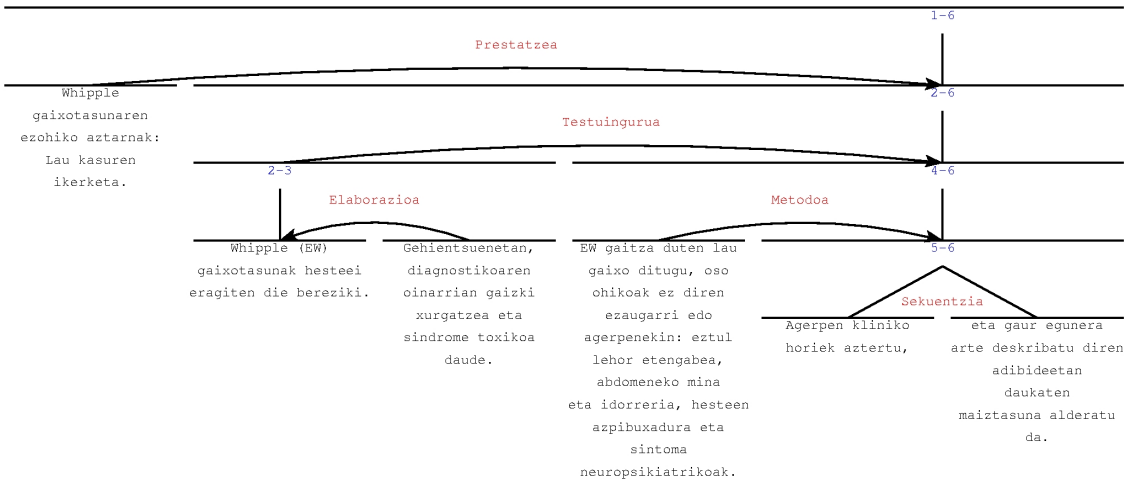
2.1. edukia: Whipple gaixotasunaren testu bat.

```
1 Whipple gaixotasunaren ezohiko aztarnak. Lau kasuren ikerketa.
2 Whipple (EW) gaixotasunak hesteei eragiten die bereziki. Gehientsuenetan, diagnostikoaren
      oinarrian gaizki xurgatzea eta sindrome toxikoa daude. EW gaitza duten lau gaixo ditugu, oso
       ohikoak ez diren ezaugarri edo agerpenekin: eztul lehor etengabea, abdomeneko mina eta
      idorreria, hesteen azpibuxadura eta sintoma neuropsikiatrikoak. Agerpen kliniko horiek
      aztertu, eta gaur egunera arte deskribatu diren adibideetan daukaten maiztasuna alderatu da.
```

Erlazioen efektuak aztertzeko, 2.1. edukian agertzen den testua adibide gisa erabiliko da. Testua segmentutan banatu da, beraien artean *span*ak osatu dira, eta EDUak eta *span*ak erlazionatu dira koherentzia erlazioen bidez. Emaitza 2.2. irudian azter daiteke. Erlazioen efektuen ulermena errazteko, 2.1. taula aztertuko da. Lehenengo zutabean 2.2. irudiko zein erlazio adieraziko den agertuko da, bigarren zutabean erlazioa lotzen duen ezkerreko EDU edo *span*a eta eskuinaldekoa hirugarren zutabean agertuz, eta laugarren zutabean erlazioaren efektua azalduko da. Ezkerreko eta eskubiko zatiak nukleo ala satelite diren adieraziko da zuhaitzaren geziaren helmuga eta jatorria aztertuz, hurrenez hurren.

Erlazioa	Ezkerreko testua	Eskumako testua	Efektua
Elaborazioa	**Nukleoa**. Whipple (EW) gaixotasunak hesteei eragiten die bereziki.	**Satelitea**. Gehientsuenetan, diagnostikoaren oinarrian gaizki xurgatzea eta sindrome toxikoa daude.	Satelitean aurkeztutako egoerak nukleoko ezaugarriren bat garatzen duela onartzen du irakurleak. Irakurleak garatutako elementua edo gaia identifikatzen du.
Metodoa	**Satelitea**. EW gaitza duten lau gaixo ditugu, oso ohikoak ez diren ezaugarri edo agerpenekin: eztul lehor etengabea, abdomeneko mina eta idorreria, hesteen azpibuxadura eta sintoma neuropsikiatrikoak.	**Nukleoa**. Agerpen kliniko horiek aztertu, eta gaur egunera arte deskribatu diren adibideetan daukaten maiztasuna alderatu da.	Irakurleak onartzen du satelitean aurkezturiko metodoak edo instrumentuak nukleoa posible egiten duela.
Sekuentzia	**Nukleoa**. Agerpen kliniko horiek aztertu,	**Nukleoa**. eta gaur egunera arte deskribatu diren adibideetan daukaten maiztasuna alderatu da.	Nukleoen arteko segida erlazioa ezagutzen du irakurleak.

2.1. taula: 2.2. irudian agertzen diren erlazio batzuen efektuaren azalpena



2.2. irudia: Gaixotasun testu bateko RST zuhaitza.

2.2 Artearen egoera

RST teorian hainbat lan eta aurrerapen egin dira, batez ere sortu diren corpusei esker. Corpusa giza etiketatzaileek etiketatutako testuen RST zuhaitzen datu-multzoa da. Corpus on batean hainbat etiketatzailek egindako kontribuzioak daude, testuak hainbat arlotakoak izanik, horien artean medikuntza, literatura, berriak, iritzi artikuluak, eta abar. Baldintza hau betetzen duen corpusa badago euskararako: Euskal RST Treebanka [Iruskieta et al., 2013], Ixa taldeak sortua[1].

Corpusak baliatuz RSTko aplikazioak sor daitezke. Adibidez, sentimenduen analisirako [Kraus and Feuerriegel, 2017], iritzi eta produktuen berrikuspen faltsuak detektatzeko [Popoola, 2017], eta RSTetan koherentzia aztertzeko [Skoufaki, 2020] proiektuak aurkitzen dira. Euskarari dagokionez, sentimenduen analisirako [Alkorta et al., 2019] eta laburpenak egiteko [Atutxa et al., 2021] proiektuak ere egin dira. Aplikazio gehiago biltzen dituen dokumentua aztertu nahi izanez gero, ikusi [Taboada and Mann, 2006a].

Simon Fraser Univertsityk mantentzen duen dibulgazioko RST web orrialdea[2] aipatu beharra dago ere, informazio ugari lor daitekeelako RSTri buruz. Gainera, RSTko hainbat publikazio biltzen dituen bibliografia[3] ere badute.

Proiektu honetan RST erlazioak era automatikoan etiketatu nahi dira, eta hori lortzeko baditugu RSTko zuhaitzak, segmentuak eta erlazioak iragartzen dituzten hainbat lan. Askok *laburtu-desplazatu* trantsizioetan oinarritutako parserrak erabiltzen dituzte, *Long Short-term Memory* [Hochreiter and Schmidhuber, 1997] sare neuronalekin batera. 2.2. taulan ingelesezko RST zuhaitzak iragartzeko sortu diren hainbat lanen zehaztasunak aurkezten dira. *Span* metrikak EDU eta *span*en arteko multzokatzea ebaluatzen ditu, Nuklearitate metrikak erlazio baten bidez lotutako EDU edota *span*ak satelite-nukleo, nukleo-satelite ala nukleo-nukleo bi zuhaitzetan berdinak diren ebaluatzen ditu, eta Erlazio metrikak bi zuhaitzetan agertzen diren erlazioen adostasuna ebaluatzen ditu.

Euskararako, ordea, sistema gutxiago sortu dira ingelesarekin konparatuz. Horietako bat [Iruskieta and Braud, 2019] da, 78,98, 55,05 eta 34,78 puntu lortuz *Span*, Nuklearitate eta Erlazio metriketan. Sistema ahaltsuagoak lortzeko asmoa ere badago, eta horregatik

[1] Euskal RST Treebankaren corpusa hemen atzigarri: https://ixa2.si.ehu.eus/diskurtsoa/

[2] SFUren RST web orrialdea, ingelesez, euskaraz, gazteleraz, portugesaz eta frantsesaz: https://www.sfu.ca/rst/index.html

[3] RST sakontzeko bibliografia: https://www.sfu.ca/rst/05bibliographies/bibs/RST_bibliography.pdf

Erreferentzia	Span zehaztasuna	Nuklearitate zehaztasuna	Erlazio zehaztasuna
[Kobayashi et al., 2021]$_{ensemble}$	87,1	75,0	63,2
[Kobayashi et al., 2021]$_{average}$	86,8	74,7	62,5
[Yu et al., 2018]	85,5	73,1	60,2
[Kobayashi et al., 2020]	87,0	74,6	60,0
[Wang et al., 2017]	86,0	72,4	59,7
[Feng and Hirst, 2014]	85,7	71,0	58,2
[Ji and Eisenstein, 2014]	82,0	68,2	57,8
[Braud et al., 2017]	81,3	68,1	56,3
[Joty et al., 2015]	83,8	68,9	55,8
[Surdeanu et al., 2015]	82,6	67,1	55,4
[duVerle and Prendinger, 2009]	83,0	68,4	55,3
[Hayashi et al., 2016]	82,6	66,6	54,6
[Nguyen et al., 2021]	74,3	64,3	51,6
[Li et al., 2016]	82,2	66,5	51,4
[Braud et al., 2016]	79,7	63,6	47,7

2.2. taula: RST erlazioak iragartzeko sortu diren hainbat sistema, eta ingeleserako lortu dituzten zehaztasun balioak

erdibideko sistemak sortu izan dira, horien artean segmentatzailea [Iruskieta et al., 2019] eta unitate zentralaren antzematea [Atutxa et al., 2019].

RST zuhaitzak iragartzen dituzten sistemen emaitzak aztertuta (2.2. taula), ikus daiteke erlazio metrikak besteak baino zehaztasun gutxiagoa lortzen duela. Horregatik, 2021ean RSTri buruzko Shared-Task[4] bat proposatu zen, DISRPT2021 [Zeldes et al., 2021], erlazioen zehaztasun metrika hobetzeko asmoz. Kasu honetan, hainbat hizkuntzetako RST corpusak eskaini ziren, partaideek segmentazio eta erlazio iragarpen sistemak inplementatzeko asmoz.

RST erlazioen iragarpenak egiteko bi sistema aurkeztu ziren. Alde batetik, DisCoDisCo [Gessler et al., 2021] sistema daukagu, eta bestetik, DiscRel [Zeldes et al., 2021] sistema daukagu. Desberdintasun nagusia sailkatzailean dago. DisCoDisCok BERT sailkatzailea erabiltzen du, eta DiscRelek, ordea, Random Forest sailkatzailea. Erlazioetan lortzen duten zehaztasuna (hizkuntza guztien batazbestekoa) % 61,82 eta % 54,23 da, hurrenez hurren. Euskararako, DisCoDisCo sistemak DiscRelek baino 16,37 puntu gehiago lortu ditu, % 60,62 zehaztasunarekin.

DisCoDisCok euskararako emaitza hobeak lortzen ditu eta, gainera, sistemaren kodea

[4]Shared-Task ikerketa arlo baten inguruan ikerketa taldeek problema zehatzak ebazteko antolatzen diren zereginak dira.

eskuragarri dago. Hori dela eta, proiektu honetan DisCoDisCo sistema oinarritzat hartuko dugu gure sistema sortzeko, KEIEBA[5], eta bertan hobekuntza proposamenak gehituko dizkiogu erlazioen emaitza hobetzeko asmoz.

2.3 Baliabideak

Atal honetan, iragarpenak egiteko beharrezkoak diren baliabideak azalduko ditugu. Alde batetik, DisCoDisCok erabiltzen dituen BERT sare neuronalak ikusiko ditugu, bestetik, koherentzia erlazioen corpusa, azkenik DisCoDisCo sistema deskribatzeko.

2.3.1 Sare neuronalak

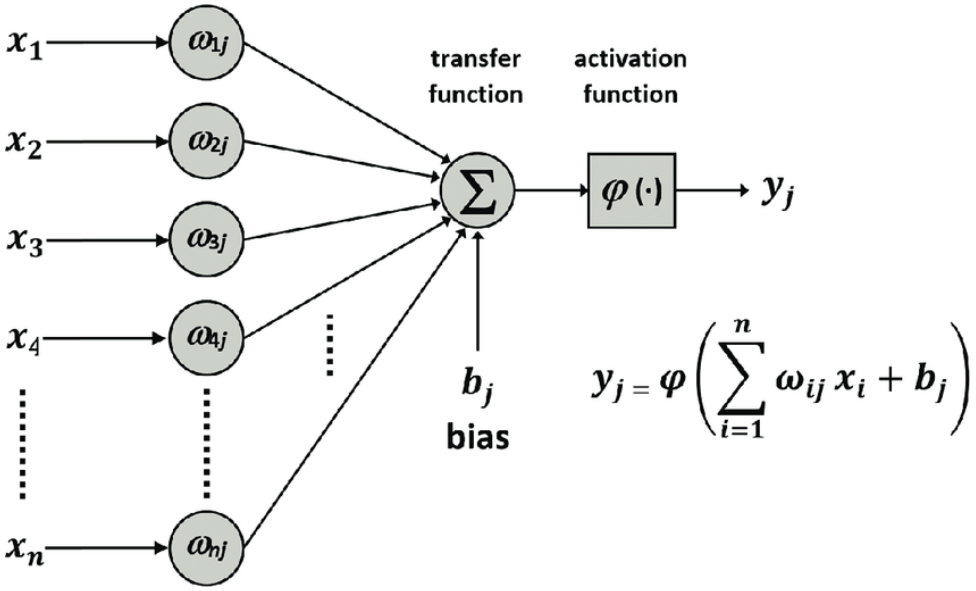
Diskurtsoko koherentzia erlazioen informazioa digitalki gorde dezakegu ordenagailuen bidez. Baina, modu horretan, kontsultatzeko informazioa besterik ez dugu izango. Ordenagailuari gai baten inguruko ezagutza eman diezaiokegu eta, honi esker, erlazioak iragarri. Horretarako, erregeletan oinarritutako sistemak erabil daitezke. Sistema mota hauetan iragarpenak lortu ahal izateko, ezagutza ordenagailuan programatu behar da erregelak zehatuz. RST erlazioetarako RASTA [Corston-Oliver, 1998] sistema dago. Era berean, ordenagailua berak bere kabuz ezagutza ikas dezan algoritmoak sortu dira, datuetatik ikasiz datu berri baten iragarpena egiteko. Prozesu honi *Machine Learning* deritzo, hau da, Ikasketa Automatikoa [Mahesh, 2020]

Proiektu honetarako Ikasketa Automatikoko teknikak erabiliko ditugu RST zuhaitzetatik koherentzia erlazioak ikasteko eta ondoren lortutako ezagutzatik iragarpenak egiteko. Dauden algoritmoen artean, sare neuronalak dira erabiliko direnak. Izan ere, azken urteotan sare neuronalek ikaragarrizko bultzada izan dute hainbat ikerketatik lortu diren sistema arrakastatsuak lortu direla eta [Dargan et al., 2019]. Bereziki, konputazio aukera handitu delako.

Sare neuronal bat, izenak aipatzen duen bezala, neuronaz osaturiko eta beraien artean konektaturiko arkitektura bat da. Bere jatorria McCulloch-Pittek giza neurona artifizialki era konputazionalean eta matematikoan aurkeztetik dator [McCulloch and Pitts, 1943]. Neurona artifizialek zenbakiekin dihardute, sarrera balioen arabera irteera balio bat ema-

[5]Izena proiektu honetako titulutik eratorria: Diskurtsoko **K**oherentzia **E**rlazioen **I**ragarpenak **E**uskararako **B**ERT sare neuronal **A**urrenetrenatua erabiliz

nez. Sare neuronal sinpleenari pertzeptroia deritzo eta neurona bakar batez osatuta dago. Pertzeptroiaren egitura 2.3. irudian ikus daiteke.

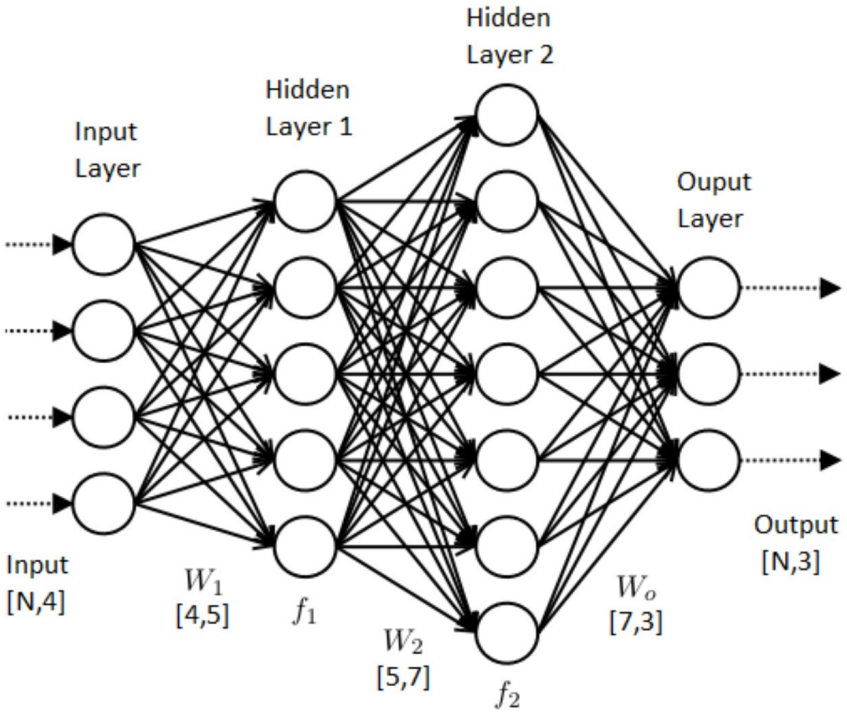


2.3. irudia: Pertzeptroiaren egitura.

Pertzeptroiak [Rosenblatt, 1958] nahi beste datu izan ditzake sarrera gisa. Datuak zenbakiak izan behar dira, bestela *One-hot encoding* aplika daiteke [Potdar et al., 2017], hau da, datuak kategoriaka antolatu behar dira bakoitzari zenbaki bat ezarriz. Ondoren, sarrera zenbakiak sarrera pisuekin biderkatzen dira, emaitzak beraien artean batuz, *bias* zenbakiarekin batera. Pisuek eta *bias* zenbakiak funtzio lineal bat osatzen dute, eta beraien balioen arabera, sarrera datuetatik emaitza bat ala beste lortuko da. Azkenik, emaitzari aktibazio funtzio bat aplikatzen zaio, azken emaitza lortuz. Aktibazio funtzioen artean, ezagunenak *sigmoidea*, *tanh* eta *ReLU* dira [Nwankpa et al., 2018].

Hala ere, funtzio lineal bakarrarekin batzutan ez da nahikoa sailkapen egokiak egiteko. Kasu honetan, funtzio lineala ez litzateke bi klaseen artean diskriminatzeko gai izango. Hori dela eta, neurona gehiagoz osatutako sare neuronalak erabiltzen dira. Ondorioz, funtzio konplexuagoak ikasteko gai dira. Sare neuronal hauek, *Multi Layer Perceptron* (MLP) deiturikoak, geruzaka antolatuta daude eta geruza bakoitza atzekoarekin eta aurrekoarekin konektatuta dago bakoitza dagokien pisuekin. Lehenengo geruza sarrerako datuei dagokie, eta azkenekoa irteera eta iragarpeneko geruzari dagokio. Erdikoak ikasteko balio dute, eta ezkutuko geruza bezala ezagutzen dira. MLPren egitura 2.4. irudian ikus daiteke.

Neuronen parametroak moldatuz, adibidez pisuak eta *bias*ak, hauek osatzen duten funtzioa egokitu dezakegu guk lortu nahi dugun irteera balioetara hurbiltzeko. Honi ikasketa



2.4. irudia: Multi Layer Pertzeptroiaren egitura.

prozesua deritzo. Ikasketari esker, datuetatik ikas dezakegu datu berrien iragarpenak egiteko. Helburua iragarpen hoberenak ematen dituzten parametroen balio eta konbinazioa lortzea da.

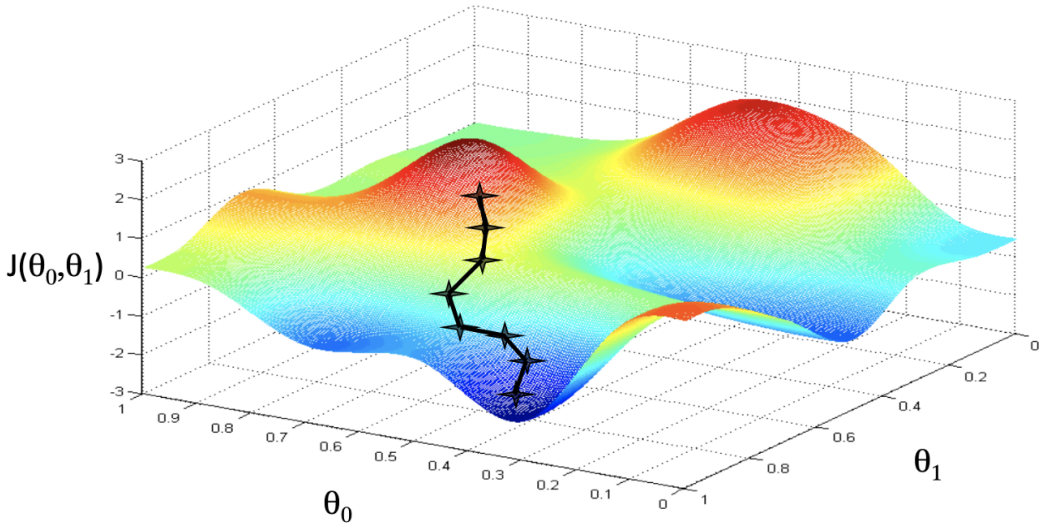
Horretarako, optimizazio algoritmoak erabiltzen dira. Lehenik eta behin galera funtzio bat ezartzen da. Galera funtzioa, parametroek dituzten balioekin iragarpenak zein onak diren esango digun funtzioa da. Hau da, lortutako eta espero den emaitzen arteko errorea kalkulatzen du, sare neuronalaren errendimendua ebaluatuz. Gure kasuan, erlazioak iragarri nahi direnez eta hauek kategoriak direnez, *Categorical Cross-Entropy* funtzioa erabil dezakegu. Funtzioa 2.1. ekuazioan adierazita dago, i kasu bakoitza, j kategoria bakoitza (kasu honetan erlazioa), y benetako kasua eta $\hat{y}$ iragarritako kasua izanik.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(\hat{y}_{ij}\right) \tag{2.1}$$

Ondoren, galera funtziotik gradientea lortzen da atzeranzko propagazioa eginez (hau da, *backpropagation*) [Rojas, 2013]. Atzeranzko propagazio algoritmoak parametroen balio aldaketak galera funtzioan nola eragiten duten analizatzen du, gradientea kalkulatzeko. Era honetan, galera funtzioa minimizatzen duen noranzkoko parametro balioak hartzeko

aukera izango dugu iragarpen hobeak egiteko.

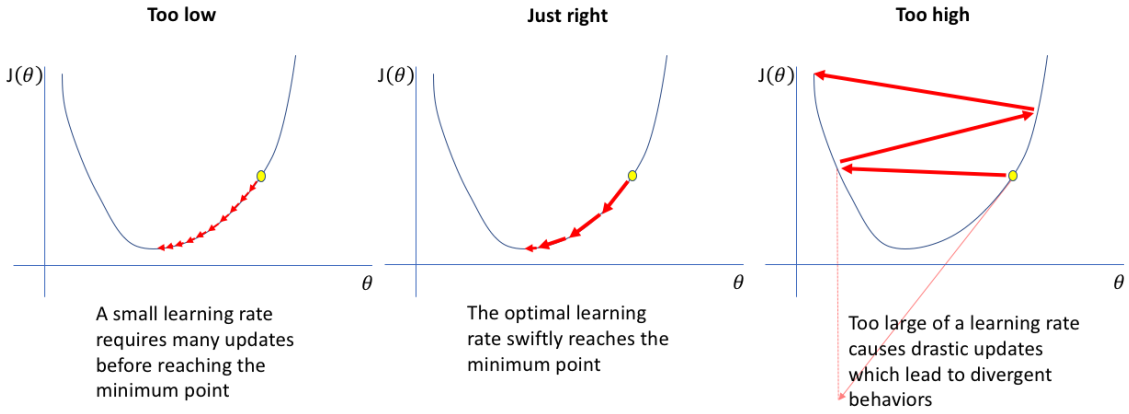
Parametroen lehenengo deribatu partziala kalkulatuz, parametroek galera funtzioan duten malda lortuko dugu. Beraz, galera funtzioa minimizatzeko noranzkoa izango dugu. Ondorioz, gradiente jaitsiera (*Gradient Descent)* algoritmoa aplikatzen da, kontrako noranzkoa jarraituz funtzioa minimizatzeko. Pausoka gradientetik jaitsiz, sare neuronalaren parametroak optimizatzen dira, galera funtzioaren balioa txikituz. Laburbilduz, pausu bakoitzeko, atzeranzko propagazioa egiten da gradientea kalkulatzeko, eta gradiente jaitsiera egiten da parametroak optimizatzeko, galera funtzioa minimizatuz. Pausuari ikasketa-tasa ere deitzen zaio (*learning-rate* ingelesez).



2.5. irudia: Gradientearen jaitsiera.

2.5. irudian galera funtzio baten adibidea azter dezakegu. Gorengo izarra da gure abiapuntua. Bertatik maldak kalkulatzen dira eta gradientetik jaisten da, izar bakoitzak eman den pausuarekin kalkulatu diren parametro balio eta galera funtzioaren balioa adieraziz. Azpien dagoen izarrak lortu den minimoa adierazten du, eta bertako parametro balioak izango dira sare neuronalean erabiliko direnak iragarpenak egiteko.

Ikasketa-tasaren balioa ere finka daiteke. Oso txikia bada, ia ziur egon gaitezke galera-funtzioko minimo batera hurbilduko dela. Hala ere, honek ikasketa denbora handitzea eragiten du, eta hori saihesteko, balioa handitu daiteke baina, handiegia bada, minimora ez heltzea eragin dezake. Hori dela eta, *learning-rate scheduler* izeneko funtzioa erabil



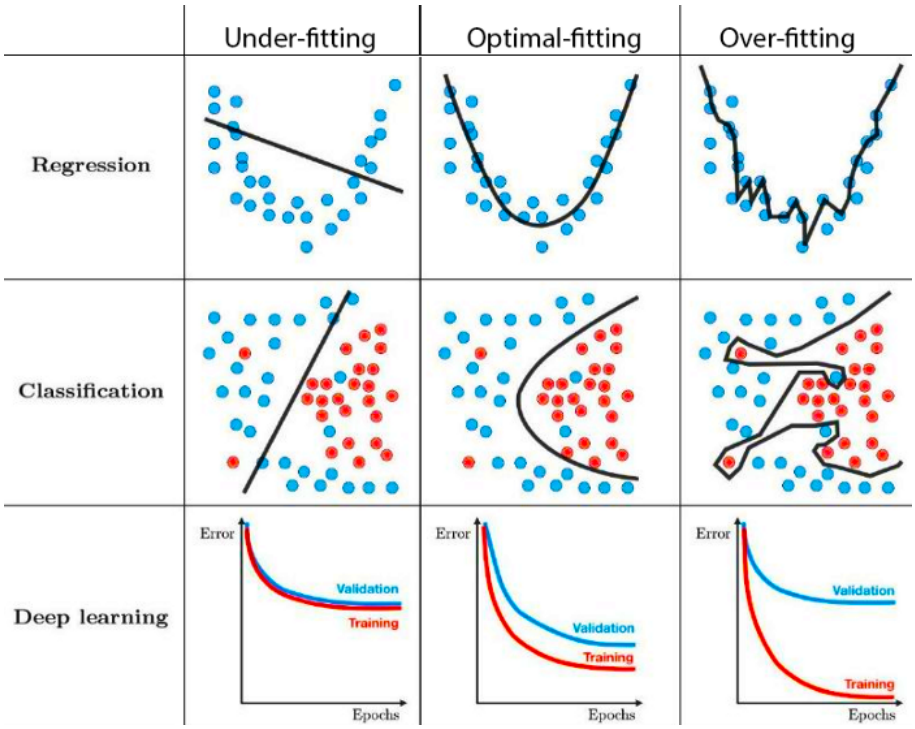
2.6. irudia: Ikasketa-tasa balioaren eragina bi dimentsioko galera funtzio batean.

daiteke. Erreminta horrek entrenamenduko lehenengo iterazioetan ikasketa-tasa handia mantentzen du, galera-funtzioko balio altuetatik ahalik eta azkarren irteteko. Iterazioak pasatzen diren heinean, ikasketa-tasa gutxituz joango da finkatutako balio batera arte, galera funtzioko minimoetara hurbiltzeko. Honen adibidea 2.6. irudian ikus daiteke.

Nahiz eta minimora heldu, horrek ez du esan nahi sare neuronal optimo bat lortu dugunik. Izan ere, ikasteko erabili dituen datuak "buruz ikasi" balitu bezala funtziona lezake eta kasu berriei iragarpen okerrak egin ditzake, batez ere kasu berria ikasitako datuetatik nahiko desberdintsua bada. Kasu honi gaindoitzea (*overfitting*) deritzo. Momentu honetan sare neuronala ebaluatzeko datu-multzoaren zehaztasuna jaisten doa entrenatzekoa hobetzen doan bitartean. Beste era batean esanda, sareak ikasi duen funtzioa oso konplexua da, datuetan ager daitekeen instantzia atipikoak (*outlier*) ikasten dituelako. Hau ekiditeko modu bat, momentu horretan sareak dituen parametroak gordetzea da. Gaindoitzearen adibide bat 2.7. irudian ikus daiteke.

Bestalde, MLPetaz gain, beste arkitektura, helburu eta funtzio desberdinak dituzten sare neuronalak ere sortu dira [Jain et al., 1996]. Batez ere, ezkutuko geruza ugari dituztenak, ikasteko kapazitatea handiagoa baitute. Azken finean, neurona gehiago izanda funtzio konplexuagoak ikas ditzakete. Hiru ezkutuko geruza edo gehiago dituzten sare neuronalei sare neuronal sakonak ere deitzen zaie, ingelesez *Deep neural network*. Era berean, ikasteko prozesuari ikasketa sakona deitzen zaio, ingelesez *Deep Learning*.

Ikasketa sakoneko sare neuronalek bultzada handia izan dute azken urteotan, eta hainbat arkitektura berri sortu dira, helburu edo ataza konkretuen ebazpena hobetzeko [Wang and Raj, 2017]. Horien artean, hizkuntzaren prozesamendurako, RST zuhaitzen erlazioak eta



2.7. irudia: Erregresio lineal problemako ikasketa falta, ikasketa optimoa eta gaindoitzea.

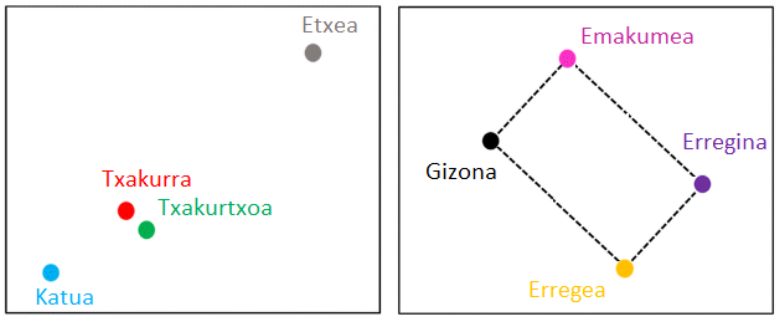
segmentuak ikasteko ere baliagarriak direla erakutsi dutenak. Sare neuronalak eta zehazki *Deep Learning* sakonki azaltzen dituen erreferentzia bibliografikoa honakoa da: [Goodfellow et al., 2016].

2.3.2 Hizkuntza-eredu neuronala: BERT

RST erlazioak iragartzeko EDUetako testuetatik ikasketa egin behar da. Hots, hizkuntza batetik egingo da ikasketa eta beraz hizkuntzaren prozesamenduko ataza izango da.

Sare neuronalek hizkuntza zenbaki eran irudikatuta izan behar dute kalkuluak egin ahal izateko. Errepresentazio honi *embedding* deritzo. Hitz bakoitza hainbat dimentsiotako bektore espazio batean kodetzen da. Era honetan, operazio matematikoen bidez bi hitz zein antzeko diren neur daiteke. Antzekotasuna neurtzeko bi bektoreen arteko distantzia euklidearra, kosinua edota produktu eskalarra erabil daiteke.

2.8. irudian *embedding*en adibide bat ikus dezakegu. Ezkerreko grafikoan hitzen *embedding* batzuk agertzen dira bi dimentsiotan adierazita. Ikus daitekeenez, “txakurra”tik hurbilen dagoen hitza “txakurtxoa” hitza da. Honek esan nahi du grafikoko “txakurra” hitzaren hitz antzekoena dela, izan ere animalia bera dira. “Katua” hitza “etxea” hitza baino



2.8. irudia: Embedding adibidea.

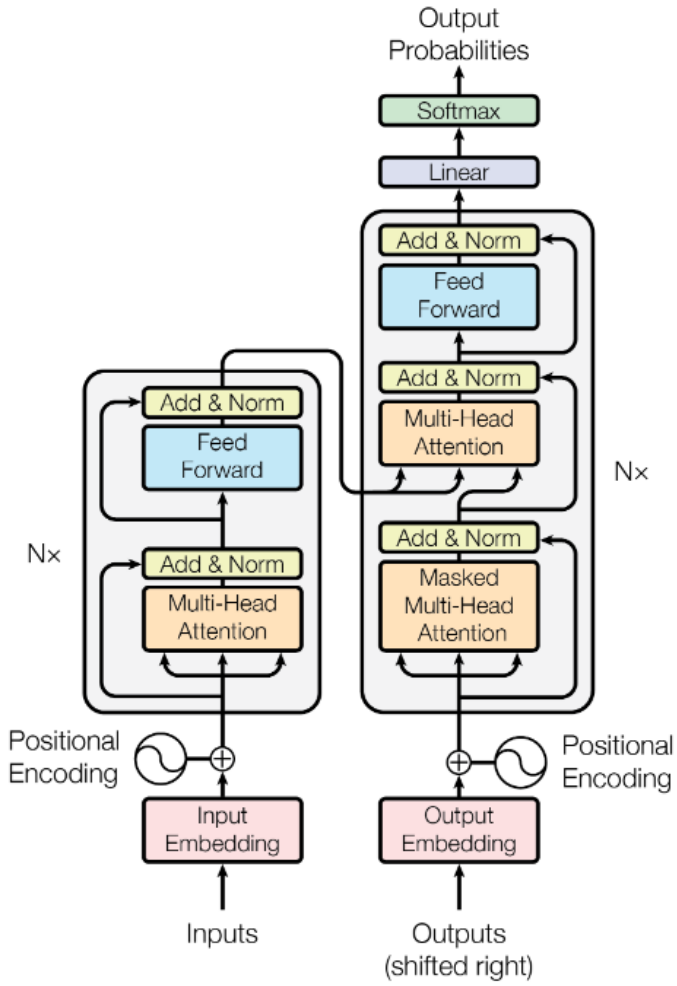
hurbilago dago "txakurra" hitzetik. Azken finean, "katua" eta "txakurra" animaliak dira, baina "etxea" eraikin bat da.

*Embedding*ei esker hitz analogoak eskura daitezke. Adibidez, 2.8. irudiko eskuineko grafikoa aztertuta, "gizona", "emakumea", "erregea" eta "erregina" hitzak daude. Ikus daiteke "gizona"tik "emakumea"ra eta "errege"tik "erregina"ra doazen bektoreak berdinak direla. Hau da, bektore horrekin gizonezko bati dagokion hitz batetik esanahi bera duen baina emakumezko bati dagokion hitz batera hel gaitezke. Normalean, operazio matematikoen bidez gauzatzen da prozesu hau. Adibidez, "erregea"ren *embedding* balioari "gizona" *embedding* balioa kendu eta "emakumea" *embedding* balioa gehituta "erregina" *embedding*ari dagokion balioa lortuko dugu.

Hizkuntzaren prozesamenduko hainbat atazarako erabiliak izan dira *embedding*ak, adibidez itzulpen automatikorako [Cho et al., 2014]. Normalean sare errekurrenteak (RNN) [Elman, 1990] erabili izan dira, esaldiak modu sekuentzialean antolatuta baitaude hitz kopuru finkorik gabe. Sare mota hauetan ezkutuko geruzetako neuronak errekurrenteak dira, hau da, uneko iterazioan ezkutuko geruzen neuronetako balioak irteera neuronara pasatzeaz gain, neurona berera bueltatzen dira sarrera gisa hurrengo iterazioan, sarrera datuekin batera.

Transformerrak [Vaswani et al., 2017], Google Brain taldeak 2017an aurkeztua, sare errekurrenteetatik pausu bat aurrerago doaz hizkuntzaren prozesamenduko atazak burutzeko. RNNekin bezalaxe *embedding*ekin eta datu sekuentzialekin funtzionatzen dute baina, aitzitik, datuak, denak batera prozesatzen dira. Transformerrak kodetzaile eta deskodetzaile geruzak ditu. Kodetzaileari esker sarrera datuen kodeketak zati aipagarrienak izango ditu. Deskodetzailean kontrakoa gauzatzen da, kodetzailean gordetako informaziorik aipagarriena datu gehiagoz osatuz. Era honetan, sareak datu garrantzitsuenak ikasten ditu. Hau lortzeko atentzio mekanismoak erabiltzen dira. Mekanismo horiek sekuentzia baten to-

ken posizio desberdinak erlazionatzen dituzte sekuentziaren errepresentazioa ikasteko. Hau da, sekuentziako atalik garrantzitsuenak ikasten dituzte. Transformerraren arkitektura 2.9. irudian azter daiteke.

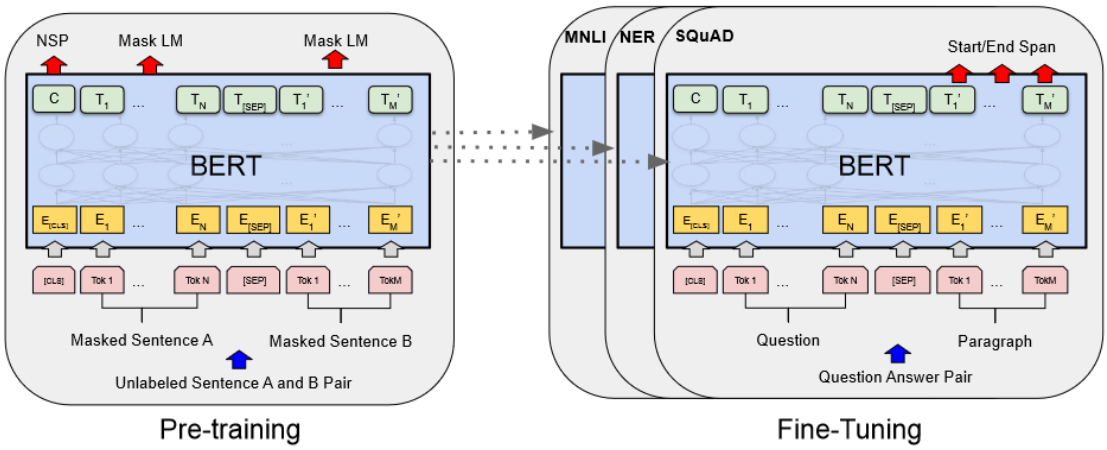


2.9. irudia: Transformerraren arkitektura.

Transformerretan oinarrituz, 2018an Google AI Language taldeak BERT [Devlin et al., 2018] hizkuntza-eredua proposatzen zuen. Izena ondorengo ingeleseko siglatik dator: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, euskaraz transformerretarako kodetzaile bidirekzionalen errepresentazioa. Hizkuntza-modelo honen erabilera artearen egoera bihurtu da hainbat hizkuntzaren prozesamenduko atazetan.

BERT sistema hizkuntzaren errepresentazioa ikasteko helburuarekin sortu da. Honekin, transferentzia bidezko ikasketa (*transfer learning*) aplikatzen da beste ataza bat ebazteko berrentrenatuz. Prozesu honi birdoitzea (*fine-tuning*) deritzo. Era honetan, alde bate-

tik, hizkuntzako errepresentazioa oinarritzat hartzen duenez, ez du zerotik ikasi beharrik izango, kostu konputazionala murriztuz. Bestetik, ez da beharrezkoa arkitekturan moldaketarik egitea, prozesua erraztuz. Prozesu hau 2.10. irudian ikus daiteke, BERTen arkitekturarekin batera.

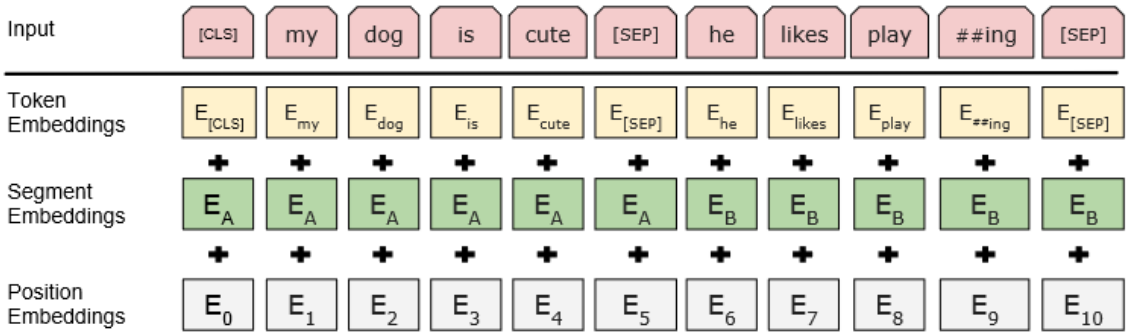


2.10. irudia: BERT arkitektura eta birdoitzea aurreentrenamenduaren ostean.

Arkitekturari dagokionez, hainbat geruzetako Transformerretako kodetzaile bidirekzionalez eta atentzio mekanismoez osatuta dago. 30000 tokenez osatuta dagoen WordPiece [Wu et al., 2016] *embedding*ak erabiltzen ditu. *Embedding* hauek hitzak zati txikiagotan banatzen ditu jatorrizko esaldia berreskuratzeko aukera izateko. Adibidez, "euritakoarekin" tokena "euritako" eta "##arekin" bezala banatzen da, biak konbinatuz berriro hitz originala lortuz.

BERT egituraren sarrera sekuentziaz osatuta dago. Sekuentzia esaldi bakarraz ala bi esaldiz osatuta egon daiteke. Sekuentzia bakoitzeko token bereziak erabiltzen dira: [CLS] tokena agertuko da edozein sekuentziaren lehenengo posizioan, eta [SEP] tokena esaldiak bereizteko bi egonez gero. Token bakoitzari dagoeneko ikasitako *embedding* balioak gehitzen zaizkio, lehenengo esaldian (A) ala bigarren esaldian (B) dagoen arabera. Gainera, token bakoitzaren posizioaren arabera, posizio *embedding* balio bat ere gehituko zaio. Lortutako emaitza *embedding* balio hauekin lan egingo du BERTek. Honen eskema 2.11. irudian ikus daiteke.

Tamaina desberdineko bi BERT erabili dira. Txikiena BERT_{BASE} da, 12 Transformer kodetzaile, 768 dimentsio eta 12 atentzio mekanismorekin, guztira 110 milioi parametroekin. Handiena BERT_{LARGE} da, 24 Transformer kodetzaile, 1024 dimentsio eta 16 atentzio mekanismorekin, guztira 340 milioi parametroekin. Aurreentrenamendua ondorengo bi



2.11. irudia: BERT sarrera gisa erabiliko dituen embeddingak, token, segmentu eta posizio embeddingeko batura izanik.

atazekin gauzatu da:

- **Maskaradun hizkuntza-eredua**. Esaldietako testuingurua ikasteko asmoz, ataza honetan sarrera sekuentzia bakoitzeko tokenen % 15 ezkutatzen dira eta sistemaren helburua jatorrizko tokena asmatzea da. Ezkutuko diren tokenak ausaz aukeratzen dira, sekuentziako kopuruaren % 15a mantenduz. Token hauetako bakoitza % 80 probabilitatearekin [MASK] tokenarekin ordeztuko da, % 10 probabilitatearekin ausazko token batekin ordeztuko da eta % 10 probabilitatearekin token berbera mantenduko da.

- **Hurrengo esaldiaren iragarpena**. Ataza honetan esaldien arteko erlazioa ikasi nahi da. Horretarako, A eta B esaldiekin osaturiko instantziak sortu dira. % 50 kasuetan B Aren hurrengo esaldia izango da "IsNext"ekin etiketatuta, eta beste % 50 kasuetan B corpusaren ausazko esaldi bat izango da "NotNext"ekin etiketatuta.

Nahiz eta aurreentrenamendu ostean birdoitzea aplika daitekeen beste atazetara egokitzeko, kasu honetan euskarazko erlazioen iragarpenerako, ezin da BERT eredu hau erabili ingelesarekin entrenatu delako. Hori dela eta, euskararekin entrenatu diren BERT ereduak erabili behar dira.

Proiektu honetan BERTeus, RoBERTa-EusCrawl eta IXAmBERT euskarazko BERT sistemak erabiliko ditugu. DisCoDisCo sisteman erabili dena BERTeus da, eta gure KEIE-BA sisteman ere erabiliko dugu. RoBERTa-EusCrawl eta IXAmBERT, aldiz, ez dira beste RST erlazioen iragarpen lanetan erabili eta ebaluatu. Horregatik, lanaren zatirik garrantzitsuenetariko bat da.

BERTeus [Agerri et al., 2020] euskararekin aurreentrenatu den BERT eredu bat da. Oinarrizko BERT_{BASE}ren arkitektura berdina erabili da, hobekuntzekin eta FastText *embedding*ekin [Bojanowski et al., 2017]. Aurreentrenamendua egiteko BERT sistema originalaren bi ataz berberak erabili dira. Corpusaren inguruan, aldizkarietako eta egunkarietako testuak hartu dira, Euskal Wikipediarekin batera. Guztira 224.6 milioi tokenez osatuta dago corpusa.

RoBERTa-EusCrawl [Artetxe et al., 2022] euskararekin aurreentrenatu den RoBERTa [Liu et al., 2019] eredua da. Horretarako, EusCrawl [Artetxe et al., 2022] corpusa sortu eta erabili da. Corpus hau interneteko euskarazko 33 web-orrialdetako testuez osatuta dago, nagusiki berriak. 12528 dokumentuz osatuta dago, guztira 423 milioi token izanik.

RoBERTak beste maskaratze estrategia bat jarraitzen du hizkuntza-eredua sortzeko. Sekuentzia bakoitza 10 aldiz bikoizten da eta, ondorioz, sekuentzia bakoitza 10 era desberdinetan maskaratuta egongo da. Gainera, maskaratzea entrenamendu puntuan gauzatzen da. Hortaz, entrenamendu iterazio bakoitzean sekuentzia bakoitzeko 10 maskaratze desberdin sortuko dira. Bestalde, ez da aurreentrenatu hurrengo esaldiaren iragarpen atazarako, emaitza hobeak lortzen baititu bakarrik maskaradun hizkuntza-eredu atazarekin entrenatuta. Arkitekturari dagokionez, BERT_{LARGE} bezalakoa da.

IXAmBERT [Otegi et al., 2020] eredua mBERT [Devlin et al., 2018][6] ereduan oinarritzen da. mBERT beste hizkuntzekin eskuragarri jartzeko, BERT sistema bera erabiltzen dute baina 100 hizkuntzetako Wikipedia corpusak era jarraituan corpus gisa erabilita.

Aitzitik, euskararen emaitzak hobetzeko asmoz IXAmBERTek hiru hizkuntza bakarrik ditu: euskara, ingelesa eta gaztelera. Euskarako BERTeuseko corpus bera erabiltzen da, ingeleserako eta gaztelerarako, ordea, Wikipedia corpusa. Galdera-erantzun iragarpen atazerako aurreentrenatu da. Horretarako *embedding* geruza bat gehiago gehitu zaio, History Answer Embedding [Qu et al., 2019] hain zuzen ere. *Embedding* horiek elkarrizketa historialaren parte diren tokenei garrantzia ematen diete.

Laburpen gisa erabiliko diren BERT ereduen ezaugarriak 2.3. taulan azter daitezke.

[6] https://github.com/google-research/bert/blob/master/multilingual.md

Ezaugarria	**BERTeus**	**RoBERTa-EusCrawl**	**IXAmBERT**
Erreferentzia	[Agerri et al., 2020]	[Artetxe et al., 2022]	[Otegi et al., 2020]
Oinarria	BERT	RoBERTa	mBERT
Corpusa	Aldizkari, egunkari euskarazko testuak + Euskal Wikipedia	33 euskarazko web-orrialdetako testuak	Aldizkari, egunkari euskarazko testuak + Euskal Wikipedia + Gaztelerako Wikipedia + Ingelesezko Wikipedia
Aurre-entrenamendua	Maskaradun hizkuntza-eredua + hurrengo esaldiaren iragarpena	Maskaradun hizkuntza-eredua	Galdera-erantzun sistema
Arkitektura	BASE	LARGE	BASE

2.3. taula: Erabiliko diren BERT ereduen ezaugarri nagusiak

2.3.3 Corpusa

Proiektu honetan erabiliko dugun corpusa DISRPT2021 Shared-Taskean proposatutakoa[7] da, euskararen kasuan Euskal RST Treebank iturri gisa izanik. Azken finean, gure KEIEBA sistemak oinarrian DisCoDisCo sistema erabiltzen du, eta ebaluaziorako corpus bera erabilita hobekuntzak kuantifikatu daitezke bi sistemen emaitzak aztertuta. Corpusak hurrengo ezaugarriak ditu[8]:

- **Esaldi kopurua:** 2360
- **Token kopurua:** 45780
- **Dokumentu kopurua:** 164
- **EDU unitate kopurua:** 4202
- **Erlazio kopurua:** 2533
- **Erlazio motak:** 29

[7] DISRPT2021 Shared-Taskeko corpusak hemen eskuragarri: https://github.com/disrpt/sharedtask2021

[8] Datu hauen iturria [Zeldes et al., 2021] da. Egileek jarraitu duten kontaketa metodoaren arabera 2533 erlazio daude. Erlazioak instantziaka kontatzen baditugu, ordea, 3825 erlazio daude. Instantzien araberako kontaketa jarraituko dugu erlazioen distribuzioa eta kopurua aztertzeko (2.5. taula)

Corpuseko elementu bakoitza hiru fitxategiren bitartez adierazten da: *conllu*, *tok* eta *rels*. Dokumentuek token bakoitzaren informazioa, dokumentu bakoitzaren tokenizazioa eta segmentu bakoitzaren hasiera adierazpena eta segmentuen arteko erlazio informazioa adierazten dute, hurrenez hurren.

Ikasteko erabiliko den fitxategia *rels* izango da. Edukia *tsv* formatuan du, hau da, tabulazioz bereizitako zutabeak, lerro bakoitzean instantzia bat egonik. Instantzia bakoitza 12 zutabez osatuta dago. 2.2. irudian agertzen den *Testuingurua* erlazioaren instantzia fitxategi honetan nola agertzen den 2.4. taulan ikus daiteke.

Zutabea	Balioa
doc	GMB_Medikuntza_0503
unit1_toks	10-19
unit2_toks	61-65
unit1_txt	Whipple (EW) gaixotasunak hesteei eragiten die bereziki .
unit2_txt	Agerpen kliniko horiek aztertu ,
s1_toks	10-19
s2_toks	61-77
unit1_sent	Whipple (EW) gaixotasunak hesteei eragiten die bereziki .
unit2_sent	Agerpen kliniko horiek aztertu , eta gaur egunera arte deskribatu diren adibideetan daukaten maiztasuna alderatu da .
dir	1>2
orig_label	testuingurua
label	background

2.4. taula: 2.2. irudian agertzen den *Testuingurua* erlazioaren instantzia *rels* fitxategian

Lehenengo zutabeak, *doc*, instantzia zein dokumenturi dagokion adierazten du. *unit1_txt* lotutako ezkerreko EDU ala *span*aren testua adierazten du, *unit1_toks*en *.tok* fitxategian EDU ala *span* honi dagokion token tartea agertuz. *unit1_txt* testuko esaldi osoa *unit1_sent*-ean agertzen da, *s1_toks*en bere token tartea agertuz. *unit2_txt*, *unit2_toks*, *s2_toks* eta *unit2_sent* zutabeek emandako deskribapen berak baina lotutako eskuineko EDU ala *span*-ari dagokie. *orig_label*ek eta *label*ek bi segmentuak lotzen dituen erlazioa adierazten dute, euskaraz eta ingelesez. *dir*ek noranzkoa adierazten du, jakiteko nukleoa ala satelitea zein den. “>” karakterea agertzen bada, nukleoa bigarren unitatea izango da eta satelitea lehenengoa, aldiz, “<” agertzen bada kontrako kasua izango da. Erlazio multinuklearren kasuan ere, nahiz eta biak nukleo izan “<” karakterea ere agertzen da. Dena den, erlazioaren izena aztertuta erlazioa multinuklearra den ala ez jakin dezakegu (ikusi A. eranskina).

Hala ere, datu-baseak ez du RST zuhaitzen egitura eta sakonera mantentzen. Alegia, zuhaitzeko erlazio bakoitzak lotzen dituen segmentuek ez dute *span* hori osatzen duen

testu osoa, baizik eta nukleo den sakonerako EDUko testua. Aurreko adibidearekin jarraituz eta 2.2. irudian agertzen de *Testuingurua* erlazioa aztertuz, satelitea $span_{2,3}$ da. Bere azpian *Elaborazioa* erlazioaren nukleoa EDU_2 dago, beraz hori izango da *unit1_txt* zutabean agertuko den testua. *unit2_txt*ko testua lortzeko *Testuingurua* erlazioko nukleoko azpiko *Metodoa* erlazioko nukleotik abiatu behar da ($span_{5,6}$). Bertan *Sekuentzia* erlazioa dago. Erlazio mota hau multinuklearra denez, ezkerreko nukleoa aukeratuko da, EDU_5, eta hori da agertuko den testua.

Ikasketa gauzatzeko asmoz eta beste sistemetako emaitzekin konparatzeko asmoz, corpusa hiru zatitan banatuta eskaintzen da: *train* ikasketa eta entrenamendua egiteko, *dev* ikasketa nola doan aztertzeko eta *test* behin ikasketa bukatuta lortutako eredua ebaluatzeko. *Train*, *dev* eta *test* 116, 24 eta 25 dokumentuz osatuta daude. Erlazioen distribuzioa hiru zatietan 2.5. taulan ikus daiteke.

Corpuseko *test* zatia errespetatzea nahitaezkoa izango da beste sistemekin edota guk sortuko ditugun beste ereduekin, emaitzak konparatu ahal izateko. Horregatik, sortutako ereduek aipatutako corpuseko distribuzio hau jarraituko dute. Bestalde, balioztatze-gurutzatuko teknika ere erabiliko dugu, noski *test* zatia berdin mantenduz.

2.3.4 RST erlazioak iragartzeko sistema: DisCoDisCo

DisCoDisCo [Gessler et al., 2021], Georgetown unibertsitateko Corpling Lab taldeak proposatua, RST erlazioak iragartzen dituen sare neuronala da. Eraikitako sistema eta kodea GitHuben eskuragarri dago[9].

DisCoDisCo AllenNLP liburutegiarekin [Gardner et al., 2017] sortu da, hizkuntzaren prozesamenduko atazetarako garatua izan dena eta PyTorch [Paszke et al., 2019] liburutegia oinarri duena. AllenNLPri esker, programazio eta esperimentazio fitxategiak bereizi daitezke. Ondorioz, esperimentazio fitxategian konfigurazio aldagaiak alda daitezke oinarrizko sistemaren kodea ikusi eta ukitu gabe.

PyTorchek eskaintzen dituen funtzioak erabiliz programatzen da eredua. Garatu diren funtzioak datu-basearen irakurketaz eta BERTaren entrenamenduaz arduratzen dira. Funtzio nagusiek alias bat izango dute liburutegiarekin API bat bezala erabili ahal izateko.

Esperimentazio fitxategian entrenamendurako aldagai guztiak finkatzen dira. Horien artean, erabiliko diren entrenamendu atributuak, iterazio kopurua, iterazioko instantzia sorta

[9] DisCoDisCo sistemaren errepositorioa: `https://github.com/gucorpling/DisCoDisCo`

Erlazioa	**Guztira**	**Train**	**Dev**	**Test**
elaborazioa	811	543	128	140
prestatzea	372	237	62	73
lista	320	204	62	54
helburua	247	169	28	50
zirkunstantzia	224	149	27	48
ondorioa	209	141	34	34
metodoa	175	115	23	37
testuingurua	174	125	20	29
kausa	160	98	25	37
konjuntzioa	159	102	32	25
ebaluazioa	141	99	26	16
sekuentzia	128	81	24	23
kontzesioa	115	81	17	17
kontrastea	109	72	16	21
interpretazioa	81	55	13	13
birformulazioa	66	39	14	13
justifikazioa	64	44	12	8
baldintza	46	25	12	9
antitesia	44	33	6	5
arazo-soluzioa	43	28	7	8
ebidentzia	38	20	10	8
disjuntzioa	20	11	6	3
ahalbideratzea	19	18	1	0
laburpena	16	13	0	3
motibazioa	15	9	4	2
bateratzea	11	6	4	1
ez-baldintzatzailea	8	6	1	1
alderantzizko-baldintza	8	8	0	0
aukera	2	2	0	0

2.5. taula: Erlazioen maiztasuna corpusean eta train, dev eta test zatietan

kopurua, corpusaren irakurketa eta atributuen lorpen funtzioen aliasak, BERT eredua, optimizatzailea, ikasketa-tasa...

Azkenik, bash lengoaian programaturiko script batekin sistema martxan jartzen da, bai ereduak entrenatzeko bai entrenatutako ereduen emaitzak lortzeko. AllenNLPko *train* eta *test* funtzioei deituz konfigurazio fitxategian ezarritako aldagai balioekin liburutegia bera arduratzen da programatutako funtzioen exekuzioaz, kasu honetan ereduak ikasteaz eta entrenatzeaz, eta ereduak ebaluatzeaz, hurrenez hurren.

Entrenamendua BERTeus [Agerri et al., 2020] hizkuntza-ereduarekin gauzatzen da. Erlazioak iragartzeko atazaren birdoitzea gauzatzeko, lehenik eta behin sarrera sekuentzia finkatu da. Ereduak bi unitate onartzen dituenez, instantzia bakoitzeko hura osatzen duten bi segmentuek osatuko dute sekuentzia, beti ere beraien artean [SEP] tokena egonez hasierako [CLS] tokenarekin batera.

Ondoren, sistema atributuekin aberasten da. Erlazioen norabidea pseudo-tokenak erabiliz ezartzen da. Erlazioa ezkerretik eskuinera badoa (instantzian 1>2 agertzen da), orduan "}" eta ">" tokenekin inguratzen da lehenengo unitatea. Adibidez: "[CLS] } hasi #ko gara ? > [SEP] ez [SEP]". Aldiz, kontrako kasuan "<" eta "{" tokenekin inguratzen da bigarren unitatea. Adibidez: "[CLS] mediku #aren #era noa [SEP] < gaixo nago #elako { [SEP]". Era honetan, erlazio norabidearen irudikapena sarrera sekuentzian adierazi daiteke.

Atributua	**Deskribapena**
Distance	Lehenengo eta bigarren unitateen artean dauden segmentu kopurua.
Same Speaker	Ea lehenengo eta bigarren unitatea hizlari berarenak diren ala ez.
Case*	Unitatea titulua den edota izen berezi bat duen adierazten du.
Children*	Unitateak zenbat azpiunitate dituen (*span* izan behar da, EDU bada 0 izango ditu).
Genre	Dokumentuaren generoa.
Lexical Overlap	Hitz errepikakorrak lehenengo eta bigarren unitateen artean.
Discontinuous*	Unitatearen tokenak aldamenekoak diren ala ez.
Position*	Unitatearen posizioa dokumentuan, 0.0tik 1.0rako balioen artean, ezkerrago ala eskuinago dagoen adieraziz.
Is Sentence*	Unitatea esaldi oso bat den ala ez.
Doc Length	Dokumentuaren luzera tokenetan.
Length Ratio*	Unitatearen luzeraren ratioa dokumentuarekiko.

2.6. taula: DisCoDisCo sisteman erlazioak iragartzeko aberasteko atributuak

Erabiliko diren beste atributuek, berriz, ez dituzte sarrera sekuentziako bi esaldiak zuzenean erlazionatzen. Hori dela eta, sistema aberasteko, ez dira sarrera sekuentzian gehituko

baizik eta kodetzaile geruza baino lehen, [CLS] ondoren. Atributu hauek 2.6. taulan ikus daitezke. Asteriskoa dutenak bi alditan aplikatzen dira, unitate bakoitzeko.

Hala ere, hizkuntzaren arabera, emaitza hobeak lortzeko atributu bakoitzaren kontribuzioa desberdina da. Egindako esperimentazioaren arabera hizkuntza bakoitzeko atributu zehatz batzuk aukeratu dituzte. Euskararen kasuan, bakarrik bigarren unitateari dagokion *position* atributua erabili da. Noski, erlazioen norabidea beti erabiltzen da.

DisCoDisCo euskarazko corpusarekin eta BERTeus ereduarekin exekutatuz gero % 59.59-ko zehaztasuna lortzen da. Egileek erabilitako aldagaiekin (2. unitateko *position* aldagaiarekin) % 60.62ko zehaztasuna lortzen da, 1.03 puntu hobetuz.

3. KAPITULUA

Euskarazko RST erlazioak iragartzeko sisteman moldaketak eta hobekuntzak

Behin RST, BERT, corpusa eta DisCoDisCo sistemaren funtzionamendua aztertu dugula, kapitulu honetan moldaketak proposatuko ditugu gure KEIEBA sistema sortzeko, euskararako lortzen diren emaitzak hobetzeko asmoz.

3.1 Hobekuntza moldaketak

3.1.1 Atributu berriak

Jatorrizko DisCoDisCo sisteman euskararako erabiltzen den atributu bakarra bigarren unitateari dagokion *position* da. Honen arrazoia, egileek aipatzen dutenez, hizkuntza batzuekin atributu guztiak erabiltzeak gaindoitzea eragiten duela da, guztien erabilerak eragiten duen dimentsionalitateagatik. Dena den, euskararen kasuan dagoeneko atributu bakar horrekin gaindoitze egoerara heltzen da entrenamendua. Hau da, entrenamenduan % 100eko zehaztasunera heltzen da. Are gehiago, atributurik gabe erabilita ere heltzen da, eta ereduaren ebaluazioak emaitza hobeak ematen ditu atributua erabilita baino.

Dena den, gaindoitze egoerara heldu baino lehen gordeko ditugu ereduak, iragarpen hobeak lortzeko. Hori dela eta, eskuragarri dauden atributu guztiak erabiliko dira KEIEBA sisteman, zehazki 2.6. taulan agertzen direnak. Gaindoitzearen arazoa konpontzea ez da espero (izan ere, gaindoitzea atributu gabe eta bat erabilita agertzen bada, atributu gehia-

go erabilita ere gaindoitzea mantenduko da), baina gaindoitze egoerara heldu baino lehen lortutako ereduekin geratuko garenez, uste dugu atributu berriek zehaztasun altuagoa lortzen lagundu dezaketela. Azken finean, beste esperimentu batzuetarako atributu berriek garrantzia izan lezakete ebaluazioan eta emaitza hobeak lor litezke. Adibidez, corpusa beste era batera antolatzen bada, atributuetan moldaketak eginez, atributu berriak erabiliz edo beste BERT hizkuntza-eredua erabiltzen bada.

3.1.2 RoBERTa-EusCrawl eta IXAmBERT BERT ereduen erabilera

Jatorrizko DisCoDisCo sistemak BERTeus erabiltzen du. Hizkuntza-eredu honetaz gain, 2.3.2. atalean aipatutako RoBERTa-EusCrawl eta IXAmBERT ere erabiliko ditugu gure KEIEBA sisteman. Interesgarria izango da erlazioen iragarpeneko birdoitzean bi eredu hauen errendimendua aztertzea. Alde batetik, RoBERTa-EusCrawl ez da aurreentrenatu hurrengo esaldiaren iragarpen atazarako, eta desabantaila nabaria izan daiteke, baina euskararako erabiltzen duen corpusa aberasgarriagoa da. Beste alde batetik, IXAmBERT galdera-erantzun atazarako aurreentrenatu da, baina hurrengo esaldiaren iragarpen ataza ez bezala, ez dauka iragartzeko etiketarik.

3.1.3 Testuen generoa

Bistan da testu genero bakoitzak idazteko estilo bat duela. Kontatu nahi denaren araberakoa da hori, ez baita era berean idatziko testu zientifiko ala errezeta bat, adibidez. Hori dela eta, erlazio mota desberdinak lortuko ditugu RST analisia genero bakoitzeko testu bati eginez gero.

DisCoDisCo sistemak genero atributua erabiltzen du ikasteko, baina euskarako corpusean ez. Horregatik, KEIEBA sisteman moldaketak gauzatuko ditugu euskararekin genero atributua erabil dezan. Generoa lortzeko, corpusak Euskal RST Treebanka du iturri gisa, eta bertatik eskura dezakegu testuen generoa. Izan ere, dokumentu bakoitzak bere izenean kode bat du generoa adierazteko. Beraz, alde batetik corpuseko dokumentu izenei kode ondoan dagokion genero izena gehitu zaio, eta corpusa irakurtzeko programan gehitutako generoa atributu bezala gordetzen da.

Hainbat genero balio izango ditugu iragarpenean laguntzeko. Literatura, liburuen deskribapenaz osatutako testuak (28 dokumentu); laburpena, testuen laburpenaz osatutako testuak (24 dokumentu); medikuntza, 2000 eta 2008 artean *Medical Journal of Bilbao*

aldizkarian euskaraz argitaratutako artikuluetako laburpena (20 dokumentu); osasuna, giza osasunaz diharduten testuak eta deskribapenak (20 dokumentu); terminologia, 1997an UZEIk antolatutako *International Conference on Terminology* konferentziako jardunaldiko terminologiari buruzko testuak (20 dokumentu); zientzia, 2008an Euskal Herriko Unibertsitateak antolatutako *Research Conference*n parte hartutako artikuluen testuak (20 dokumentu); informatika, informatikaren inguruan diharduten testuak (19 dokumentu); eta bestelakoak, aurreko kategoriatakoak ez diren testuak (13 dokumentu).

3.1.4 Hitzetako informazioa

Corpusa atributu gehiagoz osatu daiteke iragarpenak hobetzeko. Kasu honetan, corpuseko instantzia bakoitza hitzetako informazioarekin bete da, lehenengo eta bigarren unitateko testuak baliatuz. Informazioa lokailu eta aditz hitzetatik eskuratuko da.

Aditza esaldiko hitz garrantzitsuena da. Izan ere, aditzik gabe ezin da esaldi bat sortu, gramatikako dependentzia zuhaitzetan erroa baita [Aranzabe et al., 2004]. Semantikaren aldetik aditzak ekintza (*irakurri*), egoera (*izan*) edo gertaera (*irabazi*) adierazten du. Sintaxiaren aldetik, ordea, aditzak denbora, kasua, pertsona eta beste atributu gehiago batzuk adieraz dezake.

Datuetako *conllu* fitxategian, segmentu bakoitzeko token bakoitzaren informazio gehigarria eskuragarri dago. Hala nola, forma, lemma (hau da, hitzaren forma kanonikoa), *Part-Of-Speech* (hau da, gramatika kategoria), burua, dependentzia, ezaugarriak... Erabiliko diren parametroak ondorengoak dira:

- **conj**. Hemen lokailuak gordeko ditugu. Horretarako unitatearen tokenen *Part-Of-Speech* aztertuko dugu eta “CCONJ” denean, tokenak gordeko ditugu. Adibide bezala, 3.1. edukia jarraituz, 21. IDa duen instantziak bere *Part-Of-Speech* balioa “CCONJ” da (UPOS zutabea), beraz FORM zutabeko tokena gordeko dugu, “baina”.

- **vvoice**. Unitatearen token baten *Part-Of-Speech* “VERB” denean, ezaugarriak zutabeko “Voice” balioa gordeko dugu. “Voice”k aditzak adierazten duen ekintzaren eta partaideen arteko erlazioa deskribatzen du. Hau da, aditza aktiboa ala pasiboa den. Adibide bezala, 3.1. edukia jarraituz, 8., 2., eta 5. IDa duten instantziek “VERB” dute UPOS zutabean, beraz ezaugarriak zutabea (FEATS) aztertuko dugu. Lehenengoak “Voice” ezaugarria du, eta “Cau” balioarekin geratuko gara. Balio honek aditza kausatiboa dela adierazten du.

- **vcase**. Unitatearen token baten *Part-Of-Speech* "VERB" denean, ezaugarriak zutabeko "Case" balioa gordeko dugu. "Case"k hitzen eginkizuna adierazten laguntzen du. Adibide bezala, 3.1. edukia jarraituz, 8., 2., eta 5. IDa duten instantziek "VERB" dute UPOS zutabean, beraz ezaugarriak zutabea (FEATS) aztertuko dugu. Bigarrenak "Case" ezaugarria du, eta "Abs" balioarekin geratuko gara. Honek esan nahi du aditza kasu absolutiboan dagoela.

- **vmod**. Unitatearen token baten *Part-Of-Speech* "VERB" denean, ezaugarriak zutabeko "Mood" balioa gordeko dugu. "Mood" aditzak azpikategoriatan sailkatzeko erabiltzen da. Adibide bezala, 3.1. edukia jarraituz, 8., 2., eta 5. IDa duten instantziek "VERB" dute UPOS zutabean, beraz ezaugarriak zutabea (FEATS) aztertuko dugu. Hirugarrenak hainbat ezaugarri ditu FEATS zutabean, eta adierazitakoen artean "Mood" ezaugarria du, beraz "Ind" balioarekin geratuko gara. Honek aditzaren azpikategoria indikatiboa dela adierazten du.

3.1. edukia: conllu fitxategiko instantzia adibideak.

```
ID  FORM    LEMMA   UPOS    XPOS    FEATS   HEAD    DEPREL  DEPS    MISC
21  baina   baina   CCONJ   CCONJ   _   32  cc  _   BeginSeg=Yes
8   adierazten  adierazi    VERB    VERB    Voice=Cau  5   conj    _   _
2   finkatzeko  finkatu VERB    VERB    Case=Abs|Definite=Ind  7   advcl   _   _
5   dauden  egon    VERB    VERB    Aspect=Prog|Mood=Ind|Number[abs]=Plur|Person[abs]=3    6
    acl _   _
```

Unitate bakoitzeko aipatutako aldagaien lehenengo bi agerpenak gordetzen dira, beraz, aldagai bakoitzeko lau agerpen egongo dira. Hauek corpuseko *rels* fitxategian gehitu dira zutabe moduan, eta sisteman moldaketa egin da zutabe horiek kontuan har ditzan.

Hala ere, aipatu beharra dago berez ez dugula corpuseko *conllu* fitxategitik informazioa lortu, baizik eta instantzia bakoitzeko unitateetatik gure *conllu* fitxategia lortu dugula, UDpipe tresnari esker [Straka and Straková, 2017]. Tresna honek testu batetik *conllu* fitxategia lor dezake. Gure kasuan instantziaka joango gara unitate bakoitzeko *conllu* informazioa lortzen eta corpusean gehitzen.

Era honetan, alde batetik, uneko unitateko *conllu* informazioa lortzen dugu berehala eta ez dugu behar datuetako *conllu* fitxategian unitatea bilatu behar datuak eskuratzeko. Hori dela eta, programaren errendimendua hobea da azkarrago lortuko dugulako beharrezko informazioa. Bestetik, tresna berritzen doa, eta gerta liteke erabilitako azken tresnako euskararako ereduak *conllu*n datu zehatzagoak ematea corpuseko *conllu*ko informazioa baino.

3.1.5 Erlazioak taldekatuta

Proposamen honetan, erlazioak taldekatzea proposatzen da. Nahiz eta multzokatzea nola egin guk erabaki dugun, ideia hau dagoeneko beste lan batzuetan aplikatua izan da [Mann and Thompson, 1988] [Iruskieta, 2014]. Taldeak antzeko efektuak sortzen dituzten erlazioez osatuta egongo dira. Ikasterako orduan, corpusean agertzen diren erlazioak bakoitza dagokion taldearen izenarekin ordeztuko dira.

Taldea	**Erlazioak**
Aurkakotasun Taldea	Antitesia
	Kontzesioa
	Kontrastea
Kausa Taldea	Kausa
	Interpretazioa
	Justifikazioa
	Motibazioa
	Ondorioa
Baldintza Taldea	Baldintza
	Ez-baldintzatzailea
Bateratze Taldea	Konjuntzioa
	Bateratzea
	Lista
Elaborazioa Taldea	Elaborazioa
	Ebidentzia
Birformulazioa Taldea	Birformulazioa
	Laburpena

Talde gabe mantendutako erlazioak
Testuingurua
Zirkunstantzia
Disjuntzioa
Ahalbideratzea
Ebaluazioa
Metodoa
Aukera
Prestatzea
Helburua
Sekuentzia
Arazo-soluzioa
Alderantzizko-baldintza

3.1. taula: Erlazioen taldekatzea

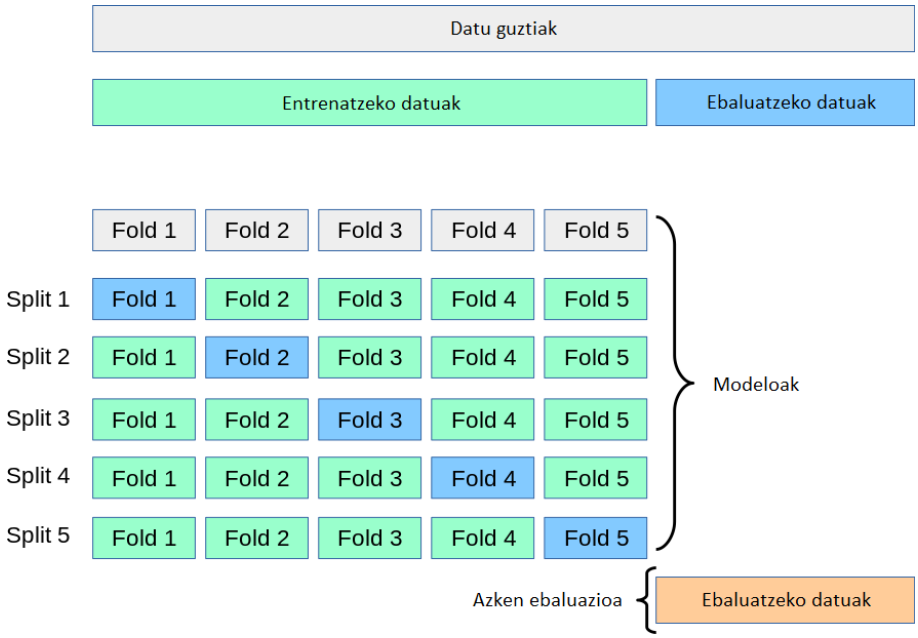
Argi dago erlazio originalen iragarpenak galduko ditugula bakarrik zein talderi dagokion

jakinez, baina honekin iragarpenen kalitatea igotzea espero da. Noski, zehaztasuna ere jaitsi daiteke, baina igotzekotan, ez du esan nahi iragarpena hobea denik. Izan ere, taldearen errendimendua konparatu beharko litzateke erlazio originalak iragartzen dituen eredu batekin. Adibidez, zehaztasun globala igo da baina talde baten zehaztasuna jaitsi da, talde horretako erlazio bakoitzak taldekatu baino lehen lortzen zuen zehaztasunarekin konparatuz.

Beraz, ataza honen helburua ez da izango zehaztasun altuagoa lortzea, baizik eta erlazioak banaka izanik eta taldeka lortzen diren emaitzak aztertzea. Alegia, lortzen diren ondorioetatik talde batzuk soilik, guztiak ala talderik ez erabiltzea erabaki daiteke sistemarako. Erabiliko den taldekatze estrategia 3.1. taulan azter daiteke.

3.1.6 Balioztatze gurutzatua

Balioztatze gurutzatua (*cross-validation)* datuen laginketa metodoa da [Hastie et al., 2009], gaindoitzea ekiditeko helburua duena. Gaindoitzea ekidinez, emaitza hobeak iragarriko dira eta, hori dela eta, teknika hau inplementatuko dugu.



3.1. irudia: k-Fold Cross-validation, 5 multzotakoa.

Erabiliko dugun metodoa *K-fold cross-validation* da. Entrenatzeko datuak K multzotan (*fold*) banatzen dira eta k iterazio aplikatzen dira. i iterazio bakoitzean eredu bat sortzen

da, non i multzoa eredua ebaluatzeko erabiliko den eta gainontzeko multzoak eredua entrenatzeko. 3.1. irudian 5 multzoko balioztatze gurutzatuko adibidea ikus daiteke.

k balio handia erabiltzen bada alborapena txikituko da baina bariantza handituko da [Kohavi et al., 1995]. k balioa txikia bada, kontrako kasua izango dugu. Ohikoena $K = 10$ erabiltzea da, eta hori da hain zuzen erabiliko duguna.

Kontuan hartu behar da azken iragarpena lortzeko, ebaluazioko datuetako instantziak erabiliko direla soilik, eta ez *k-fold cross-validation* gauzatzean eredu bakoitzari esleitu zaizkion zatiko ebaluazio instantziak. Izan ere, instantzia horiek entrenatzeko fitxategitik datoz eta, gainera, beste multzoek entrenamendurako erabili dituzte.

Ataza honen garapena 3 pausotan gauzatu da, pausu bakoitzeko script bat inplementatuz:

1. **Datu-basea**. Edozein k-rako multzoak sortuko dira. Hasteko, entrenatzeko datuetako dokumentuak k zatitan bereiziko dira. Zati bakoitzean joango diren dokumentuak aukeratzea modu ordenatuan ala ausaz egin daiteke. Ondoren, k karpeta sortuko dira, bakoitza 0tik $k$ra izendatuz. Azkenik, i karpeta bakoitzean i zatian dauden dokumentuak *test* bezala gordeko dira, gainontzekoak *train* bezala gordez.

2. **Entrenamendua**. k eredu sortuko dira, i karpeta bakoitzean dauden *train* eta *test* dokumentuak erabiliz. Datuetako *dev* dokumentu berberak erabiliko dira eredu bakoitzaren ikasketa nola doan aztertzeko.

3. **Ebaluazioa**. Behin k ereduak lortu ditugula, bakoitzetik iragarpenak lor daitezke. Hortaz, ebaluatzeko datuetan iragarpenak egingo ditugu i eredu bakoitzarekin. Hala ere, iragarpen bakarra nahi dugu. Horretarako, k iragarpenen artean azken iragarpena erabaki behar da. Bi hurbilpen erabiliko ditugu: aldi gehienetan iragarri dena emaitza izatea (“maximoarekin geratu”), ala iragarri diren aukeren artean, bakoitza emaitza izateko probabilitateak kalkulatu eta horiek kontuan hartuta bat ausaz erabaki (“erruletaren metodoa”).

3.2 Esperimentazioa

Aipatutako esperimentuak KEIEBA sisteman inplementatu dira baina DisCoDisCo sistema originalaren ereduko hiperparametro berberak erabiliko dira, emaitzak konparatu ahal izateko.

100 zikloz entrenatuko da eredu bakoitza, ziklo bakoitzean 634ko datu multzoa erabiliz. 12 zikloko pazientzia izango dute. Era honetan, entrenamendu batean 12 ziklo jarraituetan ereduko zehaztasuna ez bada hobetzen, entrenamendua geldituko da. Erabiliko den optimizatzailea gradientearen jaitsierarako *adamw* [Loshchilov and Hutter, 2017] da. Ikasketa-tasa $2e-5$ izango da. Hala ere, ikasketa-tasaren balioa jaisten joango da ikasi ahala *learning-rate scheduler*ari esker, $5e-7$ ikasketa-tasak har dezakeen balio minimoa izanik. Honen laburpena 3.2. taulan ikus daiteke.

Hiperparametroa	**Balioa**
Zikloak	100
Datu-multzoa zikloko	634
Pazientzia	12
Optimizatzailea	Adamw
Ikasketa-tasa	$2e-5$
Ikasketa-tasa minimoa	$5e-7$
Learning-rate scheduler	True

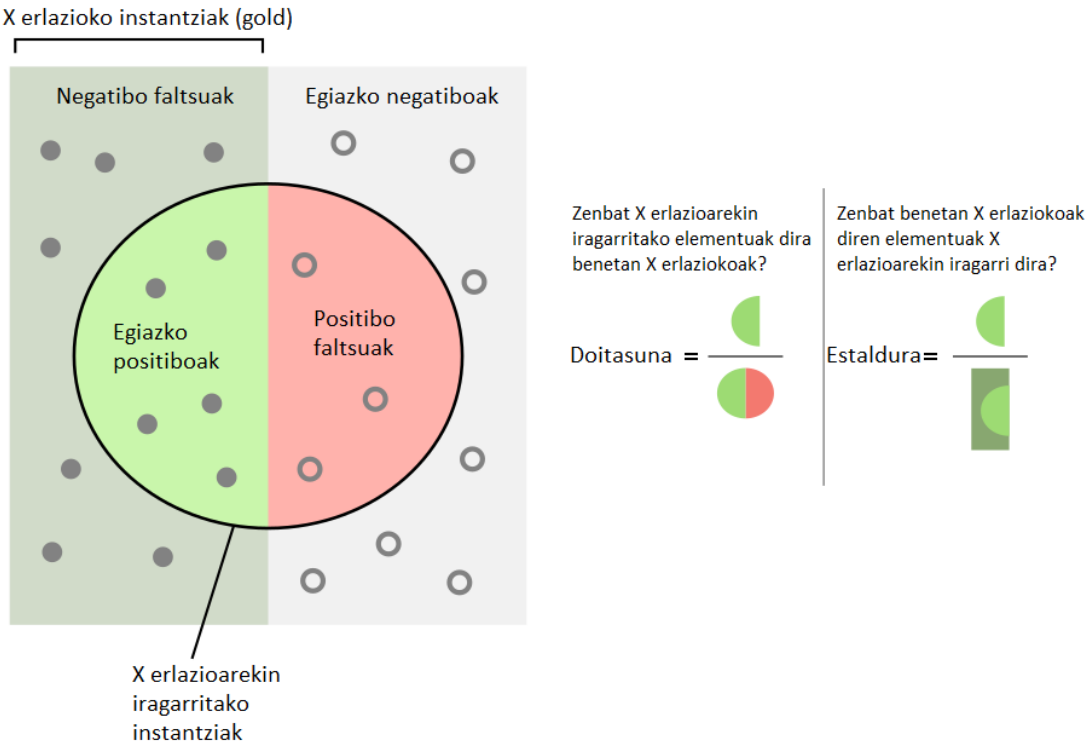
3.2. taula: Esperimentazioko hiperparametroak, DisCoDisCon.

Entrenatuko den eredu bakoitza ebaluatzeko eta beraien artean konparaketak egiteko erabiliko den metrika zehaztasuna (*accuracy*) izango da. Ereduak iragarri behar dituen datu kopurutik zenbat iragarri dituen ondo adierazten du zehaztasunak. Hau da, *test*ean agertzen diren instantzia bakoitzeko erlazioa (benetakoa, *gold*a) eta *test*eko instantzia bakoitzari iragarritako erlazioa zenbat instantzietan berdina izan den. Zehaztapena metrika erabiliko dugu entrenamendu bakoitzeko eredu hoberenarekin geratzeko. Horretarako ziklo bakoitzean uneko eredua ebaluatzean zehaztasun handiagoa lortzen badu gordeko dugu aurretik gordetako hoberena ezabatuz.

Dena den, ez da erabiliko dugun metrika bakarra. Erlazioen zehaztasuna F1-puntuazio metrikarekin kalkulatuko da. Metrika hau doitasuna (*precision*) eta estalduraren (*recall*) arteko bataz besteko harmonikoa da [Goutte and Gaussier, 2005]. Doitasuna hautemandako elementu guztietan zuzen hautemandako elementuen proportzioa da eta estaldura hauteman beharreko elementu guztietatik hautemandako elementuen proportzioa da.

Definizioz, X erlazioarekin iragarritako instantziak *gold*ean ere X erlaziokoak direnak egiazko positiboak (TP) dira. *Gold*ean beste erlaziokoak badira, positibo faltsuak dira (FP). Aldiz, X erlazioarekin iragarri ez diren instantziak *gold*ean ere X erlaziokoak ez badira, egiazko negatiboak (TN) izango dira. *Gold*ean X erlaziokoak badira, negatibo faltsuak (FN) dira. Hau jakinik, estaldura eta doitasunaren eskema 3.2. irudian azter daiteke, eta F1-puntuazioaren formula 3.1. ekuazioan.

$$
\begin{aligned}
Doitasuna &= \frac{TP}{TP+FP} \\
Estaldura &= \frac{TP}{TP+FN} \\
\text{F1}-\text{puntuazioa} = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura} &= \frac{TP}{TP+\frac{FP+FN}{2}}
\end{aligned}
\tag{3.1}
$$



3.2. irudia: Doitasuna eta estalduraren diagrama.

4. KAPITULUA

Emaitzak

Kapitulu honetan jatorrizko sistemaren emaitzak (4.1. atala) eta proposatutako ideiekin sortutako sistemarekin (KEIEBA) lortu diren emaitzak aurkeztuko dira. Lehendabizi, sisteman gehitutako atributuekin entrenatutako ereduen emaitzak aztertuko ditugu (4.2. atala); ondoren, BERT eredu desberdinekin entrenatutako ereduak (4.3. atala); horren ostean erlazioak multzokatuta erabilita lortutako emaitzak (4.4. atala); eta azkenik *cross-validation* teknika erabiliz lortzen diren emaitzak (4.5. atala).

4.1 Jatorrizko sistema

Hobekuntza proposamenekin lortutako emaitzak aztertu baino lehen, egileek DisCoDisCorekin lortutako emaitzak eta gure saiakerarekin (hau da, DisCoDisCo gure kabuz exekutatuz edota KEIEBA sistema hobekuntza barik) lortutako emaitzak aztertu ditugu. Atributu barik eta 2. unitateko *position* atributua erabilita konparatuko dira. Zehaztasunak 4.1. taulan ikus daitezke.

Ezaugarriak	**Zehaztasuna (egileena)**	**Zehaztasuna (gurea)**
DisCoDisCo jatorrizko sistema, atributu barik	% 59,59	% 58,26
DisCoDisCo jatorrizko sistema, 2. unitateko *position* atributuarekin	% 60,62	% 59,29

4.1. taula: Jatorrizko sistemaren zehaztasun emaitzak

Jatorrizko sistema atributu gabeko kasuan egileek % 59,59ko zehaztasuna lortu dute eta gure kasuan % 58,26 lortu dugu, hau da, 1,33 puntu gutxiago. 2. unitateko *position* atributua erabilitako kasuan 1,33 puntu gutxiago ere lortu dugu gure saiakeran, % 59,29 lortuz egileek % 60,62 lortu duten bitartean. Gure kasuan bi saiakera hauek 3 alditan errepikatu ditugu eta hiruretan zehaztasun bera lortu dugu. Gerta daiteke egileek esperimentuak hainbat alditan errepikatzea eta kasuren batean optimizatzaileak minimo lokal baxuago batera heltzea lortzea, ondorioz zehaztasun altuagoa lortuz.

Dena den, argi dago parametroen erabilerak ereduen zehaztasuna handitzen dutela. Izan ere, egileen kasuan 1,03 puntuz hobetzen da, zehaztasuna % 59,59tik % 60,62ra igoz; eta gure kasuan hobekuntza 1,03 puntukoa da, zehaztasuna % 58,26tik % 59,29ra igoz. Hori dela eta, 2. unitateko *position* atributua erabiltzen duten ereduak erabiliko ditugu oinarri gisa proposatutako hobekuntzekin lortutako emaitzak konparatzeko. Egileen (DIS deituko diogu) nahiz guk entrenatutakoaren (DIG deituko diogu) emaitzak kontuan hartuko ditugu. Gure saiakera, DIG eredua, izango da oinarrizko KEIEBA sistema eta bertan aipatutako hobekuntzak gehituko ditugu eredu deribatuak lortuz.

4.2 Parametro berriak

Atal honetan KEIEBA sisteman parametro berriak erabilita lortu diren emaitzak aztertuko dira: DisCoDisCo sisteman erabili ez diren jatorrizko atributuak erabilita, testuen generoa adierazten duen atributua erabilita eta POS etiketen informazioa erabilita. Atributu berriren batek zehaztasuna handitzen duen aztertu nahi da, emaitzak hobetzeko asmoz. Emaitzak 4.2. taulan ikus daitezke.

Eredua	**Ezaugarriak**	**Zehaztasuna**
DIS	DisCoDisCo, 2. unitateko *position* atributuarekin	% 60,62
DIG	KEIEBA, 2. unitateko *position* atributuarekin	% 59,29
ATR	DisCoDisCo jatorrizko sisteman euskararako erabili ez diren parametroak erabilita (KEIEBA)	% 59,29
ATG	ATRko atributuak + Testuen generoa (KEIEBA)	% 59,29
POS	POS etiketen informazioa erabilita (KEIEBA)	% 58,55
AGP	ATR + ATG + POSeko atributuak (KEIEBA)	% 59,88

4.2. taula: Parametro berriak erabilita lortu diren emaitzak

Emaitzak aztertuta, jatorrizko DIS ereduaren zehaztasuna izan da altuena. Atributu berriak erabiltzen dituzten ATR, ATG, POS eta AGP ereduek ez dute DISren zehaztasuna

gainditu, 1,33, 1,33, 2,07 eta 0,74 puntugatik, hurrenez hurren. Haatik, DISren parekoa den guk entrenatutako DIG ereduarekin konparatzen badugu, ATR eta ATG ereduek DIGren zehaztasun bera lortu dute, % 59,29, POSek % 58,55 lortu du (0,74 puntu gutxiago), eta APGk % 59,88ko zehaztasuna (0,59 puntu gehiago). Dena den, argi dago parametro berrien erabilerak zehaztasuna handitzen laguntzen duela, guk entrenatutako jatorrizko sistemak parametrorik gabe % 58,26ko zehaztasuna lortzen baitu, aipatutako ereduen artean baxuena.

Atributuei dagokionez, generoa atributua erabiltzeak berdin diola dirudi. Izan ere, ATG ereduak ATR ereduko atributu berdinak erabiltzen ditu generoa gehituta, eta zehaztasun bera lortu dute (% 59,29). Beste alde batetik, POS ereduak ATR baino 0,74 puntu gutxiago lortu ditu. Ondorioz, iragarpenean ATR ereduko parametroek POS ereduko parametroak baino gehiago laguntzen dutela esan nahi du.

Hala ere, AGP eredua aztertuta, testuen generoa eta POS etiketen informazioaren gehikuntzak emaitza altuagoa lortzea eragin du, guk entrenatutako ATR, ATG eta POS ereduen zehaztasunarekin konparatuz 0,59, 0,59 eta 1,33 puntuz handituz. Beraz, testuen generoa eta hautatutako POS etiketen informazioa batera erabiltzea baliagarria da emaitzak hobetzeko.

4.3 BERT eredua aldatuta

Guk entrenatu ditugun KEIEBA eredu desberdinen artean, proposatutako atributu berriak batera erabiltzen dituen AGP eredua izan da emaitza hobeak lortu dituena. Horregatik, AGPn erabilitako atributuak erabiliko ditugu beste BERT ereduekin entrenamenduak egiteko. Emaitzak 4.3. taulan ikus daitezke.

Eredua	**Ezaugarriak**	**Zehaztasuna**
DIS	BERTeus, DisCoDisCo jatorrizko sistema eta 2. unitateko *position* atributuarekin	% 60,62
AGP	BERTeus, ATR + ATG + POSeko atributuak (KEIEBA)	% 59,88
REC	RoBERTa-EusCrawl, AGPko atributuak (KEIEBA)	% 39,00
IMB	IXAmBERT, AGPko atributuak (KEIEBA)	% 60,47

4.3. taula: BERT eredu desberdinak erabilita lortu diren emaitzak

BERTeus eredua IXAmBERT ereduarekin ordeztuz, zehaztasuna % 59,88tik % 60,47ra igotzea lortu da, hau da, 0,59 puntu igo da. Posible da IXAmBERT galdera-erantzun atazarako aurreentranetua izateak lagundu izana.

Hala ere, ez dugu zorte bera izan RoBERTa EusCrawl ereduarekin, % 39.00 zehaztasuna lortu duelako. Eredu honek LARGE arkitektura erabiltzen duenez (BERTeusek BASE arkitektura du) espero genuen hizkuntzaren errepresentazio handiago bat izatea eta, ondorioz, emaitza hobeak lortzea. Gainera, RoBERTa beste atazetarako maskaradun-ereduarekin aurreentrenatzea nahikoa da emaitza onak lortzeko. Beraz, badirudi kasu honetarako beharrezkoa dela hurrengo esaldiaren iragarpen atazarako entrenatua izatea zehaztasun altuagoak lortzeko.

4.4 Erlazioak multzokatuta

Nahiz eta oraindik jatorrizko DIS ereduen puntuazioa gainditu ez izan, proposamenei esker gure jatorrizko DIG ereduaren zehaztasun puntuazioak gainditzea lortu dugu. Erlazioak multzokatuz ([Mann and Thompson, 1988]n eta [Iruskieta, 2014]n oinarrituta) zehaztasunak nahiko handitzen dira, hauek 4.4. taulan ikus daitezke.

Eredua	**Ezaugarriak**	**Zehaztasuna**
DIS	Erlazio multzo gabe, DisCoDisCo jatorrizko sistema, BERTeus eta 2. unitateko *position* atributuarekin	% 60,62
AGP	Erlazio multzo gabe, BERTeus, ATR + ATG + POSeko atributuak (KEIEBA)	% 59,88
IMB	Erlazio multzo gabe, IXAmBERT, AGPko atributuak (KEIEBA)	% 60,47
MGE	Erlazio multzokatuta, BERTeus, ATGko atributuak (KEIEBA)	% 63,86
MPO	Erlazio multzokatuta, BERTeus, AGPko atributuak (KEIEBA)	% 65,63

4.4. taula: Erlazio taldeak erabilita lortu diren emaitzak

MGE eta MPO ereduek beste ereduek baino zehaztasun handiagoa lortzen dute, baina kontuan hartu behar da taldekatzeak iragartzeko etiketa gutxiago izatea sortzen duela eta, ondorioz, ikasketa kapazitatea handitzea. Hori dela eta, gainera, ezin ditugu taldekatze eta ez-taldekatze ereduak zehaztasun metrikarekin konparatu jakiteko zein izan litekeen hobeagoa. Hau da, aldi gehiagotan asmatzen duen sistema bat baina erlazio zehatzak ez dakizkiena (taldea iragarriko du), gutxiagotan asmatzen duen sistema bat baina erlazio zehatza iragartzen duena hobea den esatea ez da erraza.

Horregatik, F1-puntuazioa kontuan hartuko dugu, ereduak erlazioka konparatzeko. MPO eta AGP ereduak konparatuko ditugu, biek BERT berdina erabiltzen dutelako. Horre-

la, ziur egon gaitezke desberdintasunak ez dituela beste BERT eredu batek eragin. AGP ereduko erlazioen F1-puntuazioa MPO ereduan erlazio bakoitzari dagokion taldearen F1-puntuazioarekin konparatuko da. Era honetan, taldea ala erlazio solteak izatea iragarpenerako onuragarriagoa den jakingo dugu. Emaitzak 4.5. taulan azter daitezke. Berdez MPOk AGP baino 0.1 puntu baino gehiagoko desberdintasuna lortu badu koloreztatuko da, 0.1 puntu baino gutxiago urdinez, 0 puntu baino gutxiago horiz eta -0.1 puntu baino gutxiago gorriz.

MPO	<- F1	Desb.	F1 ->	AGP
Aurkakotasun Taldea	0,64	0,47	0,17	Antitesia
		0,08	0,56	Kontzesioa
		0,26	0,38	Kontrastea
Kausa Taldea	0,61	0,14	0,47	Kausa
		0,21	0,4	Interpretazioa
		0,26	0,35	Justifikazioa
		0,61	0	Motibazioa
		0,13	0,48	Ondorioa
Baldintza Taldea	0,71	0,01	0,7	Baldintza
		0,7	0	Ez-baldintzatzailea
Bateratze Taldea	0,57	0,24	0,33	Konjuntzioa
		0,57	0	Bateratzea
		0,03	0,54	Lista
Elaborazioa Taldea	0,68	-0,01	0,69	Elaborazioa
		0,18	0,5	Ebidentzia
Birformulazioa Taldea	0,35	0,11	0,24	Birformulazioa
		0,35	0	Laburpena
Testuingurua	0,53	0,01	0,52	Testuingurua
Zirkunstantzia	0,66	0,05	0,61	Zirkunstantzia
Disjuntzioa	0,86	-0,14	1	Disjuntzioa
Ebaluazioa	0,44	0,01	0,43	Ebaluazioa
Metodoa	0,66	0,02	0,64	Metodoa
Prestatzea	0,86	0,02	0,84	Prestatzea
Helburua	0,85	0,04	0,81	Helburua
Sekuentzia	0,46	-0,02	0,48	Sekuentzia
Arazo-soluzioa	0,18	0,18	0	Arazo-soluzioa

4.5. taula: F1-puntuazioen arteko diferentziak erlazio bakunetako ereduaren eta taldekatutako erlazio ereduaren artean.

Desberdintasunak aztertuta, orokorrean taldekatzeak erlazioen F1 puntuazioa igo du, iragarpenak hobetuz. Hala ere, taldeen kasuan F1 puntuazioaren igoera askoz nabariagoa da. Beraz, alde batetik, sortutako taldeak egokiak direla ondorioztatu dezakegu. Beste

alde batetik, etiketa gutxiago izateak iragarpenean hobekuntza oso gutxi eragiten duela (talderik gabeko erlazioak ikusita, *disjuntzioa* eta *arazo-soluzioa* izan ezik besteetan 0,05 puntu baino gutxiagoko desberdintasuna dago) esan dezakegu, eta multzo bereko kasuek elkar laguntzea dela zehaztasuna igo duena.

Talderik gabeko erlazioek pixka bat hobetzen dute (0,05 baino gutxiago), *disjuntzioa* eta *arazo-soluzioa* salbu. *Disjuntzioa*ren kasuan F1-puntuazioak altu jarraitzen du MPO ereduan (0,86), eta AGPn 1 lortzen zuen. Honek esan nahi du AGPn sistema erlazio honekin ez dela nahasten, hau da, ez dagoela ez positibo faltsurik, ez negatibo faltsurik erlazio honekin. *Arazo-soluzio*ko kasuan, ordea, AGPko 0tik MPOko 0,18ra igotzea lortu da. Puntuazio baxua bada ere, 0 ez izateak gutxienez erlazio honetako egiazko positiboak egon direla berresten du.

Taldeen kasuan, hobekuntza nabaria *Birformulazio Taldea*rekin lortu dugu. Bi erlazioz eratuta dago: *birformulazioa* eta *laburpena*, 0,24 eta 0 puntuko F1 lortu dutenak AGP ereduan. Talde honek MPO ereduan 0,35 puntu lortu ditu. Beraz, bi erlazioen informazioa baliatuz eta talde bezala erabiliz iragarpenak hobetzea lortu da. Hobekuntza garrantzitsua ikusten dugu, egia da ez dela izan oso altua hobekuntza baina puntu baxuak zituzten erlazioekin izan da, iragarpena hobetuz.

*Baldintza Talde*ko hobekuntza, ordea, ez da izan oso interesgarria. Taldeak 0,71ko puntuazioa lortu du MPOn, osatzen duten erlazioak 0,7 eta 0 puntu izanik AGPn. Taldea 0,01 puntuz hobetu da *baldintza* eta *ez-baldintzatzailea* kasuak multzokatuz. Hau da, pisu gehienak *baldintza* du eta *ez-baldintzatzailea*rekin konbinatuz taldea bakarrik 0,01 igotzen da. Azken finean, *ez-baldintzatzailea* erlazioak taldean ez du taldeko emaitza ia hobetzen, eta multzo gabe erabilita ez da erlazio honetako egiazko kasu bat ere ondo iragartzen. Laburbilduz, taldea erabiltzea ala taldeko erlazioak solte erabiltzeak ez du emaitzetan desberdintasunik eragingo.

*Elaborazioa Taldea*rekin 0,68 puntu lortu dira MPOn, taldeko *elaborazioa* eta *ebidentzia* erlazioek 0,69 eta 0,5 puntu lortuz. Kasu honetan taldeak 0,01 puntu gutxiago lortu ditu *elaborazioa* solte erabilita baino, baina 0,18 puntu gehiago *ebidentzia* solte erabilita baino. Kasu honetan agian hobe litzateke erlazio solteak erabiltzea talde erlazioa baino. Desberdintasunak ikusita interesgarriagoa izan daiteke erlazio solteak erabiltzea, horrela iragarpenean erlazio zehatzak lortuko genituzke, talde izen bat lortu ordez.

Aldiz, *Aurkakotasun Taldea*, *Kausa Taldea* eta *Bateratze Taldea* hirukoteak multzoko erlazio F1 puntuazio handiagoa lortu dute solte erabilita baino, gehienetan 0,1 puntu hobetuz.

4.5 Balioztatze-gurutzatu teknika erabiliz

Balioztatze-gurutzatua aurretik aipatutako eredu konfigurazio berdinekin aplikatu da, teknika honek ekar dezakeen zehaztasun hobekuntza aztertzeko. Emaitzak 4.6. taulan azter daitezke. N zutabean aurreko ereduetan lortutako zehaztasuna dago, eta BG zutabean eredua balioztatze-gurutzatua aplikatuta entrenatuta lortzen den zehaztasuna dago.

Eredua	**Ezaugarriak**	**Z. N**	**Z. BG**
DIS	DisCoDisCo jatorrizko sistema, BERTeus eta 2. unitateko *position* atributuarekin. Balioztatze-gurutzatuaren emaitza metodoa: maximoarekin geratu	% 60,62	% 62,24
DIS	DisCoDisCo jatorrizko sistema, BERTeus eta 2. unitateko *position* atributuarekin. Balioztatze-gurutzatuaren emaitza metodoa: erruletaren metodoa	% 60,62	% 54,42
ATG	DisCoDisCo jatorrizko sisteman euskararako erabili ez diren parametroak erabilita, testuen generoa gehituta, BERTeus (KEIEBA)	% 59,29	% 63,27
AGP	POS etiketen informazioa erabilita eta ATGko parametroak, BERTeus (KEIEBA)	% 59,88	% 62,24
IMB	AGPko atributuak, IXAmBERT (KEIEBA)	% 60,47	% 62,54
MPO	Erlazio multzokatuta, BERTeus, AGPko atributuak (KEIEBA)	% 65,63	% 67,55

4.6. taula: Ereduen zehaztasuna balioztatze-gurutzatua erabiliz

Hasteko, balioztatze-gurutzatuan inplementatutako bi iragarpen motak probatuko ditugu: “maximoarekin geratu” eta “erruletaren metodoa”. Horretarako jatorrizko DIS konfigurazioarekin bi eredu entrenatuko ditugu, ikusteko zein balioztatze-gurutzatuko aukera egokiagoa izan litekeen erlazioen iragarpenak hobetzeko. Balioztatze-gurutzaturik gabe % 60.62ko zehaztasuna du ereduak, erruletaren metodoa aplikatuta % 54.42 eta maximoarekin geratuz % 62.24 zehaztasuna.

Maximoarekin geratuz 1.62 puntu hobetzen dugu, erruletaren metodoarekin aldiz 6.2 puntu okertu da. Honek ez du esan nahi maximoarekin geratuz erruletaren metodoa aplikatu beharrean beti emaitza hobeak lortuko direnik. Izan ere, emandako emaitzen araberakoa izango da eta, hori dela eta, baliteke beste problema batean erruletaren metodoak errendimendu hobea izatea. Kasu honetarako maximoarekin geratuz emaitza hobea lortzen da, eta hori da proposatutako ereduei aplikatuko zaion metodoa iragarpenak lortzeko balioztatze-gurutzatuarekin entrenatzean.

Balioztatze-gurutzatuak ere atributu berriak erabilita zehaztasuna hobetzea lortzen du. ATG ereduan % 59.29tik % 63.27rako hobekuntza lortzen da, 3.98 puntuko igoerarekin. AGP ereduak gutxiago hobetzen du, 2.23 puntuko igoera % 59.88tik % 62.24ra. Fenomemo bera dugu IMB ereduarekin (IXAmBERT erabiltzen duena) eta MPO ereduarekin, erlazio multzokatuekin entrenatu dena, 2.07 puntu (% 60.47tik % 62.54ra) eta 1.92 puntu (% 65.63tik % 67.55ra) hobetuz, hurrenez hurren.

Argi dago balioztatze-gurutzatuaren proposamenarekin emaitzak hobetzen direla, izan ere eredu konfigurazio guztietan igoera lortu baita. Gainera, jatorrizko DIS ereduko zehaztasuna gainditzea lortu da balioztatze-gurutzatua maximoarekin geratuz erabilitako eredu konfigurazio guztietan.

ATG (BG erabilita)	<- F1	Desb.	F1 ->	ATG (BG gabe)
Antitesia	0,44	0,11	0,33	Antitesia
Kontzesioa	0,7	0,09	0,61	Kontzesioa
Kontrastea	0,26	-0,13	0,39	Kontrastea
Kausa	0,47	0,04	0,43	Kausa
Interpretazioa	0,45	0,2	0,25	Interpretazioa
Justifikazioa	0,47	0,16	0,31	Justifikazioa
Motibazioa	0	0	0	Motibazioa
Ondorioa	0,51	-0,01	0,52	Ondorioa
Baldintza	0,8	0,05	0,75	Baldintza
Ez-baldintzatzailea	0	0	0	Ez-baldintzatzailea
Konjuntzioa	0,46	0,05	0,41	Konjuntzioa
Bateratzea	0	0	0	Bateratzea
Lista	0,57	0,11	0,46	Lista
Elaborazioa	0,69	0,03	0,66	Elaborazioa
Ebidentzia	0,36	0	0,36	Ebidentzia
Birformulazioa	0,25	-0,04	0,29	Birformulazioa
Laburpena	0	0	0	Laburpena
Testuingurua	0,56	-0,06	0,62	Testuingurua
Zirkunstantzia	0,69	0,07	0,62	Zirkunstantzia
Disjuntzioa	1	0,5	0,5	Disjuntzioa
Ebaluazioa	0,47	-0,04	0,51	Ebaluazioa
Metodoa	0,66	0,08	0,58	Metodoa
Prestatzea	0,86	0	0,86	Prestatzea
Helburua	0,84	0,05	0,79	Helburua
Sekuentzia	0,5	0,04	0,46	Sekuentzia
Arazo-soluzioa	0,2	0,2	0	Arazo-soluzioa

4.7. taula: ATG eredu konfigurazioa balioztatze-gurutzatuarekin aplikatuta eta aplikatu gabe erlazio bakoitzak lortutako F1-puntuazioa eta beraien arteko desberdintasuna

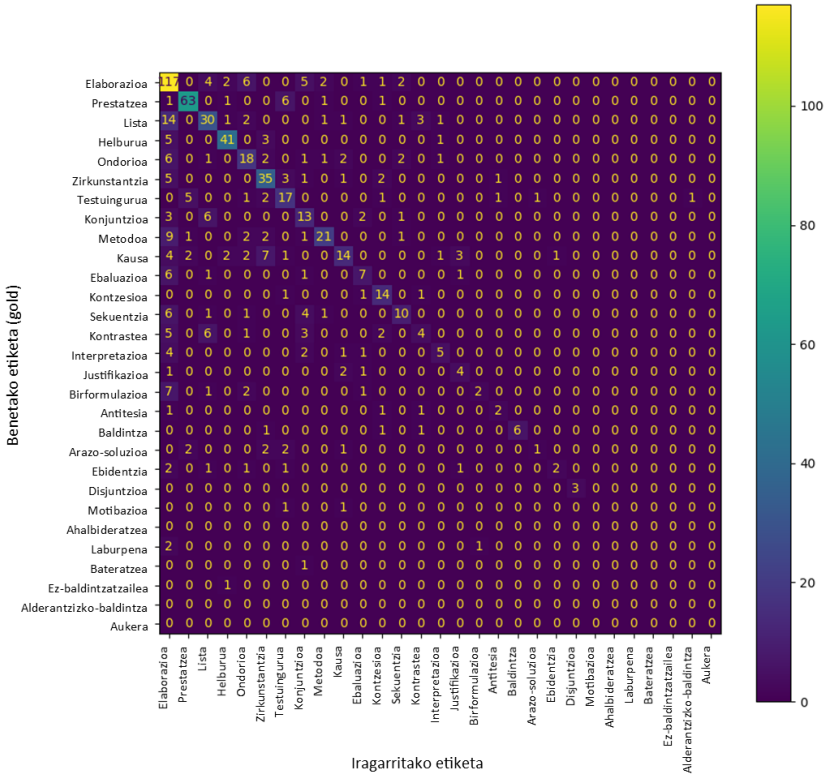
Balioztatze-gurutzatuak erlazio bakoitzean eragiten duen onura ala kaltea F1-puntuazioarekin azter daiteke 4.7. taulan. Konparaketa ATG ereduarekin gauzatu da, gehien hobetu dena delako. 0ko F1-puntuazio lortzen zuten 5 erlazioen artean bakarrik *arazo-soluzioa* erlazioaren puntuazioa igo da balioztatze-gurutzatuarekin. Bakarrik 5 erlaziok emaitza baxuagoak lortu dute (*kontrastea*, *ondorioa*, *birformulazioa*, *testuingurua* eta *ebaluazioa*, horiz eta gorriz koloreztatuak). 2 erlazioetan (*ebidentzia* eta *prestatzea*) ez da egon desberdintasunik. Gainontzeko 14 erlazioetan hobekuntza egon da. 0.5 baino gutxiagoko F1 lortzen zuten 5 erlazioek (*antitesia*, *interpretazioa*, *justifikazioa*, *lista* eta *disjuntzioa*, berdez koloreztatuak) 0.1 puntu baino gehiago hobetzea lortu du balioztatze-gurutzatuak. Beste 9 erlazioek (urdinez koloreztatuak), orokorrean, 0.5 puntu baino gehiago lortzen zuten eta balioztatze-gurutzatuarekin 0.1 puntu baino gutxiago hobetzea lortu dute.

Amaitzeko, sisteman iragarpenak lortzeko bai ATG aukera ala MPO aukera erabiliko genuke balioztatze-gurutzatuarekin, zehaztasun handienak lortu dituztenak direlako. ATG ereduarekin iragarpenean erlazio konkretuak lortuko genituzke zehaztasun txikiagorekin, MPO ereduan, ordea, zehaztasun handiagoa lortuko genuke, baina iragarpenean kasu batzutan erlazio taldeak lortuko genituzke erlazio konkretuak jaso ordez. Ikerketaren ala helburu akademikoaren arabera, eredu bat ala bestea erabiliko litzateke. Erabakia errazteko asmoz, ATG eta MPO ereduen konfusio-matrizeak aurkezten ditugu, 4.1. irudian eta 4.2. irudian hurrenez hurren.

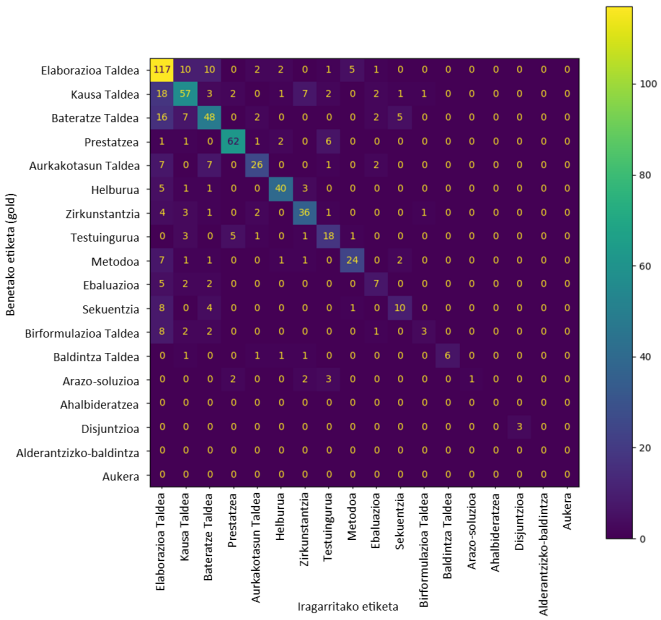
Konfusio-matrizearen interpretazioa nahiko sinplea da. Erlazio bat hautatuta, bere ilarak negatibo faltsuak adierazten ditu, eta bere zutabeak positibo faltsuak. Erlazio baten ilara eta zutabea elkartzen diren posizioan egiazko positiboa adierazten da. Hau jakinik, erlazio bakoitzaren doitasuna eta estaldura kalkula liteke. Ereduaren zehaztasuna diagonalaren batura instantzia kopuruarekiko da.

Taldekatu gabeko erlazioak aztertuta, *zirkunstantzia*, *metodoa* eta *testuingurua* erlazioek egiazko positibo gehiago lortu dituzte MPO ereduan ATG ereduarekin alderatuz, 1, 3 eta 1 hurrenez hurren. *Prestatzea* eta *helburua* erlazioek egiazko positibo 1 gutxiago izan dute MPOn, eta *ebaluazioa*, *sekuentzia* eta *arazo-soluzioa* egiazko positibo kopuru bera lortu dute bi ereduetan. Beraz, taldekatu gabeko erlazioetan antzeko zehaztasuna lortuko da bi ereduetan.

Multzokatutako erlazioen kasuetan, bi ereduetan neurri batean ala bestean iragarpenetan fenomeno antzekoak aurki daitezke. Adibidez, MPOn *Bateratze Taldea*k negatibo faltsu gehienak ditu *Elaborazio Taldea*rekin. ATGn *Bateratze Talde*ko *lista* eta *konjuntzioa* negatibo faltsu gehienak *elaborazioa* erlazioarekin ditu.



4.1. irudia: ATG (balioztatze-gurutzatuarekin) ereduaren konfusio-matrizea.



4.2. irudia: MPO (balioztatze-gurutzatuarekin) ereduaren konfusio-matrizea.

Esan bezala, ikerketaren ala helburu akademikoaren arabera, MPO edo ATG eredua aukeratuko da. Erlazio konkretuetan interesa izatekotan, erlazio horiek egiazko positibo gehien dituen eredua hauta liteke. Adibidez, *helburua* erlazioan interesa izatekotan ATGrekin geratuko ginateke. Hala ere, erlazio solteetan egiazko positibo desberdintasuna hain baxua izanik, bai MPO bai ATG ereduak erabil litezke, erlazio horietan antzeko zehaztasuna espero baita. Erlazio batean negatibo faltsuak interesatzen ez bazaizkigu, erlazio horietan gutxien dituen eredua aukeratuko genuke. Adibidez, *Bateratze Taldea*k 32 negatibo faltsu ditu MPOn, eta ATGn taldeko *lista*, *konjuntzioa* eta *bateratzea*, 24, 12 eta 1 hurrenez hurren, guztira 37 izanik. Erlazio horietan negatibo faltsu gutxien iragarri nahi izanez gero, taldearekin geratuko ginateke, hots, MPO eredua.

5. KAPITULUA

Ondorioak

5.1 Proiektuaren Ondorioak

Proiektu honetan, RST teoriako diskurtso erlazioak eta sare neuronalen teoria azaldu dugu, hauekin erlazioak iragartzen dituzten garatutako sistemak aipatuz. DisCoDisCo sistema deskribatu dugu, eta emaitzak hobetzeko proposamenak aurkeztu gure KEIEBA sistemarekin. Azkenik, proposamenekin esperimentazioa gauzatu da eta emaitzak bildu ditugu.

Ez dago zalantzarik proiektuaren helburua lortu egin dela. Izan ere, jatorrizko DisCoDisCo sistemaren emaitza hobetu dugu hainbat KEIEBA eredutan. Horien artean ATG (balioztatze-gurutzatuarekin) eredua dugu, % 60,62 zehaztasunetik % 63,27 zehaztasunera arte igoz. Gainera, beste BERT ereduak erabilita eta erlazioak multzokatuz ere zehaztasun handiagoa lortu dugu, balioztatze gurutzatuaren teknikaz baliatuz.

Hala ere, lan honen ekarpena ez da soilik izan sistemaren emaitzak hobetzea, baizik eta hobekuntza proposamenekin esperimentazioa gauzatzea. Era honetan, emaitzen zehaztasun aldaketa aztertu dugu eta zein aukerak gehien laguntzen duen ikusi dugu.

5.2 Etorkizuneko lana

Atributu gehiago erabiltzea

Emaitzak hobetzeko esperimentu gehiago egin litezke. Horien artean, atributu gehiago erabiltzea da asmoa. Aditzetako informazio gehiago erabil daiteke, ala aditzekin egin den bezalaxe, izenetako informazioa eskuratzea. Informazioaz gain, atributu berriak gehitu daitezke, kontaketa kopuruak izan daitezkeenak, adibidez lokailu kopurua. Beste alde batetik, testu osotik atributu orokorrak ere lor daitezke. Adibidez, testuari sentimenduen analisia aplikatu jakiteko segmentua positiboa, negatiboa ala neutrala den, eta hori atributu bezala erabili.

Instantzien spanak osatzea

2.3.3. atalean aipatu dugun bezala, corpusak ez du RST zuhaitz egitura eta sakonera mantentzen. Horregatik, corpuseko instantzietan erlazioak lotzen dituen segmentuetan ez da testu osoa agertzen, baizik eta *span* horretako nukleo den sakonerako EDUko testua. Modu honetara erabili ordez, instantzietako segmentuetan *span* hori osatzen duen testu osoa jar daiteke. Horrela, sistemak testu osoa izango luke eta gerta liteke testu gehiagorekin BERTek testuinguru gehiago izatea bi segmentuen arteko erlazioa ikasteko. Honetarako, corpus originala eskuratu behar da, RST egitura osoa izateko eta *rels* fitxategian segmentu bakoitzaren testu osoa jarri.

Zuhaitza lortu irudi bezala

Aurreko proposamena gauzatzen bada, RST egitura osoa izango genuke *rels* fitxategian, eta bertatik zuhaitzaren irudi bat lortzeko aukera izango dugu, *rs3* formatura pasatzen duen programa bat sortuz. Era honetan, era bisualean aurkeztu daiteke zuhaitza iragarri diren erlazioekin, RSTTool [O'Donnell, 2000] tresna erabiliz.

Eranskinak

A. ERANSKINA

RST erlazioen taulak

Hona hemen euskarazko RST erlazioen arauak eta efektuak, [Iruskieta, 2014]tik aterata.

A.1 Aurkezpenezko erlazioak euskaraz

Aurkezpenezko erlazioen definizioak			
Erlazioa	**Arauak Sn eta Nn**	**Arauak S-Nn**	**Efektua**
Antitesia	N-n: idazleak N-rekiko aldeko iritzia du	N eta S aurkaritzako erlazioan daude eta bateraezinak dira; beraz, ezinezkoa da biekiko (N eta S) aldeko iritzia edukitzea. S-ren aldeko iritzia izatean, ezin da N-ren aldeko iritzia izan. Aurkaritza erlazio horretan, irakurleak N-rekiko duen aldeko iritzia handitzen du	Bateraezinak diren bi egoeren aurrean irakurlearen iritzi positiboa handitzen da N-rekiko

Testuingurua	N-n: S irakurri arte irakurleak ez du N ulertuko guztiz	S irakurtzeak N edo N-ko elementuren bat hobeto ulertzea dakar	N hobeto ulertzeko irakurlearen aldeko iritzia handitzen da
Kontzesioa	N-n: idazleak N-rekiko aldeko iritzia du; S-n: idazleak ez du erakusten S onartzen ez duenik	Idazlearentzat N eta S onargarriak izan badaitezke ere, bien artean aurkakotasunak egon daitezkeela onartzen du; irakurleak aurkakotasunezko egoera hori onartzean, N-rekiko aldeko iritzia handitzen du	N eta S onargarriak izan arren, aurkakotasunak daude bi egoeretan. Aurkakotasun horrek irakurlearengan N-rekiko aldeko iritzia handitzea dakar

Ahalbideratzea	N-n: gauzatu gabeko ekintza bat aurkezten da N-n irakurleari zuzendua (eskaintza baten onarpena barnean duela); S-n: Irakurleak S ulertzean N-n aurkezten den gauzatu gabeko ekintza gauzatzeko aldeko iritzia handitzen du	Proposaturiko ekintza gauzatzeko irakurlearen aldeko iritzia handitzen da	N gauzatu gabe dago eta S-k N gauzatzeko modua erakusten du
Evidentzia	N-n: gerta liteke irakurleak ez izatea froga nahikorik N-rekiko aldeko iritzia izateko; S-n: irakurleak S-rekiko aldeko iritzia du.	Irakurleak S ulertzean N-rekiko aldeko iritzia handitzen du	Irakurleak S-n ditu frogak N sinesteko edo N-rekiko aldeko iritzia handitzeko
Justifikazioa	Ez dago baldintzarik	Irakurleak S ulertzean, idazleak N aurkezteko egokitasuna areagotzen da	Irakurleak idazleari N aurkezteko egokitasuna onartzen dio

Motibazioa	N-n: N gauzatu gabeko ekintza da eta irakurlea ekintzaren egilea (eskaintza baten onarpena ere badago)	S ulertzeak irakurleari N-n proposatu zaion ekintza egiteko aldeko iritzia edo nahia handitzen dio	N-n proposaturiko ekintza egiteko irakurlearen gogoa handitzen du
Prestatzea	Ez dago baldintzarik	Testuan S dago N-ren aurretik; S-rekin irakurleari N-n dagoen informazioa aurreratzen edo interesa pizten dio idazleak	Irakurleari N irakurtzeko, interesa, prestutasuna edota orientazioa handitzen zaio
Birformulazioa	Ez dago baldintzarik	S-rekin N-n dagoena beste era batera formulatzen da. Testuaren tamainari dagokionez, S eta N tamaina berekoak dira; baina nukleartasunari dagokionez, N idazlearentzat S baino garrantzitsuagoa da	Irakurleak S N-ren bestelako formulaziotzat hartzen du
Laburpena	N-n:unitate bat baino gehiagoz osatuta egon behar da	N-n idatzitakoaren sintesia agertzen da S-n; beraz, S-ko informazioa N-koa baino laburragoa da	Irakurleak S-n dagoena N-n dagoenaren laburpena dela onartzen du

A.1. taula: Euskarazko aurkezpenezko erlazioen arauak eta efektuak

A.2 Edukizko erlazioak euskaraz

Edukizko erlazioen definizioak			
Erlazioa	Arauak Sn eta Nn	Arauak S-Nn	Efektua
Zirkunstantzia	S-n: S gauzatuta dago	Irakurleak S-n deskribatutako zirkunstantzietan interpretatu behar du N	N interpretatzeko zirkunstantzia S-k ematen diola onartzen du irakurleak
Baldintza	S-n: S egoera hipotetikoa, etorkizunekoa edo gauzatu gabekoa da	N gauzatuko da, baldin eta S gauzatzen bada	N S-k baldintzatzen duela onartzen du irakurleak

Elaborazioa	Ez dago baldintzarik	N-n aurkeztutako gaiaren edo egoeraren ezaugarriren bat garatzen da S-n edo N-tiko inferentzia aurkezten da S-n, erlazio hauen arabera: multzoa :: kidea; abstraktua :: adibidea; osoa :: zatia; prozesua :: urratsa; objektua :: atributua; orokorra :: espezifikoa	S-n aurkeztutako egoerak N-ko ezaugarriren bat garatzen duela onartzen du irakurleak. Irakurleak garatutako elementua edo gaia identifikatzen du
Ebaluazioa	Ez dago baldintzarik	Idazleak N-rekiko duen aldeko iritzia aurkezten du S-k	S-n N ebaluatzen dela onartzen du irakurleak
Interpretazioa	Ez dago baldintzarik	Idazleak N-n ez dauden ideiak erlazionatzen ditu S-rekin	N-n ez dauden ideia-multzoak S-rekin erlazioa duela onartzen du irakurleak
Metodoa	N aktibitatea da	S-n metodoa edo instrumentua aurkezten da, zeinaren bidez N egikaritzen den	Irakurleak onartzen du S-n aurkezturiko metodoak edo instrumentuak N posible egiten duela

Kausa	N-n: N-n arrazoia aurkezten da	N gauzatzeko arrazoia S-n agertzen da; S aurkeztu gabe irakurleak ezingo luke jakin zergatik gertatu den N; N S baino garrantzitsuagoa da idazlearen helburuetarako	N-n gertatzen den egoeraren kausa S dela onartzen du irakurleak
Ondorioa	S-n: S-n ekintza edo egoera sortu den arrazoia aurkezten da	N-k eragin zezakeen S; S baino garrantzitsuagoa da N idazlearen helburuetarako	Irakurleak onartzen du N dela S-ren kausa edo S dela N-ren ondorioa
Aukera	N-n: N gauzatu gabeko egoera da; S-n: S gauzatu gabeko egoera da.	N gauzatzeak S gauzatzea galarazten du	N-ren gauzatzeak S-ren ez gauzatzearekin duen mendeko erlazioa onartzen du irakurleak
Helburua	N-n: N aktibitatea da; S-n: S gauzatu ez den egoera da	N-ko aktibitatearekin gauzatuko da S	N-ko aktibitatea S gauzatzeko egin dela onartzen du irakurleak

Arazo-soluzioa	S-n: S-n arazoa aurkezten da	N da S-n aurkezturiko arazoaren soluzioa	S-n aurkezturiko arazoaren soluzioa N dela onartzen du irakurleak
Ez-baldintzatzailea	S-n: baliteke S-k N-ren egikaritzean eragitea	S-k ez du N baldintzatzen	Irakurleak onartzen du N ez duela S-k baldintzatzen
Alderantzizko Baldintza	Ez dago baldintzarik	N gauzatzen da baldin eta ez bada S gauzatzen	N gauzatuko dela onartzen du irakurleak baldin eta bakarrik S ez bada gauzatzen

A.2. taula: Euskarazko edukizko erlazioen arauak eta efektuak

A.3 Multinuklear erlazioak euskaraz

Erlazio nukleoaniztunen definizioak		
Erlazioa	Arauak nukleo bakoitzean	Efektua
Konjuntzioa	N guztiek osotasun bat osatzen dute. Osotasun horretan N baten rola beste N guztiekin konparagarria da.	N-ek osotasun bat osatzen dutela eta elkarren artean erlazionaturik daudela ezagutzen du irakurleak

Kontrastea	Bi N daude, ez gehiago; N horien arteko erlazioak honakoak izan daitezke: a) antzekotzat ulertzen dira zenbait ezaugarritan, b) ezberdintzat ulertzen dira zenbait ezaugarritan, c) ezberdintasuna da beste antzeko ezaugarriekin konparatzen dena.	Aurkaritzazko konparazioaren berdintasunak eta ezberdintasunak onartzen ditu irakurleak
Disjuntzioa	N bat beste(ar)en alternatibatzat (ez derrigorrez esklusiboa) aurkezten da	N guztiak alternatiboak direla onartzen du irakurleak
Bateratzea	Ez dago baldintzarik	Ez dago baldintzarik
Lista	N guztiek elkarren artean ezaugarriren bat konpartitzen dute eta, gainera, N guztiek zerrenda bat osatzen dute	Zerrenda bateko elementuak direla ezagutzen du irakurleak
Birformulazio nukleoaniztuna	N bat berregiten da beste N batekin eta idazlearen helburuetarako bi N horien garrantzia maila berekoa da	Berregindako N-ek garrantzia bera dutela ezagutzen du irakurleak
Sekuentzia	N guztien artean segida erlazioa dago	N guztien arteko segida erlazioa ezagutzen du irakurleak

A.3. taula: Euskarazko multinuklear erlazioen arauak eta efektuak

Bibliografia

[Agerri et al., 2020] Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

[Alkorta et al., 2019] Alkorta, J., Gojenola, K., and Iruskieta, M. (2019). Towards discourse annotation and sentiment analysis of the Basque opinion corpus. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 144–152, Minneapolis, MN. Association for Computational Linguistics.

[Aranzabe et al., 2004] Aranzabe, M. J., Arriola, J. M., and de Ilarraza, A. D. (2004). Towards a dependency parser for basque. In *Proceedings of the Workshop on Recent Advances in Dependency Grammar*, pages 41–48.

[Artetxe et al., 2022] Artetxe, M., Aldabe, I., Agerri, R., Perez-de Viñaspre, O., and Soroa, A. (2022). Does corpus quality really matter for low-resource languages?

[Atutxa et al., 2019] Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., and Iruskieta, M. (2019). Towards a top-down approach for an automatic discourse analysis for basque: Segmentation and central unit detection tool. *PLOS ONE*, 14(9):1–25.

[Atutxa et al., 2021] Atutxa, U., Molina-Villegas, A., and Iruskieta Quintian, M. (2021). Generación automática de meta-resúmenes para la evaluación del manejo de estructuras discursivas y coherencia en el alumnado. *Procesamiento del Lenguaje Natural*, pages 165–175.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Braud et al., 2017] Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual rst discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*.

[Braud et al., 2016] Braud, C., Plank, B., and Søgaard, A. (2016). Multi-view and multi-task training of RST discourse parsers. In *Conference on Computational Linguistics (CoLing)*, pages 1903 – 1913, Osaka, Japan.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

[Corston-Oliver, 1998] Corston-Oliver, S. (1998). Identifying the linguistic correlates of rhetorical relations. In *Discourse Relations and Discourse Markers*.

[Dargan et al., 2019] Dargan, S., Kumar, M., Ayyagari, M. R., and Kumar, G. (2019). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, pages 1–22.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[duVerle and Prendinger, 2009] duVerle, D. and Prendinger, H. (2009). A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 665–673, Suntec, Singapore. Association for Computational Linguistics.

[Elman, 1990] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

[Feng and Hirst, 2014] Feng, V. W. and Hirst, G. (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

[Gardner et al., 2017] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform. In *arXiv*.

[Gessler et al., 2021] Gessler, L., Behzad, S., Liu, Y. J., Peng, S., Zhu, Y., and Zeldes, A. (2021). DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[Goutte and Gaussier, 2005] Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[Hayashi et al., 2016] Hayashi, K., Hirao, T., and Nagata, M. (2016). Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles. Association for Computational Linguistics.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

[Iruskieta, 2014] Iruskieta, M. (2014). *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean*. PhD thesis, Euskal Herriko Unibertsitatea.

[Iruskieta et al., 2013] Iruskieta, M., Aranzabe, M., Diaz de Ilarraza, A., Gonzalez-Dios, I., Lersundi, M., and Lopez de Lacalle, O. (2013). The rst basque treebank: an online search interface to check rhetorical relations. In *IV Workshop A RST e os Estudos do Texto*, page 40–49, Fortaleza, CE. Sociedade Brasileira de Computaçao. Outubro 21-23.

[Iruskieta et al., 2019] Iruskieta, M., Bengoetxea, K., Atutxa Salazar, A., and Diaz de Ilarraza, A. (2019). Multilingual segmentation based on neural networks and pretrained word embeddings. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 125–132, Minneapolis, MN. Association for Computational Linguistics.

[Iruskieta and Braud, 2019] Iruskieta, M. and Braud, C. (2019). EusDisParser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 62–71, Minneapolis, MN. Association for Computational Linguistics.

[Jain et al., 1996] Jain, A., Mao, J., and Mohiuddin, K. (1996). Artificial neural networks: a tutorial. *Computer*, 29(3):31–44.

[Ji and Eisenstein, 2014] Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

[Joty et al., 2015] Joty, S., Carenini, G., and Ng, R. T. (2015). CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

[Kobayashi et al., 2020] Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., and Nagata, M. (2020). Top-down rst parsing utilizing granularity levels in documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106.

[Kobayashi et al., 2021] Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., and Nagata, M. (2021). Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.

[Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

[Kraus and Feuerriegel, 2017] Kraus, M. and Feuerriegel, S. (2017). Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118.

[Li et al., 2016] Li, Q., Li, T., and Chang, B. (2016). Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas. Association for Computational Linguistics.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization.

[Mahesh, 2020] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.

[Mann and Thompson, 1988] Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8 (3)*, pages 243–281.

[Marcu, 2000] Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Nguyen et al., 2021] Nguyen, T.-T., Nguyen, X.-P., Joty, S., and Li, X. (2021). Rst parsing from scratch. *arXiv preprint arXiv:2105.10861*.

[Nwankpa et al., 2018] Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning.

[O'Donnell, 2000] O'Donnell, M. (2000). Rsttool 2.4: A markup tool for rhetorical structure theory. In *Proceedings of the First International Conference on Natural Language Generation - Volume 14*, INLG '00, page 253–256, USA. Association for Computational Linguistics.

[Otegi et al., 2020] Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 436–442, Marseille, France. European Language Resources Association.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning

library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[Popoola, 2017] Popoola, O. (2017). Using Rhetorical Structure Theory for detection of fake online reviews. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 58–63, Santiago de Compostela, Spain. Association for Computational Linguistics.

[Potdar et al., 2017] Potdar, K., Pardawala, T. S., and Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9.

[Qu et al., 2019] Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

[Rojas, 2013] Rojas, R. (2013). *Neural networks: a systematic introduction*. Springer Science & Business Media.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408.

[Skoufaki, 2020] Skoufaki, S. (2020). Rhetorical structure theory and coherence break identification. *Text and Talk*, 40:99–124.

[Straka and Straková, 2017] Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

[Surdeanu et al., 2015] Surdeanu, M., Hicks, T., and Valenzuela-Escárcega, M. A. (2015). Two practical Rhetorical Structure Theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado. Association for Computational Linguistics.

[Taboada and Mann, 2006a] Taboada, M. and Mann, W. (2006a). Applications of rhetorical structure theory. *Discourse Studies - DISCOURSE STUD*, 8:567–588.

[Taboada and Mann, 2006b] Taboada, M. and Mann, W. (2006b). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies - DISCOURSE STUD*, 8.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.Ñ., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

[Wang and Raj, 2017] Wang, H. and Raj, B. (2017). On the origin of deep learning.

[Wang et al., 2017] Wang, Y., Li, S., and Wang, H. (2017). A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

[Yu et al., 2018] Yu, N., Zhang, M., and Fu, G. (2018). Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Zeldes et al., 2021] Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.