eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Gradu Amaierako Lana

Informatika Ingeniaritzako Gradua

Konputazioa

Entitateen eta erlazioen erauzketa testu klinikoetan datu etiketatu gutxirekin

Josu Loidi Gorostidi

Zuzendariak
Oier Lopez de Lacalle
Aitziber Atutxa Salazar
Koldo Gojenola Galletebeitia

2023ko ekainaren 20a

Esker onak

Eskerrak eman nahi dizkiet Aitziber Atutxa eta Koldo Gojenolari proiektu hau egiteko gai nintzela pentsatzeagatik eta egiteko aukera niri emateagatik.

Honetaz gain, eskerrak eman nahi dizket Aitziber Atutxa, Koldo Gojenola eta baita Oier Lopezi ere proiektu hau egitean eman didaten laguntzagatik eta norabidea galduta nenbilenean berriro ere bidea non zegoen aurkitzen laguntzeagatik.

Gehitzeko, BioDonostiako Ander Alberdik lideratutako mediku taldeari eskerrak eman beharrean nago. Batetik, proiektua haiek proposaturikoa delako. Eta bestetik, datu-multzoa zerotik sortzeko konpromisoa hartu dutelako. Eskertzekoa da beraien bizitza profesionalean duten lan guztia edukita ere, lan karga gehigarri hori bere gain hartzea ikerketaren alde.

Eta nola ez, eskerrak eman behar dizkiet familia eta ingurukoei babesa emateagatik. Baita, egiten ari nintzen lanean interes minimo bat erakutsi duten pertsona guztiei ere. Barkatu zenbait momentutan astunegia izan banaiz.

Laburpena

Osakidetzako larrialdi zerbitzuetan arazo bat topatu da. Gaixo bat bertara heltzen denean, medikuek gaixoak dituen sintomen inguruko galderak egiten dizkiote eta honen erantzunak lengoaia naturala erabiliz testu klinikoetan erregistratzen dituzte. Kontua da, zenbait kasutan galdetu beharreko daturen bat ez dutela galdetzen edota ez dutela behar bezala idatzita uzten. Hau problema bat da, etorkizunean diagnosi okerrak egiteko aukera sortzen baitu.

Pazientea bularraldeko minarekin joango balitz ospitalera, sendagileak bederatzi datu erregistratu beharko lituzke. Honelako kasu batean beharrezkoa da hau egitea, bularraldeko mina gaixotasun hilgarri askoren aurrekaria izan baitaiteke. Hau jakinda, Hizkuntza Naturalaren Prozesamenduko teknikak erabiliz, bularraldeko minarekin erlazionatutako txosten klinikoetan aipatu beharreko datuen presentzia eta absentzia detektatzen duen sistema bat garatu nahi izan da.

Baina, hau garatzeko arazo nagusi bat egon da: datu etiketatuen urritasuna. Ataza hau ikasteko eskuragarri dagoen corpusa oso txikia da. Egoera hau behin baino gehiagotan errepikatzen da medikuntzaren domeinuan. Testu kliniko elektroniko ugari topatu arren, etiketatu gabe egon ohi dira. Hori dela eta, arazoari aurre egin nahi izan zaio. Honelako eszenatoki bat planteatzen den kasuan, alegia, domeinua oso murritza denean (medikuntzaren kasua) eta datu urritasuna nabaria denean, zein bide posible har daitezkeen ikertu da.

Domeinu orokorrean, datu-urritasun eszenarioetan, erlazio erauzketa ataza inferentzia ataza batean bihurtzea lagungarria dela frogatuta dago. Proiektu honen helburuetako bat, medikuntza domeinuan ondorio berdinak mantentzen direla ikustea izan da. Inferentzian oinarritutako eredu batek benetan ataza ikasteko gaitasuna ote duen frogatu da. Baita, datu gutxi edukita ataza birmoldatu gabe lortuko liratekeen ereduak aztertu ere. Bukatzeko, bi estrategiekin lortutako emaitzak alderatu dira.

Eskuraturiko emaitzak interesgarriak izan dira. Ikasketa ataza birmoldatuta, datu etiketatu gutxirekin medikuntza domeinuan sistema ahaltsuak garatzeko aukera dagoela ikusi da. Bestalde, ikasketa ataza birmoldatu gabe, hau da, erlazio erauzketa sailkapen ataza tradizional gisa planteatuta, zenbait erlaziorekin oso instantzia etiketatu gutxi behar direla ikusi da.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. Sarrera

1.1 Motibazioa

Hizkuntza Naturalaren Prozesamendua (HNP) adimen artifizialeko adar garrantzitsu bat da, zeinaren helburua sistema informatikoek gizakion hizkuntza prozesatu eta ulertzea den. Informazio kliniko elektronikoaren zati handi bat testu libre gisa gorderik dago. HNPko teknikei esker, testu hauetatik automatikoki informazioa erauzi eta egituratu daiteke. Informazio egituratu hau aplikazio askotarako erabil daiteke, besteak beste, kodifikaziorako, diagnostikorako, pronostikorako, irakaskuntzarako edota saiakuntza klinikoak maneiatzeko. Aplikazio horietako bat sintomen analisia da.

Pertsona batek larrialdi zerbitzuak bisitatzen dituenean, medikuak gaixoaren informazioa jasotzen du diagnosia burutzeko. Informazio erauzketa hau, anamnesia deritzona, galdera-erantzunen bitartez gertatzen da. Gaixoak medikuari aipaturiko sintomak historia klinikoetan gordetzen dira hizkuntza naturalean idatzirik. Zenbaitetan hizkuntza hori anbiguo eta konplexua suertatzen da. Baina, aldi berean, informazio erabakigarria gordetzen du diagnostikorako, tratamendurako, pronostikorako eta prozesuaren balorazio ekonomikorako. Honenbestez, sintomen analisi zehatzak egiten dituzten sistemak, osasun egoera hobetzen lagun dezakete.

Historikoki, sintomen analisia aditu klinikoen eskuzko berrikuspenaren bidez egin izan da. Halere, eskuragarri dagoen historia kliniko elektroniko kantitatea dela eta, ataza hau era automatiko edo erdiautomatiko batean burutzen duten HNP sistemen bilaketa martxan jarri da. Bide honetan, hainbat lan egin dira [1], zeinetan aztertu nahi izan den posible ote den esklerosi anizkoitza duten gaixoak identifikatzea historia kliniko elektronikoan deskribatutako sintometan oinarrituz, hesteetako gaixotasun inflamatorioaren sintoma iradokitzaileak identifikatuz, edota testuetatik minaren eta intentsitatearen aipamenak ateraz. Horretarako, izendun entitateen, adierazpen tenporalen eta erlazioen erauzketa funtsezkoak dira eta garapenean dagoen ikerketa arlo baten parte dira [2].

Elkarrizketa klinikoetan, sintomen jatorria ezagutu eta gaixotasun larriak baztertzeko helburuarekin, medikuek gaixoei alarma-sintomen inguruan galdetzen die [3]. Ondoren, gaixoek emandako erantzun guztiak lengoaia naturalean idazten dituzte. Hala eta guztiz ere, batzuetan, informazio hau ez da galdetzen edo ez da behar bezala erregistratzen. Honek, etorkizunean diagnosi okerrak egitea edota gaixotasun larriak oharkabean pasatzea eragin dezake. Sintomak ez erregistratzeak ez du esan nahi hauen inguruan galdetu ez denik, eta beharbada medikuak gogoan mantentzen du gaixoak esandako guztia. Baina, zenbait kasutan gaixoen jarraipena luzatu egiten da eta profesional ezberdinen artean egiten da. Horregatik, behar-beharrezkoa da gaixoaren informazio guztia erregistraturik mantentzea.

Proiektu hau, hasiera batean, gaixotasun konkretu batean oinarritu da: bularraldeko mina. Halere, proiektuaren garapena aurrera joan ahala, helburuak aldatuz joan dira eta gaixotasun ezberdinetara hedatu dira esperimentuak. Zehazki, proiektuaren bigarren partea gaixotasun arraroen ezaugarriak erauztera bideratu da.

Bularraldeko mina izaten da larrialdi zerbitzuetara joateko arrazoi errepikatuenetako bat. Kasu hauetan, medikuek diagnosi azkarrak egin behar dituzte; min hori gaixotasun hilgarri askoren aurrekaria izan baitaiteke. Esate baterako, sindrome koronario akutua, aortaren disekzioa, biriketako enbolia, tentsio-pneumotorax-a, taponamendu perikardikoa eta mediastinitis-a. Bularraldeko minaren hasiera-data, bere mota (ziztatzailea, zapaltzailea, erremina, pleuritikoa...), irradiazioa (besoak, lepoa, bizkar-aldea...), kokapena (zentrua, ezkerraldea, eskuinaldea...), denboran zehar duen bilakaera (konstantea, aldizkakoa...), iraupena (iraupen denbora), faktore astungarri eta arintzaileak (mina zerrekin hobetzen eta okertzen den), mina noiz agertzen den (korrika hastean, eseritakoan...) eta sintoma gehigarriak (eztula, sukarra, goragalea...) funtsezkoak dira diagnostiko zuzena egiteko eta arrazoi larriak baztertzeko. Horiexek dira medikuak bularraldeko mina duen gaixo bat iristean erregistratu beharko lituzkeen alarma-sintomak. Baina, esan bezala, zenbaitetan behar bezala erregistratu gabe gelditzen dira.

Arazo honen konponbide posible bat, historia klinikoak lengoaia naturalean idatzi beharrean era egituratu batean egitea da. Baina, helburu horrekin sortutako sistema askok ez dute arrakasta handirik izan [4]. Alde batetik, medikuei lan-karga gehigarri bat eskatzen die. Beste aldetik, historia klinikoa behar bezala azaltzeko beharrezkoa den malgutasuna falta zaie. Arrazoi hauek medio, medikuek nahiago dute testu librea. Informazio guztia modu koherentean idazteko aukera ematen baitie. Beraz, puntu honetan sartzen da Hizkuntza Naturalaren Prozesamendua, eszenario ezberdinetan erabili dena [1] [5]. Informazio egituratu eta ez-egituratua konbinatzen dituzten sistemek bi metodoen onurak batzen dituzte.

1.2 Helburua

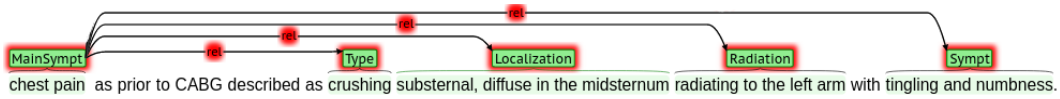
Proiektuaren helburua, Hizkuntza Naturalaren Prozesamenduko teknikak erabiliz, domeinu klinikoan entitateen eta erlazioen erauzketa egiten duen sistema bat garatzea da.

Zehazki, 1.1 sekzioan aipaturiko bularraldeko minaren alarma-sintomak detektatzen dituen sistema bat garatzeko ezinbestekoa den Hizkuntzaren Prozesamenduan oinarritutako tresna sortu nahi da. Honetaz gain, sortuko den erreminta ebaluatu nahi da. Entrenatzeko corpus handi bat edukiz gero, erraza da hori egitea. Baina, txikia izanez gero, beharrezkoa da tresna corpus handiago baten gainean probatzea (2.3 sekzioan informazio gehiago).

Sortzen den tresnak aplikazio ugari izango ditu. Batetik, praktika klinikoan gehien ahazten diren datuak zein diren jakin ahal izango da eta horrek profesionalen heziketa hobetzeko bidea ekarriko du. Bestetik, testuak denbora errealean aztertzen dituzten eta *checklist* gisa funtzionatzen duten sistemak garatzeko aukera egongo da. Hauek, testuan daturen bat faltako balitz, medikuari abisu bat pasako liokete.

Sortuko den tresna gai izango da, entitate klinikoak identifikatzeko eta elkarren artean erlazionatzeko. Kasu honetan, bularraldeko minaren aipamenak eta aipamen bakoitzarekin erlazionaturiko alarma-sintomak. Sistemak, 1.1 irudian ikus daitekeen moduko emaitzak lortzeko ahalmena izango du.

Planteaturiko atazak zailtasun ugari ditu. Normalean, testu klinikoetan ez da bularraldeko minaren inguruko informazioa bakarrik agertzen. Gaixoa, patologia bat baino gehiagorekin iris daiteke larrialdi zerbitzuetara. Testuetan honelako esaldiak aurki daitezke: *Astelehenetik, gaixoak bularraldeko mina du eta arnasa hartzeko zailtasunak ditu atzotik.*



1.1 Irudia: Lortu nahi den emaitza

Bertan, bi hasiera-data daude, baina bakar bat da bularraldeko minarekin lotuta dagoena. Beraz, bi lan egin behar dira: hasiera-datak identifikatu eta bularraldeko minarekin erlazionaturik dagoena zein den erabaki.

Datu urritasunak eta domeinu klinikoak beste zailtasun gehigarri bat suposatzen dute. Edozein ikaskuntza automatiko gainbegiratuk, entrenatzeko eskuz anotatutako datu ugari behar ditu. Aurrerago azalduko den moduan, kasu honetan oso datu etiketatu gutxi daude eskuragarri. Beraz, honek beste helburu bat ekarri du: entitate eta erlazioak datu urrirekin erauztea.

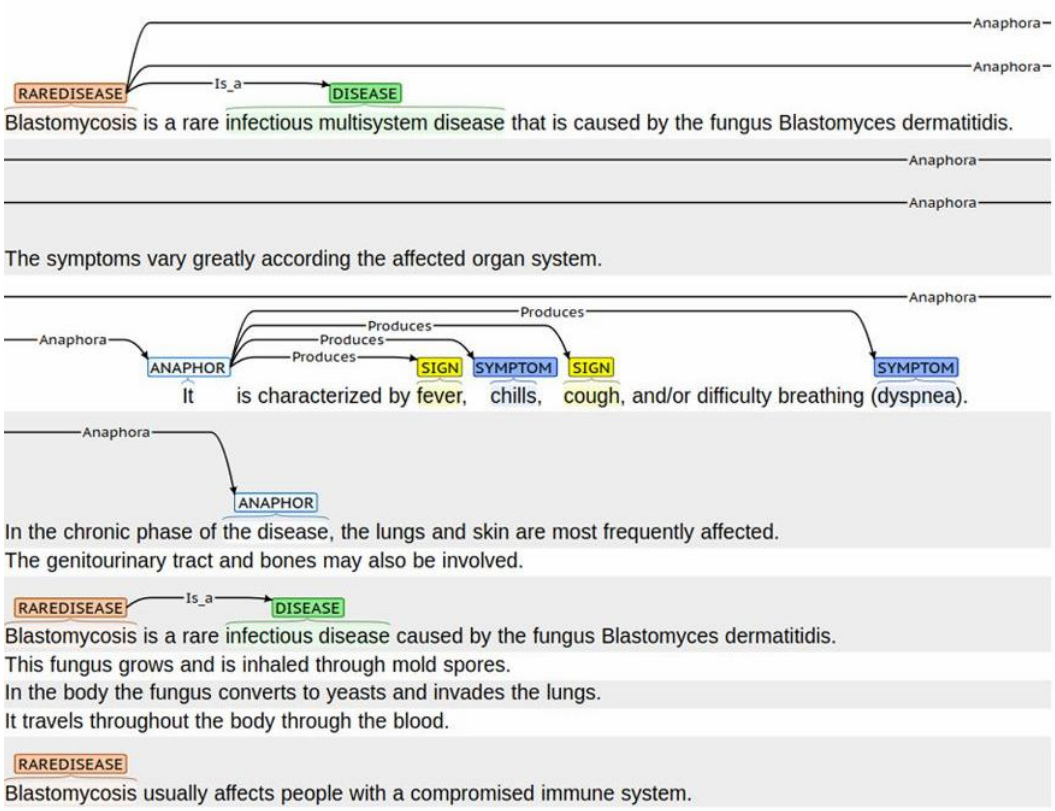
Azken urteetan, hizkuntza eredu aurre-entrenatuek ikasketaren zati bat modu ez gainbegiratuan egitea ahalbidetu dute. Alegia, hizkuntzari eta ez ataza konkretu bati dagozkion hitz/hitz zatien distribuzioak aldez aurretik ikastea (aurre-entrenamendu fasea). Honela, hizkuntza eredu hauek ikasitako distribuzioak moldatu daitezke ataza konkretuen datu anotatuekin entrenamendua jarraituz (fine-tuning edo doiketa fasea). Hizkuntza ereduen entrenamenduan domeinua ondo islatuta egotea garrantzitsua da. Domeinu klinikoa askotan infra-errepresentatuta egoten da. Bestalde, doikuntza fasean ataza konkretuan eskuz anotatutako datuen kantitateak ere eragin handia izaten du emaitzetan. Izatez, zenbat eta datu gehiago erabili, orduan eta emaitza hobeak lortzen dira.

Hau guztia jakinda, honako lau betebehar hauek finkatu dira:

- Bularraldeko minaren aipamenak eta datu kritikoak identifikatzen dituen sistema bat edo gehiago garatzea.
- Bularraldeko minaren aipamenak testuan agertzen diren datu kritikoekin erlazionaturik ote dauden esaten duen sistema bat edo gehiago garatzea.
- Garaturiko sistema ezberdinen konparazioak egitea eta ataza bakoitzerako sistema hoberenak aukeratzea.
- Behin bi sistemak aukeratuta, bi sistemak batzea ataza orokorra burutzen duen sistema lortzeko.

Ataza orokorra egoki burutzen duen sistema bat lortzea interesgarria izango da helburu ezberdinetarako:

- Erabakiak hartzeko prozesu ezberdinei oinarri bat emateko:
  - Diagnostikoa
  - Pronostikoa
  - Ekonomikoa
- Informazio klinikoaren erregistroa estandarizatzeko, elementu deskribatzaile kritiko guztiak biltegiratzen direla ziurtatuz.



1.2 Irudia: Medikuntzako entitate eta erlazio erauzketa adibide bat

- Estatistika eta datu meatzaritza gaitzeko eta osasunerako eta bioinformatikarako informazio egituratua emateko.
- Etorkizunean beste patologia batzuentzat erreminta orokortuak sortzeko. Adibidez, abdominaletako min eta sindrome traumatikoarentzat. Zeinak bularraldeko minarekin batera, larrialdi zerbitzuetara joateko arrazoi errepikatuenetakoak diren.
- Domeinu bereko antzeko atazak burutzen dituzten sistemak garatzeko. Ataza berdintsu ugari topatzen dira medikuntzan. Adibidez, testuetan agertzen diren gaixotasunak euren ezaugarriekin lotzeko ataza (ikusi 1.2 irudia). Funtsean 1.1 eta 1.2 irudietan lortzen diren emaitzak oso antzekoak dira. Batentzat sortzen den ereduak, bestearen oinarri bezala balio dezake.

1.3 Dokumentuaren antolaketa

Kapitulu honetaz gain, beste bederatzi aurki daitezke dokumentu honetan. Bigarrengoan atal honetan definituriko proiektua aurrera eraman ahal izateko garaturiko plangintza aurkezten da. Hirugarrengoan, proiektuaren testuingurua ulertzen laguntzen duten oinarri teoriko batzuen azalpena ematen da. Laugarrengo kapituluan, proiektua garatzeko erabilitako corpus eta metodo ezberdinak aurkezten dira. Bosgarrenean, proiektuaren garapen prozesu guztia azaltzen da, hartutako erabaki, erabilitako tresna eta garaturiko modelo guztien inguruko informazioarekin. Seigarren atalean, lorturiko emaitza guztien analisi

sakon bat egiten da eta hauetatik atera daitezkeen ondorioak aztertzen dira. Zazpigarrenean, bigarren atalean definituriko plangintza zenbateraino bete den aurkezten da. Zortzigarren kapituluan berriz, proiektu guztia laburbiltzen duten ondorio nagusiak ateratzen dira eta etorkizunera begira egin daitezkeen lanak aipatzen dira. Azkenik, amaierako kapituluetan, informazio gehigarria eskuratzeko eranskinak eta erabilitako bibliografia agertzen dira.

2. Plangintza

Kapitulu honetan, aurreko atalean definitu den proiektua aurrera eraman ahal izateko garaturiko plangintza aurkezten da. Plangintza hau hiru atal ezberdinetan banatzen da. Hasteko, lanaren deskonposaketa eredua aurki daiteke. Alegia, proiektu handia zein zatitan banatu den. Jarraian, ataza bakoitza noiz eta zenbat denboraz egingo den adierazten dituzten *gantt*-en diagrama eta denbora taulak daude. Azkenik, proiektuan topa daitezkeen arrisku ezberdinak eta hauek agertuz gero hartuko diren bideak agertzen dira. Proiektu honetan, bereziki, azken atal horri eman zaio garrantzia.

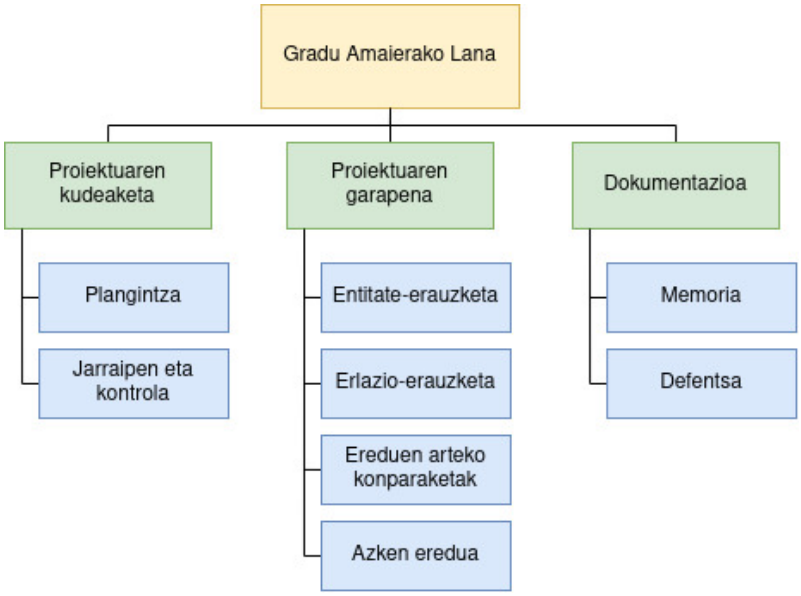
2.1 Lanaren deskonposaketa egitura

Gradu Amaierako Lana, hiru azpiproiektu nagusitan banatu da: proiektuaren garapena, dokumentazioa eta proiektuaren kudeaketa. Zati bakoitzean, lan-pakete ezberdinak definitu dira eta lan-pakete bakoitzeko ataza desberdinak. 2.1 irudian lan guztiaren deskonposaketa biltzen duen LDE diagrama aurkitzen da.

2.1.1 Proiektuaren garapena

Proiektuaren muina biltzen duen eta iraupen luzeena duen fasea da. Bertan, proiektu honetan garatu nahi den sistema sortzeko beharrezkoak diren pausu guztiak biltzen dira. Zehazki, lau lan-paketetan bereizi da, 1.2 atalean zehazturiko betebeharrekin estuki loturik daudenak.

1. Entitate-erauzketa (EnE): Bularraldeko minaren aipamenak eta datu kritikoak identifikatzen dituen sistema bat edo gehiago garatzeko eginbehar guztiak sartzen dira bertan. Beraz, irteera moduan gutxienez entitate-erauzle bat lortuko da.
    - EnE.1: Gisa honetako atazak egiteko dauden modu ezberdinak ikertu.
    - EnE.2: Corpusa egokitu probatuko diren teknika desberdinetarako.
    - EnE.3: Teknika bakoitzarekin entrenamenduak egin sistema ezberdinak lortzeko.
2. Erlazio-erauzketa (ErE): bularraldeko minaren aipamenak testuan agertzen diren datu kritikoekin erlazionaturik ote dauden esaten duen sistema bat edo gehiago garatu ahal izateko burutu beharreko lanak biltzen ditu. Irteeran, gutxienez erlazio-erauzle bat lortuko da.
    - ErE.1: Gisa honetako atazak egiteko dauden modu ezberdinak ikertu.
    - ErE.2: Corpusa egokitu probatuko diren hurbilpen desberdinetarako.
    - ErE.3: Teknika bakoitzarekin entrenamenduak egin sistema ezberdinak lortzeko.



2.1 Irudia: LDE diagrama

3. Ereduen arteko konparaketak (EAK): Bi atazetako bakoitza (entitate-erauzketa eta erlazio-erauzketa) egiteko zein sistema den egokiena erabakitzeko egin beharreko lanak. Irteera gisa, ataza bakoitzeko sistema bat lortuko da.

    - EAK.1: Ataza bakoitza ebaluatzeko zein ebaluazio metrika erabiliko den erabaki.
    - EAK.2: Sistema bakoitza aukeraturiko ebaluazio metrikarekin ebaluatu eta emaitzekin konparazio taulak eta grafikoak egin.
    - EAK.3: Taula eta grafiko ezberdinak aztertuta sistema eta hurbilpen hoberenak zein diren erabaki.

4. Azken eredua (AE): Behin bi sistema hoberenak edukita bi atazak burutzen dituen sistema garatzeko eginbeharrak. Irteeran, sistema orokorra lortuko da.

    - AE.1: Entitate-erauzlearen irteerak erlazio-erauzlearen sarrera izateko nola eraldatuko diren pentsatu eta azken emaitzak zein formatutan emango diren erabaki.
    - AE.2: Bi sistemak batzen dituen eredua garatu.
    - AE.3: Sortu den sistema ebaluatu.

2.1.2 Dokumentazioa

Fase hau bi zatitan banatzen da: memoriaren idazketa eta defentsaren prestaketa. Hau da, behin proiektua amaiturik, egindako lana erregistraturik uzteko burutu beharreko ekintzak.

1. Memoria (M): Memoria idaztearekin loturiko eginkizun guztiak daude bertan. Irteera gisa proiektuaren nondik norako guztiak biltzen dituen dokumentua lortuko da.

    - M.1: Memoria on batek bete beharreko baldintzak zein diren ikertu.

- M.2: Aurreko urteetan defendatu diren memoriak gainbegiratu.
- M.3: Memoriaren eskema prestatu.
- M.4: Memoria idatzi

2. Defentsa (D): Idatzi den memoria defendatzeko egin beharreko lanak. Irteera moduan, jendaurrean egingo den defentsa lortuko da.
    - D.1: Jendaurrean azalduko diren ideia nagusiak zein izango diren erabaki.
    - D.2: Ideia nagusi bakoitzarekin gutxi gora-behera zer azalduko den erabaki.
    - D.3: Aurkezpen dokumentua prestatu.

2.1.3 Proiektuaren kudeaketa

Proiektuaren garapen egokia bermatuko duten zereginak daude bertan. Bi lan-paketetan banatzen da: Plangintza eta jarraipen eta kontrola.

1. Plangintza (P): Proiektua hasi aurretik plan egoki bat eduki ahal izateko eginbeharrak daude bertan. Irteera moduan plangintza biltzen duen dokumentua lortuko da.
    - P.1: Arazoaren azterketa sakona egin.
    - P.2: Hasierako plangintza bat garatu.
    - P.3: Plangintza eguneratu, beharrezkoa bada.
2. Jarraipen eta kontrola (JK): Proiektua benetan betetzen ari dela egiaztatzeko beharrezkoak diren lanak. Irteera gisa, proiektuaren garapenaren nondik norakoak biltzen dituen dokumentu bat lortuko da.
    - JK.1: Proiektuaren garapenari buruzko informazio garrantzitsua jaso.
    - JK.2: Proiektua ongi garatzen ari dela ziurtatzeko eta erabaki berriak hartzeko bilerak egin.

2.2 *Gantt*-en diagrama eta denbora taula

Ataza guztiak definitu ostean bakoitza noiz eta zenbat denboraz egin behar den erabaki behar da. 2.1 taulan ataza guztiak zein denboralditan egingo diren adierazten duen *gantt*-en diagrama aurkezten da. Aldiz, 2.2 taulan, ataza bakoitzarekin igaroko den denbora ikus daiteke ordutan.

Gantt diagramari dagokionez, ikusten da nahiko sekuentziala izango dela prozesua. *Proiektuaren garapena* faseko lan-paketeak elkarren segidan egingo dira, bat amaitzean bestearekin hasiz. Behin hauek egindakoan, dokumentazioaren atala egingo da. Aldiz, jarraipen eta kontrolerako zereginak proiektuaren garapen guztian zehar egingo dira.

Denboren taulan berriz, ikusten da garrantzi gehiena proiektuaren garapenari emango zaiola 200 ordurekin. Horretaz gain, dokumentazioa egiteko 70 ordu inguru beharko direla pentsatzen da. Hauei proiektuaren kudeaketako 30 orduak batuta, guztira Gradu Amaierako Lanak bete beharko lituzkeen 300 orduak lortzen dira.

Ataza	Hilabeteak						
	abu.	...	urt.	ots.	mar.	api.	mai.
EnE.1							
EnE.2							
EnE.3							
ErE.1							
ErE.2							
ErE.3							
EAK.1							
EAK.2							
EAK.3							
AE.1							
AE.2							
AE.3							
M.1							
M.2							
M.3							
M.4							
D.1							
D.2							
D.3							
P.1							
P.2							
P.3							
JK.1							
JK.2							

2.1 Taula: Proiektuaren *gantt* diagrama

2.3 Arriskuen analisia

Proiektua hain luzea izanik, hau amaitzea zalantzan jar dezaketen hainbat arrisku ager daitezke. Behar-beharrezkoa da arrisku hauek hasieratik identifikaturik edukitzea, eta hauek gertatuz gero zer egingo den argi izatea. Bi dira lan honetan identifikatu diren arrisku garrantzitsuenak.

Hasteko, proiektu hau garatu ahal izateko, behar-beharrezkoa da GPUen erabilera. Izan ere, eredu asko entrenatu beharko dira eta GPUrik gabe asko atzeratuko da lana. Zehazki, Ixa ikerketa taldeko zerbitzarietako GPUak erabiltzea erabaki da. Baina, proiektu hau 2023ko urtarriletik maiatzera bitarte garatuko da gutxi gora-behera. Denbora tarte honetan Ixa taldeko ikerlariek lan ugari edukiko dutela aurreikusten da, eta baliteke GPUrik libre ez egotea. Edo, beharbada, libre egongo diren GPUak ez dira behar adina ahaltsuak izango. Hau gertatuko balitz, Google-en GPUak erabiltzea pentsatu da, *Google Colaboratory* bidez atzitu daitezkeenak. Hori bai, kontuan izan behar dira hauek eskaintzen dituzten murriztapenak. Denbora mugatuz erabil daitezke eta ez dute memoria infinitua.

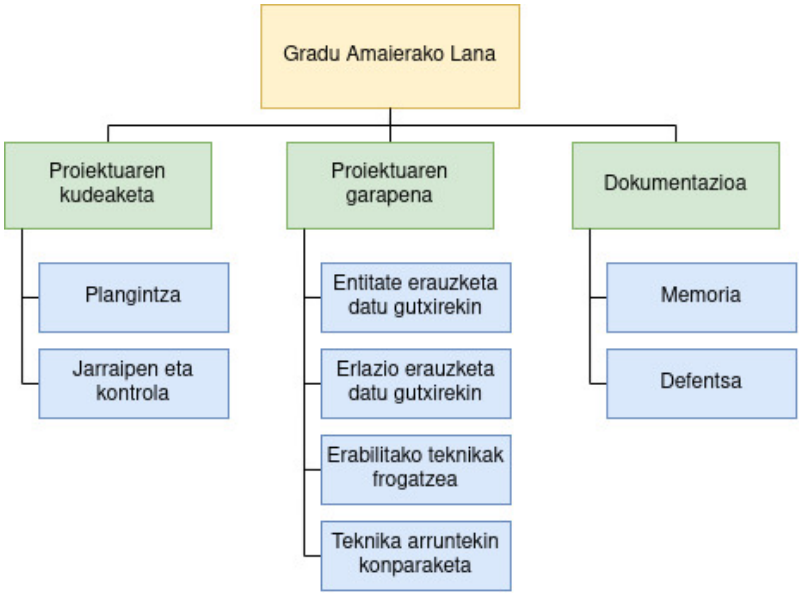
Bigarren arriskuak arazo larriagoak suposatzen ditu. Gaur arte, ez da inon aurkitu proiektu honetan planteatu den arazo konkretua burutzen duen eta argitaratuta dagoen sistemarik. Are gutxiago doiketa fasea egiteko behar den eskuz anotatutako corpusik.

Lan-paketea	Ataza	Denbora (h)
Proiektuaren garapena		200
Entitate erauzlea	EnE.1	15
	EnE.2	20
	EnE.3	35
Erlazio erauzlea	ErE.1	15
	ErE.2	20
	ErE.3	35
Ereduen arteko konparaketak	EAK.1	4
	EAK.2	10
	EAK.3	6
Azken eredua	AE.1	10
	AE.2	20
	AE.3	10
Dokumentazioa		70
Memoria	M.1	1
	M.2	2
	M.3	2
	M.4	50
Defentsa	D.1	3
	D.2	5
	D.3	7
Proiektuaren kudeaketa		30
Plangintza	P.1	6
	P.2	6
	P.3	6
Jarraipen eta kontrola	JK.1	2
	JK.2	10
Guztira		300

2.2 Taula: Ataza bakoitzari eskainiko zaion denbora

Adimen Artifizialean aberastasuna datuek ematen dute. Honelako ataza bat burutzeko, oso garrantzitsua da corpus hornitu bat edukitzea. Hori jakinda, BioDonostiako bi medikuek corpusa sortzeko konpromisoa hartu dute. Epe luzeko helburuetako bat, garatutako tresna Osakidetzako ordenagailuetan inplementatzea denez, bertako testuak lortu nahi izan dira. Baina, Datu Pertsonalak Babesteari buruzko legea (LOPDP Ley Orgánica 3/2018) dela eta, hauek lortzeko prozesua oso luzea da. 2022ko urte hasieratik hauek lortu nahian dabiltza. Testu hauen zain egoteko denborarik ez dagoenez, ingelesezko corpus publiko bateko (*MIMIC-III* [6]) testuak erabiltzea erabaki da ikerketa hasteko. Halere, testuak lortzea ez da nahikoa. Anotatu egin behar dira. Horrek lan karga handia suposatzen du. 2022ko abenduan, bi medikuek bakoitzak bere aldetik etiketaturiko 20 testu dituzte. Eta adostasuna ez da erabatekoa. Adostasun bat lortu behar da eta honetaz gain testu etiketatu gehiago nahi dira. Baina, medikuek euren bizitza profesionalean lan ugari dute eta posible da testuen etiketatzea atzeratzea.

Egoera honetan, ez da posible ikusten sistema sendo bat garatzeko ahalmena edukiko denik. Beraz, etiketatzea atzeratuko balitz, ezingo litzateke definituriko plangintza aurrera



2.2 Irudia: LDE diagrama alternatiboa

eraman. Horri aurre egiteko, plangintza alternatibo bat definitu da. Proiektuaren helburua proposatu den sistema sortzea izatetik, corpus oso txikiak edukitzean har daitezkeen bide ezberdinak eta hauen errendimenduak aztertzea izatera pasako litzateke. Azken finean, antzeko corpus batean ataza ongi burutzea lortuz gero, Osakidetzako testuak eskuragarri daudenean moldaketa bat besterik ez da egin beharko.

Proiektuaren helburu berriak hauek izango lirateke:

- Entitateen detekzioa egitea datu etiketatu oso gutxi edukita.
- Erlazioen detekzioa egitea bestelako teknika batzuk (*zero-shot* eta *few-shot*) erabilita.
- Erabilitako teknikak benetan egokiak direla frogatzea.
- Erabilitako teknikak teknika tradizionalekin konparatzea.

2.2 irudian LDE diagrama alternatiboa aurkezten da. 2.1 irudikoarekin duen ezberdintasun bakarra *Proiektuaren garapena* fasea da. Lan-pakete berri batzuk agertzen dira.

Honako hauek dira sortzen diren lan-pakete berriak eta bakoitzari dagozkion atazak:

1. Entitate-erauzketa datu gutxirekin (EnEDG): Entitate erauzketarako etiketatutako datu gutxi dauden kasuetan har daitezkeen bide ezberdinak ikertzeko egin beharreko zereginak. Irteera gisa entitate detektatzaile bat lortuko da.
    - EnEDG.1: Datu gutxi dauden kasuetan har daitezkeen bide posibleak ikertu.
    - EnEDG.2: Bide ezberdinak probatu eta emaitzak ebaluatu
2. Erlazio-erauzketa datu gutxirekin (ErEDG): Erlazio erauzketarako etiketatutako datu gutxi dauden kasuetan har daitezkeen bide ezberdinak ikertzeko burutu beharreko lanak. Irteera gisa erlazio-erauzle bat lortuko da.

Ataza	Hilabeteak						
	abu.	...	urt.	ots.	mar.	api.	mai.
EnEDG.1 EnEDG.2							
ErEDG.1 ErEDG.2							
ETF.1 ETF.2 ETF.3							
TAK.1 TAK.2 TAK.3							

2.3 Taula: *Gantt* diagrama alternatiboa

- ErEDG.1: Datu gutxi dauden kasuetan har daitezkeen bide posibleak ikertu.
- ErEDG.2: Bide ezberdinak probatu eta emaitzak ebaluatu

3. Erabilitako teknikak frogatzea (ETF): Erabilitako teknikak benetan sendoak direla egiaztatzeko balio duen eginkizun oro sartzen da bertan. Irteera gisa erabilitako teknikak egokiak diren edo ez lortuko da.

   - ETF.1: Antzeko corpus bat bilatu (handia dena).
   - ETF.2: Erabilitako teknikak corpus berriaren gainean probatu.
   - ETF.3: Emaitzak aztertu eta teknikak egokiak ote diren baloratu.

4. Teknika arruntekin konparaketa (TAK): Egoera berdinean egonez gero (datu gutxi) teknika tradizionalak aplikatuz lortuko liratekeen emaitzak ikusteko beharrezko ataza guztiak. Irteera bezala, teknika tradizionalak edo berritzaileak diren hobeak lortuko da.

   - TAK.1: Teknika tradizional bat aukeratu.
   - TAK.2: Teknika tradizionala corpus handiaren gainean probatu.
   - TAK.3: Lorturiko emaitzak aurreko teknikekin konparatu eta hoberena zein den erabaki.

Ataza hauei dagokien gantt-en diagrama 2.3 taulan topa daiteke. Ikus daitekeen moduan, berezko *gantt* diagramaren forma oso antzekoa du. Atazak dira aldatzen direnak. Aldiz, lan hauetako bakoitzari eskaintzea espero den denbora 2.4 taulan aurki daiteke. Pakete honi, hasierako plangintzan definituta dagoen moduan, 200 ordu eskainiko zaizkio. Ordu horiek zeregin berrien artean banatzen dira.

Beraz, bigarren arriskua gertatuko balitz, dagoeneko definiturik dago garatuko litzatekeen plangintza alternatiboa.

Lan-paketea	Ataza	Denbora (h)
Proiektuaren garapena		200
Entitate erauzketa datu gutxirekin	EnEDG.1	20
	EnEDG.2	30
Erlazio erauzketa datu gutxirekin	ErEDG.1	20
	ErEDG.2	30
Erabilitako teknikak frogatzea	ETF.1	5
	ETF.2	35
	ETF.3	10
Teknika arruntekin konparaketa	TAK.1	5
	TAK.2	30
	TAK.3	15

2.4 Taula: Ataza alternatibo bakoitzari eskainiko zaion denbora

3. Oinarri teorikoak

Kapitulu honetan proiektuaren testuingurua ulertzen laguntzen duten hainbat azalpen teoriko ematen dira. Egindako guztiak, oinarri teoriko bat du. Hartutako erabaki bakoitza teoriako kontzeptuetan oinarritzen da. Beraz, kapitulu honek, hurrengo kapituluetan agertuko diren edukiak hobeto ulertzeko balio du.

3.1 Zer da Hizkuntza Naturalaren Prozesamendua?

Dagoeneko, dokumentu honetan, behin baino gehiagotan errepikatu da Hizkuntza Naturalaren Prozesamendua terminoa, baina, ezer gutxi esan da honen inguruan. Adimen artifizialeko adar garrantzitsu bat da eta bere helburua sistema informatikoek gizakien hizkuntza prozesatu eta ulertzea da. Definizio formal bat ematekotan, honako hau eman daiteke:

> Hizkuntza Naturalaren Prozesamendua (HNP), edo ingelesez *Natural Language Procesing* (NLP) moduan ezagunagoa dena, teorikoki motibatutako teknika konputazionalen multzo bat da hizkuntza-analisiaren maila bat edo gehiagotan testu naturalak aztertu eta irudikatzeko balio duena, eta helburu gisa gizakiaren antzeko hizkuntzaren prozesamendu bat lortzea duena ataza eta aplikazio ezberdinak burutu ahal izateko [7].

Gizakien hizkuntza idatzia edo ahoz esana den arren, Hizkuntza Naturalaren Prozesamenduak adimen artifizialeko teknikak erabiltzen ditu mundu errealeko sarrera hartu, prozesatu eta konputagailu batek ulertzeko moduko zentzua emateko. Pertsonek zentzumen organo ezberdinak dituzten moduan (hala nola, entzuteko belarriak eta ikusteko begiak), ordenagailuek testuak irakurtzeko programak eta audioa jasotzeko mikrofonoak dituzte. Bestalde, gizakiek pentsatzeko burmuina duten bezala, konputagailuek programak dituzte sarrera prozesatzeko balio dutenak.

HNPn bi fase nagusi bereizten dira: datuen aurre-prozesaketa eta algoritmoen garapena. Datuen aurre-prozesaketa, testuak prestatzeko eta garbitzeko prozesua da, makinek hauek analizatzeko gaitasuna izan dezaten. Behin datuak prest izanda, hauek prozesatzeko algoritmoak garatzen dira [8].

Aurre-prozesaketako bi pausu garrantzitsu, tokenizazioa eta lematizazioa dira. Helburua testuak formatu erabilgarriago batera transformatzea da.

Tokenizazioa Hizkuntza Naturalaren Prozesamenduko ataza sinpleenetako bat da. Sarrerako testua, ordenagailu batentzat karakterez osaturiko kate luze bat besterik ez dena, azpiunitatetan (token deiturikoak) zatitzean datza. Aukera errazena, esaldia zuriuneak jarraituz zatitzea da, baina badaude token gisa hitzak eta puntuazio markak erauzteko gai diren tokenizatzaileak ere. Normalean, ataza konplexuagoak burutu ahal izateko aurrekari bat izaten da tokenizazioa (aurre-prozesaketaren parte alegia). Hala nola, analisi morfologikoa, hitz bakoitzari klase bat esleitzea eta analisi sintaktikoa [9].

Esaldia	Tokenizazioa eta lematizazioa
He has arrived very early!	He | has | arrived | very | early | !
	He | have | arrive | very | early | !
Your bike is better than mine.	Your | bike | is | better | than | mine | .
	You | bike | be | good | than | I | .

3.1 Taula: Tokenizazio eta lematizazio adibideak

Lematizazioa, tokenizazioaren antzera, testu-meatzaritzaren aplikazio ezberdinen aurreprozesamenduko pausu garrantzitsu bat da. Sarrerako testuko hitzen forma normalizatua bilatzeko prozesua da. Hitzen erro-bilaketaren antzekoa da, baina ez guztiz berdina. Zenbait kasutan, hitzen erroa eta forma normalizatua bat datoz. Adibidez, ingelesezko *working*, *works* eta *worked* hitzen erroa, *work* da; aldi berean hauen forma normalizatua dena. Aldiz, *computes*, *computing* eta *computed* hitzen erroa *comput* da, baina, forma normalizatua *compute* da (aditzaren infinitiboa). Lematizazioa, tokenizazioaren ostean egiten da. Tokenizazioan, testuko hitzak lortzen dira eta lematizazioan hitz bakoitzari dagokion lema edo forma normalizatua [10]. 3.1 taulan tokenizazio eta lematizazio adibide pare bat daude ikusgai.

Algoritmoen garapenerako, berriz, tresna oso erabili bat *Transformer* arkitektura da. Honek, Adimen Artifizialean oso erabiliak diren sare neuronaletan du oinarria. Sare neuronal sinpleena pertzeptroia da. Sailkatzaile lineal bat da, sarrera moduan bektore bat hartzen duena $(x_1, x_2, ..., x_n)$ eta W bektoreaz biderkatzen duena $(w_1, w_2, ..., w_n)$. Ondoren, *bias* izenez ezagutzen den konstante bat batzen dio. W eta *bias*-a entrenamendu bitartez ikasten dira [11]. Hona hemen konputatzen duen formula:

$$f(x) = W \cdot X + b$$

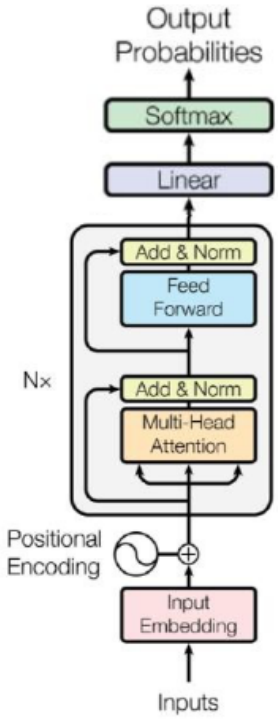
Pertzeptroiak elkarren artean batuz, *Multi Layer Perceptron*-ak (MLP) edo geruza anitzeko pertzeptroiak lortzen dira. Ikasteko gaitasun handiagoa erakusten dutenak.

Multi Layer Perceptron-en ostean *Recurrent Neural Network* (RNN) delakoak sortu ziren. Hauek, datu sekuentzialekin egiten dute lan (adibidez, testua). Sarrera hainbat zatitan banatzen da (adibidez, testua tokenetan). Zati bakoitzeko sailkapen bat egiteko aukera ematen dute. Zati baten sailkapena egitean, aurretik edo ondoren dauden zatien informazioa jasotzeko aukera ere badute. Baina, konputazionalki garestiak dira. RNN arkitektura hobetu nahian *Transformer*-ak agertu ziren.

3.2 *Transformer*

Transformer-a Hizkuntza Naturalaren Prozesamendua erabat aldatu duen arkitektura bat da. HNPko ataza askotan, hizkuntza naturalean dagoen testua tokenizatu egiten da eta token sekuentzia bat lortzen da. *Transformer*-ak sekuentzia horren transdukzioa egiteko bidea ematen du. Hau da, token sekuentzia transformatzen du ataza burutzeko informazio esanguratsuagoa ematen duen errepresentazio batera. Hau egiteko, *Transformer*-ak atentzio mekanismoak bakarrik erabiltzen ditu (*Attention is all you need*).

Oraingoz, aipaturiko sekuentziako tokenak, karakterez osaturiko "hitzak" dira. Baina, *Transformer* bati pasa ahal izateko, zenbakizko bektorez adierazita egon behar dute. Hau

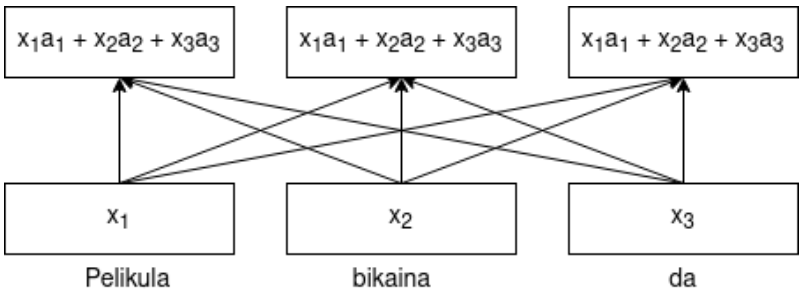


3.1 Irudia: *Transformer* arkitektura

da, token bakoitzarentzat errepresentazio numeriko bat lortu behar da (zenbakizko bektore bat), *embedding* deitzen zaiona. Modu ezberdinak daude *embedding*-ak lortzeko, baina, sinpleena hitz-zakuarena da. Hiztegi bat definitzen da, atazan ager daitezkeen token guztiekin. Token bakoitzeko, hiztegiaren luzera bera duen *one-hot encoding* erako bektore bat definitzen da embedding gisa. Bektoreak 0 balioa du gelaxka guztietan bakar batean izan ezik, tokenari dagokion posizioan 1eko bat du. Posizio hau hiztegian duen posizioak finkatzen du. Hiztegian agertzen ez diren token guztientzat UNK token berezia definitzen da. Embedding hauek, token bakoitza identifikatzeko balio dute, baina informazio gutxi ematen dute, *transformer*-ari esker eraldatu egiten dira. 3.1 irudian *transformer* arkitektura aurki daiteke.

Hasteko, sarrerako *embedding*-ei atentzio mekanismoak aplikatzen zaizkie. Atentzioa izatez, sekuentziako *embedding* guztien arteko batura haztatua da. Token bakoitzeko, *embedding* bakoitza pisu baten bidez biderkatzen da eta denen arteko batura egiten da. 3.2 irudian atentzio geruzaren eskema bat topa daiteke. Bertan agertzen diren x_i balioak, sarrerako tokenen *embedding*-ak dira. a_i balioak berriz aipaturiko pisuak dira, $[0, 1]$ tartean daude eta guztien arteko baturaren emaitza 1 da.

Pisu hauek lortzeko q eta k bektoreak erabiltzen dira. Bien arteko biderkadura eskalarra egiten da eta emaitza k bektorearen dimentsioaren erroketaren bidez zatitzen da ($\sqrt{d_k}$). q eta v entrenamenduan ikasten diren parametroak dira eta token bakoitzeko bana edukitzen dira. Bektore horiek, Q eta K matrizeetan jasotzen dira. Honetaz gain, hitzen *embedding* guztiak V matrizean jasotzen dira. Zehazki, atentzio geruza batean honako kalkulu hauek konputatzen dira:



3.2 Irudia: Atentzio geruzaren eskema

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Esan bezala, q eta k bektoreen arteko biderkadura d_k balioaren erro karratuaz zatituz lortzen dira pisuak. Ondoren, *softmax* funtzioa aplikatzen zaie. Azken horri esker, pisu guztiak $[0, 1]$ tartean egotea eta denen arteko batura 1 izatea lortzen da. Pisuen eta v bektoreen arteko biderkadura eskalarrak eginda, sekuentziako *embedding* guztien arteko batura haztatua lortzen da.

Kalkulu horiek, atentzio geruza arrunt batenak dira, baina, 3.1 irudian ikus daitekeen moduan *transformer*-ek buru anitzeko atentzio geruza dute (*Multi Head Attention*). Hau da, atentzioa h aldiz errepikatzen da Q eta K matrize ezberdinekin eta emaitzak bateratzen dira.

3.1 irudira itzuliz, atentzio geruzaren ostean normalizazio geruza bat agertzen da eta jarraian *feed forward* geruza arrunt bat. Hau ere normalizazio geruza batez jarraituta. Honetaz gain, azpimarratzekoa da bi azpiblokeek duten hondakin-konexioa (*residual contection*) [12]. *Transformer*-etan nahi adina (N_x) atentzio eta *feed forward* geruza jar daitezke.

Honela, *transformer*-ek sarrerako testuaren errepresentazio esanguratsuak lortzen dituzte. Hauek, ataza ezberdinetarako erabil daitezke. Adibidez, testuen sailkapen ataza baterako. Horretarako, nahikoa da *transformer* baten irteeran *feed forward* geruza bat eta *softmax* bat gehitzea (3.1 irudian egiten den moduan).

Transformer-ak, ataza konkretuetarako zerotik entrena daitezke. Baina, zeregin baterako ikasi denak, beste batzuetarako ere balio dezake. Edozein ikaskuntza automatiko gainbegiratuk entrenatzeko eskuz anotatutako datu ugari behar ditu. *Transformer*-ek, ikasketaren zati bat modu ez gainbegiratuan egitea ahalbidetu dute. Ataza baterako eredu batek ikasitako distribuzioak beste ataza baterako moldatu daitezke. Hasierako prozesu hori aurre-entrenamendu fase bezala ezagutzen da. Ataza berrira moldatzeko, atazaren datu etiketatuekin entrenamendua jarraitzen da. Bigarren pausu honi fine-tuning edo doiketa fasea deitzen zaio. Aurre-entrenatzen diren hizkuntza ereduen entrenamenduan domeinua ondo islatuta egotea garrantzitsua da. Domeinu klinikoa askotan infra-errepresentatuta egoten da. Bestalde, doikuntza fasean ataza konkretuan eskuz anotatutako datuen kantitateak eragin handia izaten du emaitzetan.

Aldez aurretik entrenatuta dauden eta ataza berrietarako balio duten ereduei BERT deitzen zaie. Hauetako asko topa daitezke publikatuta *Hugging Face* webgunean [13].

3.3 Izendun Entitateen Erauzketa eta Erlazio Erauzketa

Informazio Erauzketa (*Information Extraction, IE*) Hizkuntza Naturalaren Prozesamenduko ataza multzo bat da. Honen helburua, testu lau ez egituratu batetik, edozein makinak edo programak ulertuko duen informazio egituratua eskuratzea da [14]. Izendun Entitateen Erauzketa eta Erlazio Erauzketa ataza multzo honen parte dira eta normalean eskutik helduta doaz. Erlazio Erauzketa egin ahal izateko, lehenbizi Izendun Entitateen Erauzketa egin behar da. Adibide bat jartzearren, izendun entitate identifikatzaile batek, testu batean agertzen diren pertsona eta herri izenak aurki ditzake. Ondoren, erlazio erauzleak pertsona bakoitza bere jaioterriarekin erlazionatu dezake.

3.3.1 Izendun Entitateen Erauzketa

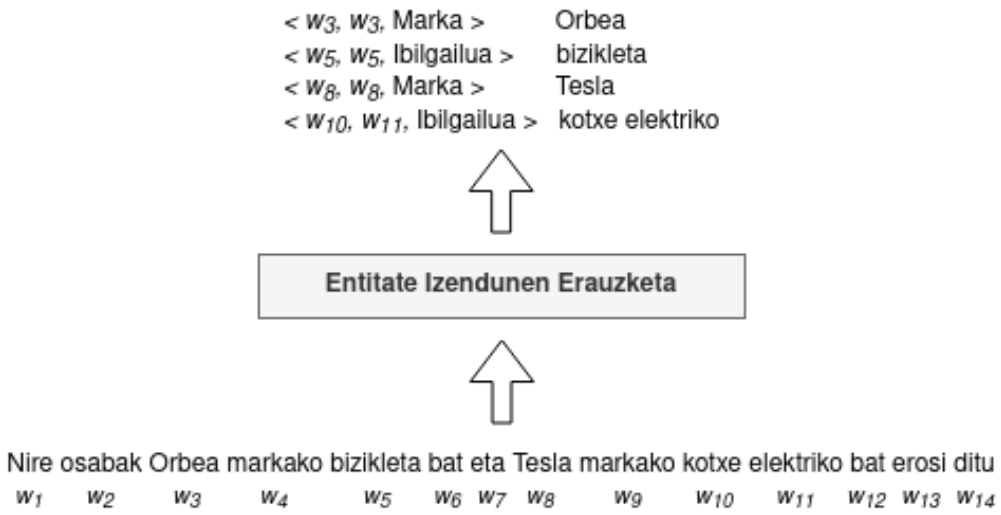
Izendun Entitateen Erauzketa, *Named Entity Recognition* (*NER*) gisa ezagunagoa dena, izendun entitateen bilaketa prozesua da. Izendun entitate bat berriz, hitz edo esamolde bat da, zeina argi eta garbi antzeko ezaugarriak partekatzen dituzten elementu multzo bateko kide den [15]. Esate baterako, *autoa*, *trena* eta *bizikleta* hitzak *ibilgailua* multzoan sar daitezkeenez, izendun entitate kontsidera daitezke. Beraz, ibilgailuak erauzten dituen sistema batek, *Nire osabak Orbea markako bizikleta bat eta Tesla markako auto elektriko bat erosi ditu* sarrera hartuta, *auto elektriko* eta *bizikleta* entitateak topatuko ditu. Ibilgailuekin egin daitekeen moduan, beste klase askorekin egin daiteke. Ataza honetan maiz agertzen diren klaseetako batzuk erakunde izenak, kokapenak, denborazko espresioak eta pertsona izenak dira. Medikuntza alorrari erreparatuz, gaixotasunak, sintomak, sendagaiak eta bestelakoak topatzen dituzten sistemak garatzea ere ohikoa da.

Formalki, token zerrenda bat emanda $s = \{w_1, w_2, ..., w_N\}$, non w_i s testuko i. tokena den, NER sistema batek $< I_s, I_e, t >$ gisako tupla zerrenda bat itzultzen du. Tupla bakoitzak, testuko izendun entitate bati egiten dio erreferentzia. Tuplako $I_s \in [1, N]$ eta $I_e \in [1, N]$ izendun entitate baten hasiera eta amaierako tokenen indizeak dira. t berriz, entitateari dagokion klasea da (aurrez definituriko klase-multzo batetik hartua) [16]. 3.3 irudian NER atazaren adibide bat ikus daiteke, bertan, esaldi batean lau izendun entitate topatu dira.

3.3.1.1 Izendun Entitateen Erauzketa egiteko estrategiak

Izendun Entitateen Erauzketa egiteko estrategia ezberdinak zabaldu dira, gehienak ikasketa gainbegiratuko metodoak izanik. Horietako bat sekuentzia-etiketatzea da. Sarrera gisa N luzerako token sekuentzia bat ematen da, $s = \{w_1, w_2, ..., w_N\}$. Hau jasota, ereduak token bakoitzari y_i etiketa bat esleitzen dio. Honela, irteera moduan N luzerako beste sekuentzia bat lortzen da, $o = \{y_1, y_2, ..., y_N\}$. Irteera horrek, BIO notazioan dagoenak, entitateen inguruko informazioa ematen du.

BIO terminoa, ingelesezko *Beginning*, *Inside* eta *Outside* hitzez osatzen da. Hau da, notazio honek sekuentziako token bakoitza entitate baten hasieran, barnean edo kanpoan dagoen adierazten du. Printzipioz, y_i bakoitzak hiru balio posible har ditzake: B, I edo O. B eta I etiketek entitateak osatzen dituzten tokenak identifikatzen dituzte. Entitateko lehen tokenari B etiketa jartzen zaio eta gainerako guztiei I. Bestalde, O etiketak, entitateetatik kanpo dauden token guztiak etiketatzeko balio du. Entitate klase bakar bat eduki beharrean, X klase multzo bat edukiz gero, B eta I etiketak B-x_i eta I-x_i etiketetan bilakatuko lirateke. Non $x_i \in X$ ezberdinak klaseen izenak diren.



3.3 Irudia: Izendun Entitateen Erauzketaren ilustrazio bat

Nire O	osabak O	Orbea B-Marka	markako O	bizikleta B-Ibilgailua	bat O	eta O
Tesla B-Marka	markako O	auto B-Ibilgailua	elektriko I-Ibilgailua	bat O	erosi O	ditu O

3.2 Taula: BIO notazio adibide bat

3.2 taulan, 3.3 irudiko adibideari dagokion BIO notazioa errepara daiteke. Bi klase ezberdintzen direnez (Marka eta Ibilgailua) B-x_i eta I-x_i motako etiketak agertzen dira.

Beraz, sarrerako sekuentziako token bakoitzari BIO etiketa bat esleitzen dion sistema bat lortu behar da. Era ezberdinak daude hori egiteko. Baina, modu sinple bat aurrez entrenatuta dagoen BERT bat erabiltzea da. Azaldu den moduan, BERT eredu batek sarrerako sekuentzia baten token bakoitzeko errepresentazio bat lortzeko gaitasuna du. Errepresentazio hori B, I eta O irteerak (edo B-x_i eta I-x_i guztiak eta O) dituen sailkatzaile baten sarrera gisa erabil daiteke. Sailkatzaile mota ugari daude, sinpleena *feed forward* geruza soil bat erabiltzea litzateke. Geruza honek, sarrera bezala errepresentazioaren elementu kopurua bezainbeste balio hartuko lituzke. Irteera moduan berriz, $2 \times |X| + 1$ balio. Token bakoitzarentzat sorturiko errepresentazioa sailkatzaile honi pasako litzaioke eta honek itzulitako etiketa esleituko litzaioke tokenari.

Honelako eredu bat entrenatzean beti galdera bera egiten da: entitate klase bakoitzeko zenbat kasu behar dira entrenamendua burutzeko? Eta erantzuna beti berdina da: ahalik eta gehien, inoiz ez da nahikoa. Oso datu etiketatu gutxi edukiz gero, ziurrenik *overfitting* arazoak agertuko dira. Hau da, ereduak entrenamenduko kasu apurrak oso ondo ikasiko ditu baina ez du orokortzeko gaitasunik lortuko. Inoiz ikusi ez duen entitate bat topatzean, ez du identifikatuko.

Ataza eta corpus bakoitza mundu oso ezberdin bat da, beraz, ezin daiteke finkatu kasu kopuru minimo bat ataza guztientzat. Baina, urteetako esperientziak esaten du Hizkuntza Naturalaren Prozesamenduko atazetan zenbat eta datu gehiago eduki orduan eta emaitza

hobeak lortzen direla.

3.3.1.2 Ebaluazio metrikak

NER sistemak ebaluatzeko, hauek lortutako emaitzak pertsona aditu batek etiketatutako datuekin konparatzen dira. Konparazioa bi modu ezberdinetara egin daiteke: kointzidentzia zehatza edo erlaxatuarekin.

Kointzidentzia zehatzen ebaluazioa

Izendun Entitateen Erauzketak inplizituki bi azpi-ataza ditu: mugak hautematea eta motak identifikatzea. Hau da, entitateak non dauden aurkitzea eta entitate bakoitza zein motatakoa den esatea. Kointzidentzia zehatzen ebaluazioan, bai mugak eta baita motak ere ondo identifikatuta egotea baloratzen da [17]. Hau da, NER sistemak bueltatzen dituen tupletako ($< I_s, I_e, t >$) hiru balioak berdinak izatea eskatzen da. Sistemak itzuliriko balioak (*system*) eta adituak etiketaturiko entitateak (*gold*) kontuan izanik, faltsu positiboak (*False Positive, FP*), faltsu negatiboak (*False Negative, FN*) eta egiazko positiboak (*True Positive, TP*) zenbatzen dira.

- Faltsu Positiboa (FP): NER sistemak itzuli duen baina *gold* patroian agertzen ez den entitatea.
- Faltsu Negatiboa (FN): NER sistemak itzuli ez duen baina *gold* patroian agertzen den entitatea.
- Egiazko Positiboa (TP): NER sistemak itzuli duen eta *gold* patroian agertzen den entitatea.

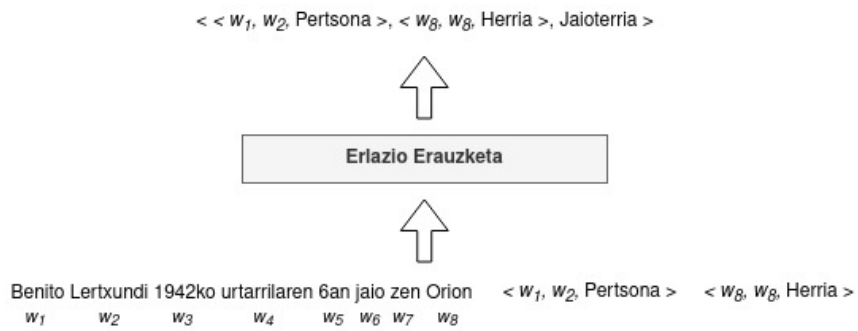
Hiru kasu ezberdin horietako agerpen kopuruak zenbatuta, doitasun eta estaldura balioak lor daitezke. Doitasun balioak, sistemak identifikaturiko entitateetatik egoki identifikatutakoen proportzioa zehazten du. Estaldura balioak aldiz, *gold* patroian definituriko entitateetatik sistemak ongi aurkitutako entitateen proportzioa. Segidan datozen hauek dira bi balio hauen formulak:

$$Doitasuna = \frac{\#TP}{\#TP+\#FP} \qquad Estaldura = \frac{\#TP}{\#TP+\#FN}$$

Bi balioak konbinatzen dituen ebaluazio metrika, hauen arteko batezbesteko harmonikoa da. Metrika berri honi *F-score* deitzen zaio. Hona hemen haren formula:

$$F\text{-}score = 2 \times \frac{Doitasuna \times Estaldura}{Doitasuna + Estaldura}$$

Entitate klase bat baino gehiago dauden kasuetan, *macro-averaged F-score* eta *micro-averaged F-score* kalkula daitezke. Hasteko, klase bakoitzeko *F-score*-a kalulatzen da eta ondoren konbinatu egiten dira. *Macro-averaged F-score*-ari dagokionez, konbinazioa batezbesteko arrunta da. Aldiz, *micro-averaged F-score*-ren kasuan batezbesteko ponderatua egiten da. Klase bakoitzari pisu ezberdin bat ematen zaio. Normalean, pisu hori klasea corpusean zein proportziotan agertzen den izaten da (klasearen kasu kopurua corpusean / kasu kopuru totala corpusean). Honela, corpusean gehien errepikatzen diren klaseei garrantzi gehiago ematen zaie eta alderantziz [16].



3.4 Irudia: Erlazio Erauzketaren ilustrazio bat

Kointzidentzia erlaxatuen ebaluazioa

Kointzidentzia erlaxatuen ebaluazioan, entitate motak ongi identifikatuta egotea baloratzen da muga zehatzak kontuan izan gabe. Hori bai, etiketatutako entitatearen eta sistemak emandako entitatearen mugak teilakatuta egotea eskatzen da. Adibidez, 3.3 irudiko adibidean, *kotxe elektriko* sailkatu beharrean *kotxe* (edo *elektriko*) ibilgailu bezala sailkatuko balu, ontzat emango litzateke. Ebaluazio metodo konplexuagoak ere proposatu dira [18]. Hauetan, mugak teilakatuta dituzten eta klase etiketa okerra duten kasuak ere ontzat ematen dira eta izendun entitateetan azpi-mota batzuk definitzen dira. Baina, metodo hauek ez dira intuitiboak eta erroreen analisia zailtzen dute. Ondorioz ez dira erabiltzen [16].

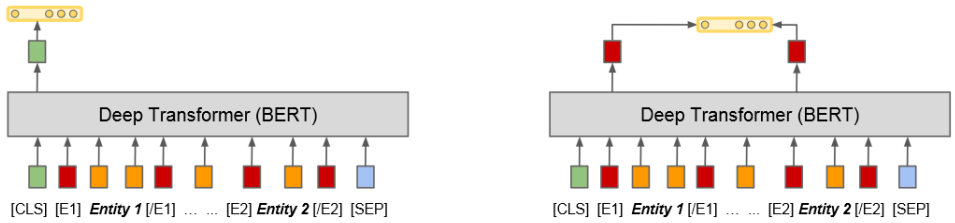
3.3.2 Erlazio Erauzketa

Erlazio Erauzketa edo ingelesez *Relation Extraction* (RE) atazak aldaera ezberdinak ditu. Sarrera moduan testu bat eta bertan agertzen diren bi izendun entitate jasotzen dira, baina, irteera ezberdinak ematen dituzten atazak defini daitezke. Batetik, sarrerako bi entitateen artean erlaziorik badagoen edo ez itzultzen duten sistemak garatu daitezke, sailkatzaile bitar gisa modelatzen direnak. Bestetik, sarrerako bi entitateen artean erlazioa existitzen dela oinarritzat hartuz, zein motatako erlazioa (aurrez definituriko klase multzo batetik hartua) den esaten duten ereduak daude. Hauek, *multiclass* sailkapeneko problema moduan garatzen dira. Azkenik, bi ataza horiek konbinatuz, hirugarren ataza bat lortzen da. Bi entitateen artean erlaziorik badagoen edo ez esatea, eta badagoen kasuetan zein motatakoa den adieraztea. Hau, *multiclass* sailkapenean oinarritzen den ereduari erlazio gabeko kasuak adierazten dituen klasea gehituta lortzen da (*no_relation*) [14].

Erlazio mota asko defini daitezke, hona hemen adibide batzuk: pertsona bat bere jaioterriarekin lotzen duen erlazioa, herrialde bat bere kapitalarekin lotzen duena, gaixotasun bat bere hasiera-datarekin lotzen duena, gaixotasun bat bere sintomarekin lotzen duena... 3.4 irudian RE atazaren adibide bat agertzen da. Bertan, *Pertsona* eta *Herria* klaseko entitateen artean *Jaioterria* motako erlazioa topatzen da.

3.3.2.1 Erlazio Erauzketa egiteko estrategiak

Erlazio Erauzketa burutzeko ere estrategia ezberdinak erabil daitezke. Horietako bat, entitate marketan (*entity marker*) oinarritutako erlazio-erauzle bat garatzea da. Laburrean azalduta, sistemari tokenizatua dagoen esaldi bat pasatzen zaio. Bertan, entitate bikotearen hasiera eta amaiera zehazten dituzten token berezi batzuk gehitzen dira (entitate markak).



(d) ENTITY MARKERS – [CLS] (f) ENTITY MARKERS – ENTITY START

3.5 Irudia: Erlazio Erauzle eredu ezberdinak

BERT eredu bati esker, token bakoitzaren errepresentazio bat lortzen da. Entitate markei dagozkien errepresentazioak erabiltzen dira erlazioa sailkatzeko. Eredu honen ideia *Matching the Blanks: Distributional Similarity for Relation Learning* artikulutik hartu da [19].

BERT ereduek, sarreran token segidak hartzen dituzte, eta token bakoitzarentzat errepresentazio bat sortzen dute. Esaldi osoaren errepresentazio bat lortu nahi bada, arrunta izaten da, sarrerako token segidaren hasieran CLS token berezi bat gehitzea. Token horri dagokion errepresentazioak nolabait esaldi guztia ordezkatzen du (azken finean token guztien informazioa jaso baitu beste errepresentazio guztiek bezala). Hasieran jartzen den moduan, amaiera adierazteko ere beste token berezi bat erabiltzen da: SEP.

Horrelako token bereziak, izatez, nahi adina eta sarrerako sekuentziako edozein tokitan gehitu daitezke. Erlazio-erauzle honetan, CLS eta SEP tokenak gehitzeaz gain, beste lau gehitzen dira: E1, /E1, E2 eta /E2. Hauei, entitate marka deitu zaie, azken finean izendun entitateak non dauden markatzen dutelako. E1 tokena lehen entitatearen lehen tokenaren aurretik jartzen da, eta /E1 lehen entitatearen azken tokenaren ostean. E2 eta /E2 tokenak bigarren entitatearekin jartzen dira modu berean. Beraz, ereduari $x = [x_0, ...x_n]$ token sekuentzia bat pasatzen zaio, non $x_0 = [CLS]$ eta $x_n = [SEP]$ diren. Honetaz gain, $x_i = [E1]$, $x_j = [/E1]$, $x_k = [E2]$ eta $x_l = [/E2]$ (edo $x_i = [E2]$, $x_j = [/E2]$, $x_k = [E1]$ eta $x_l = [/E1]$) dira, $0 < i < j - 1$, $j < k$, $k < l - 1$ eta $l < n$ izanik.

Honela, BERT eredu bati esker esaldiko token bakoitzeko errepresentazio bat edukitzeaz gain, beste sei lortzen dira. Horiek sailkatzaileen sarrera gisa erabil daitezke. Modu ezberdinetara konbina daitezke elkarrekin. Aukera bat edo beste eginda sailkatzaileak irizpide batekiko edo beste batekiko hartuko ditu erabakiak.

3.5 irudian tokenak konbinatzeko bi aukera ezberdin aurkezten dira. Ezkerreko kasuan, sailkatzaileari, CLS tokenari dagokion errepresentazioa pasatzen zaio. Nolabait esateko entitateak markaturik dituen esaldi osoaren informazioa pasatzen zaio. Aldiz, bigarren kasuan, entitateen hasierako markei dagozkien bi errepresentazioen konbinazio bat ematen zaio sailkatzaileari.

Bi metodoek emaitza onak itzul ditzakete. Baina, intuizioak esaten du bigarrenak emaitza hobeak emango dituela. Teknika horretan, bi errepresentazioen (luzera bereko bi bektore) konbinazioa egiteko aukera ezberdinak daude. *Max*, *min* eta *average pooling* estrategiak aplika daitezke. Alegia, luzera bereko hirugarren bektore bat lortzea, posizio bakoitzean gainerako bektoreetan kokapen berean dauden balioen arteko maximoa, minimoa edo batezbestekoa jasotzen duena. Hauez gain, besterik gabe bi bektoreak kateatu daitezke, luzera bikoitzeko bektore berri bat lortuz.

Irag. klasea \ Benetako klasea	A	B	C	D
A	50	3	0	0
B	26	8	0	1
C	20	2	4	0
D	12	0	0	1

3.3 Taula: Errore-matrize adibide bat

Errepresentazio hauek *feed forward* geruza bati pasata sailkapena egiten da. Geruza honek, sarreran errepresentazioen konbinazioa hartuko du eta irteera gisa erlazioa zein klasetakoa den bueltatuko du.

3.3.2.2 Ebaluazio metrikak

Azken batean, erlazio-erauzle bat *multiclass* sailkapeneko problema moduan garatzen da. Gutxienez bi klase edukiko dira (*relation* eta *no_relation*) eta gehienez nahi adina. Beraz, erlazio erauzle bat ebaluatzeko klase anitzeko sailkatzaile bat ebaluatzeko erabiltzen diren metrikak erabiltzen dira.

Hasteko, modeloak emandako emaitzak errore-matrize batean erregistratzen dira. Errore-matrizea bi sailkapenen artean konparazioa egiteko balio duen taula gurutzatu bat da. Bi horiek, sailkapen erreala (corpusean etiketatuta dagoena) eta garaturiko modeloak iragarri duena dira. Matrizean, etiketatzaileak klase bat esan duenean ereduak klase bakoitza zenbat aldiz aurreikusi duen zenbatzen da. Nahasmendurik ez sortzeko, txosten honetan zehar agertuko diren errore-matrize guztietan, zutabeek etiketaturiko klasea eta lerroek ereduaren iragarpena adieraziko dituzte.

3.3 taulan errore-matrize baten adibidea aurki daiteke. Adibide bat ematearren, bertan ikus daitekeen B zutabeko C lerroan agertzen den 2 balioak, izatez B klasekoak diren eta C gisa sailkatu diren bi kasu egon direla adierazten du. Helburua, zutabe eta lerro bakoitzeko balio altuenak diagonal nagusian (grisez markatuta dagoena) egotea da, bi sailkatzaileen artean (aditua eta garaturiko eredua) adostasuna dagoela esan nahiko baitu horrek. Dokumentu honetan topatuko diren errore matrize guztietan, zutabe bakoitzeko balio altuena grisez markatuta egongo da.

Errore-matrizeko zenbakiak erabiliz, klase bakoitzeko doitasun, estaldura eta *F-score*-a kalkulatzen dira. Klase baten doitasun balioak, klase hori esleitu zaien kasuen artean ondo sailkatu diren adibideen proportzioa adierazten du. Hona hemen formula:

$$Doitasuna(klasea = X) = \frac{TP(klasea=X)}{TP(klasea=X)+FP(klasea=X)}$$

$TP(klasea = X)$ X klasekoak diren eta X klaseko moduan sailkatu diren kasu kopurua da (errore-matrizean X zutabeko X lerroko balioa). Aldiz $FP(klasea = X)$, X klasekoak izan gabe X klasekoak direla esan den kasu kopurua (errore-matrizean X ez diren zutabe guztietako X lerroko balioen batura). 3.3 taulako adibidean B klasearen doitasun balioa honela kalkulatzen da: $Doitasuna(klasea = B) = 8/(8 + 27) = 0.23$.

Klase baten estaldura balioak aldiz, izatez klase horretakoak diren kasuen artean ongi sailkatu diren kasuen proportzioa zehazten du. Hau da bere formula:

$$Estaldura(klasea = X) = \frac{TP(klasea=X)}{TP(klasea=X)+FN(klasea=X)}$$

$FN(klasea = X)$ balio berria, X klasekoak izanda beste klase bat esleitu zaien kasu kopurua da (errore-matrizean X zutabean X ez diren lerro guztietako balioen batura). 3.3 taulako adibidean B klasearen estalduraren kalkulua: $Estaldura(klasea = B) = 8/(8 + 5) = 0.62$.

Azkenik, bi balioen arteko batezbesteko harmonikoa eginda $F\text{-}score$ balioa lortzen da. Hona hemen formula:

$$F\text{-}score(klasea = X) = 2 \times \frac{Doitasuna(klasea=X)\times Estaldura(klasea=X)}{Doitasuna(klasea=X)+Estaldura(klasea=X)}$$

3.3 taulako adibideko B klasearen *F-score*-a honela kalkulatzen da: $F\text{-}score(klasea = X) = 2 \times 0.23 \times 0.62/(0.23 + 0.62) = 0.33$.

Klase guztietako $F\text{-}score$-ak bakar batean bateratu daitezke. Batetik, *macro-averaged F-score* kalkulatzeko aukera dago balio guztien arteko batezbestekoa eginez. Bestetik, *micro-averaged F-score* lor daiteke, balio guztien arteko batezbesteko ponderatua eginez [20].

3.3.3 BRAT formatua

Izendun Entitateen Erauzketa eta Erlazio Erauzketa egiteko sortzen diren corpus asko BRAT formatuan etiketatu izan ohi dira. Testuetan entitateak eta hauen arteko erlazioak identifikatzeko aukera ematen duen tresna bat da. Testu anoatu bakoitzeko bi fitxategi ezberdin sortzen dira:

- TXT fitxategia: Testu hutsa gordetzen duen TXT formatuan dagoen fitxategia da. Bertan ez da entitaterik eta erlaziorik markatzen.
- ANN fitxategia: TXT fitxategiko testuan agertzen diren entitate eta erlazioak etiketatzeko fitxategia da (ANN formatuan dago). Informazioa lerroka gordetzen da. Lerro bakoitzean entitate edo erlazio bakar bat idatz daiteke. Hori, patroi konkretu bat jarraituz egiten da.

  Entitateen kasuan, identifikadore bat, entitatearen klasea, entitatea testuan nondik nora agertzen den (karaktereka) eta entitatea bera jartzen dira. Hona hemen pare bat adibide:

  T1 SKINRAREDISEASE 0 12 Alkaptonuria

  T2 DISEASE 23 49 genetic metabolic disorder

  ANN fitxategi berean ezin da entitate identifikadore bat behin baino gehiagotan erabili. Bestalde, azpimarratzekoa da entitate identifikadore eta klasearen artean eta entitatearen eta amaierako karaktere identifikadorearen artean tabulazio marka ($\backslash t$) erabiltzen dela eta gainerakoetan zuriune arrunt bat.

  Erlazioei dagokienez berriz, identifikadore bat, erlazioaren klasea eta erlazioaren parte diren bi entitateen identifikadoreak adierazi behar dira. Aurreko adibideko bi entitateak lotzen dituen erlazio posible bat:

R1 Is_a Arg1:T1 Arg2:T2

Ikus daitekeenez, entitateen identifikadoreak *Arg1* eta *Arg2* hitz gakoez lagunduta doaz. Ordenak garrantzia du, izan ere, erlazioaren gezia *Arg1*-etik *Arg2*-ra joaten da. Eta noski, erabilitako entitateen identifikadoreak ANN fitxategi berean definituta egon behar dute. Erlazioetan ere lehen espazioa tabulazio marka bat da eta bigarrena zuriune arrunt bat.

4.3 irudian testu oso baten ANN fitxategia ikus daiteke.

3.4 Hizkuntza Naturalaren Inferentzia

Hizkuntza Naturalaren Inferentzia, *Natural Language Inference* (NLI) gisa ezagunagoa dena, hizkuntza naturalean dagoen p premisa bat erabiliz, hizkuntza naturalean dagoen h hipotesi bat ondoriozta ote daitekeen erabakitzen duen problema da. Urteetan zehar, Adimen Artifizialean inferentzia gai garrantzitsu bat izan da, eta ikerlariek aurrerapen handiak egin dituzte dedukzio formalerako metodo automatikoen garapenean. Baina, NLIk erronka berriak proposatzen ditu: arrazoiketa formaleko kate luzeen ordez, arrazonamendu informala, ezagutza lexiko semantikoa eta hizkuntza-adierazpenaren aldakortasuna azpimarratzen dira. Ondorengo adibideak ezberdintasuna ulertzen laguntzen du:

p Inkestatutako airelinea batzuk, inflaziora doitu ondoren ere, kostuak uste baino gehiago hazten zirela ikusi zuten.

h Inkestatutako enpresa batzuk kostuen igoeren berri eman zuten.

NLI probleman, adibidea inferentzia egokitzat hartzen da. Pertsona arrunt batek p premisa entzungo balu, ziurrenik h hipotesiak jarraitzen duela onartuko bailuke. Halere, kontuan izan behar da h ez dela p-ren ondorio logiko zuzen bat. Kostuak igotzen ikusteak ez du zertan kostuen igoeren berri eman zutela esan nahi. Beharbada enpresek ez zuten kostuen inguruen informaziorik eman (estrategia moduan). Inferentzia hau onargarritzat hartzeak atazaren definizioaren informaltasuna islatzen du.

NLI problemaren bere-berezko ezaugarri bat sarrerak lengoaia naturalean idatzita daudela da. Dedukzio automatikoan egindako ikerketetan gehienetan sarrerak esanahi formala duten aurrez definitutako errepresentazio bidez adierazten direla suposatzen da. Honek ezberdintzen du NLI problema inferentzia logikotik eta honen ondorioz NLI Hizkuntza Naturalaren Prozesamenduko ataza multzoan sailkatzen da [21].

Oraingoz, premisa batekin hipotesi bat ondoriozta ote daitekeenaren eztabaidaz hitz egin da. Honek inplizituki bi klase sortzen ditu: ondoriozta daiteke eta ezin daiteke ondorioztatu. Baina, izatez hiru klase definitzen dira NLI probleman. p premisa bat eta h hipotesi bat emanda NLI sistema batek hiru erantzun hauetako bat eman dezake:

- ENTAILMENT: Emandako premisarekin hipotesia ondoriozta daitekeela adierazten du.
- CONTRADICTION: Emandako premisarekin hipotesia ezin daitekeela ondorioztatu adierazten du. Are gehiago, hipotesiak premisarekiko kontraesan bat egiten duela esaten du.

p	Mikelek 18 urte ditu eta Jonek 16.	
h	Mikel Jon baino zaharragoa da.	ENTAILMENT
h	Mikel Jon baino gazteagoa da.	CONTRADICTION
h	Mikel Pello baino zaharragoa da.	NEUTRAL

3.4 Taula: NLI problemaren hiru adibide

- NEUTRAL: Emandako premisarekin hipotesia ezin daitekeela ondorioztatu adierazten du. Baina, kasu honetan informazio faltagatik da. Alegia, baliteke hipotesia egiazkoa izatea, baina premisak ematen duen informazioarekin ezin da hori jakin.

Klase bakoitzaren esanahia hobeto ulertzeko, 3.4 taulan hiru adibide jarri dira. Premisa bakar batentzat hiru hipotesi ezberdin definitu dira. Lehen bi kasuetan zalantzarik gabe esan daiteke lehenbizikoa ondoriozta daitekeela eta bigarrena kontraesan bat dela. Hirugarrenean berriz, Pelloren adina ezezaguna denez, ezin daiteke erabakia hartu. Hori dela eta, neutrala da.

Adibidearekin jarraituz, egoera berdina adierazteko modu ugari daude. Alegia, A eta Bren arteko adin desberdintasuna adierazteko modu asko daude: *A B baino zaharragoa da*, *Ak X urte ditu eta Bk Y* (jakinda $X > Y$), *B A baino gazteagoa da*... Adierazpen guzti hauek era orokor batean ikasten dituen NLI sistema bat izango da ona.

NLI sistema batek, aplikazio ugari izan ditzake. Hasteko, nabigatzaileen bilatzaileetan bilaketa semantikoa egiteko balio du. NLI sistemei esker erabiltzaile batek *merkataritza librearen aurkako manifestazioak* bilaketa egitean *merkataritza*, *libre* edota *manifestazioak* hitz gakoak dituzten emaitzak bakarrik lortu beharrean, *Manifestariek merkataritzako oztopoak kentzeko erabakiaren aurkako esloganak oihukatu zituzten* gisako esaldiak dituzten dokumentuak ere lortuko ditu. NLI batek gaitasuna izango baitu erabiltzailearen eskaeraren eta dokumentuan dauden esaldien arteko antzekotasuna topatzeko. Bestalde, dokumentuen laburpen automatikoa egiteko balio dezake. Askotan esaldi errepikakorrak agertzen dira testuetan, eta zenbaitetan esaldi batzuk kentzeko aukera egoten da, beste esaldi batzuetatik ondoriozta daitezkeelako. NLI eredu batek esaldi horiek identifikatzen lagun dezake. Hauez gain, beste hainbat aplikazio aurki daitezke [21]. Aurrerago azalduko den moduan testuetako entitateen artean erlazioak erauzteko ere balio dezake.

3.5 *Zero-shot* eta *few-shot* ikasketak

Aurkeztu diren atazak ikasketa gainbegiratuko atazak dira. Corpus bat lortzen da, zeinetan garatu nahi den sistemak egin beharreko ataza eginda dagoen. Atazarentzat sarrera ezberdinak eta hauei dagozkien irteerak bertan egoten dira gordeta (sarrerako datuak eta klasea, testua eta entitateak, entitateak eta erlazioak...). Ikasketa algoritmo ezberdinei esker, etiketatutako datu hauek erabiliz, makinek ataza ikasten dute. Baina ikasketa hau burutzeko, normalean, aipaturiko corpusak oso handia izan behar du. Corpus bat sortzeko prozesua oso garestia da denbora aldetik, ondorioz, batzuetan ez da lortzen etiketatutako kasu bakar bat bera ere (*zero*), edo gutxi batzuk bakarrik lortzen dira (*few*). Hori gertatuz gero, makinek ikasteko beste bide batzuk aurkitu beharko dituzte.

Zero-shot ikasketa, datu etiketatu gabe ikasteko erronka da. Gizakiaren esku-hartze txiki bat inplikatzen du, eta sortzen diren ereduak aurrez entrenatutako ereduetan eta existitzen

diren bestelako datuetan oinarritzen dira. Entrenamendurako adibide etiketatuak eman beharrean, etiketatu nahi diren klase berrien deskribapen zehatza ematen zaio makinari, jada ikasita dauzkan klaseekin antzekotasunak bila ditzan. *Zero-shot* ikasketa teknikak Konputagailu Bidezko Ikusmenean, Hizkuntza Naturalaren Prozesamenduan eta Makinen Pertzepzioan erabil daitezke [22].

Suposatu artikuluak euren klasearen arabera (kirola, entretenimendua edo teknologia) sailkatzen dituen sistema bat garatu nahi dela. Teknika tradizionalei jarraituz, BERT eredu bat hartuko litzateke eta honi puntan *feed forward* geruza bat jarriko litzaioke. Corpus etiketatu bat hartuko litzateke (artikuluak eta dagokien klasea dituena) eta ereduari *fine-tuning* bat egingo litzaioke. Izendun Entitateen Erauzketa eta Erlazioen Erauzketa egiteko planteaturiko estrategia hauxe izan da. Baina, daturik eduki ezean, bide oso ezberdinak hartu behar dira. Aukera bat, zuzenean, entrenamendurik egin gabe, ereduari artikuluak eta klaseen inguruko informazioa pasatzea da.

Adibide batekin hobeto ikusten da *zero-shot* ikasketaren ideia. Honako ataza hau planteatzen da:

"Itzuli euskaratik gaztelaniara": Zer moduz zaude? -> ?

Pertsona batek, hau ikustea bakarrik nahikoa du zer egin behar duen jakiteko. Ez dauka zertan atazaren deskribapen konplexuago bat jaso beharrik, ezta egindako adibiderik ikusi beharrik ere. Esperientziarekin gauza asko ikasita ditu eta bere ezagutza nahikoa da ataza arazorik gabe burutzeko. Ez du zertan gehiago ikasi.

Adibide gehiago ere jar daitezke:

"Filma izugarri gustatu zait!-> positiboa-edo-negatiboa

Beste behin, hau irakurtzen duen pertsona batek, informazio gehiago beharrik gabe, sailkapen ataza baten aurrean dagoela ulertzen du. Bere lana, iruzkina positibo edo negatiboa den esatea da, kasu honetan positiboa izanik.

Horixe da zehazki *zero-shot* ikasketaren funtzionatzeko modua. Aurre-entrenaturiko eredu bat hartzen da ezagutza oinarri handi bat duena, testu kantitate handiak erabiliz entrenatua izan dena. Edozein atazetarako, klaseen inguruko informazio esanguratsua ematen zaio ereduari eta honek bere ezagutzan oinarrituz inferentzia egiten du.

Few-shot ikasketa, klaseko kasu kopuru txiki bat dagoenean etiketatuta egiten da. *Zero-shot* ikasketaren modu berean egiten da. Baina, hainbat kasu etiketatu daudenez, ereduari *fine-tuning* txiki bat egiten zaio. Behin eredua hobetuta, *zero-shot*-ean aplikatzen diren estrategia berberak aplikatzen dira.

Egia da eredua zenbat eta datu etiketatu gehiagoz hornitu, orduan eta emaitza hobeak lortuko direla. Gainera, batzuetan *zero-shot* eta *few-shot* estrategiek ez dute nahi bezain ondo funtzionatzen. Halere, beste zenbait egoeratan, emaitza onak lortzen dira esfortzu txiki baten truke [23].

Zero-shot eta *few-shot* ikasketak garrantzi handia dute, izan ere, mundu errealeko eszenario askotan ez da bideragarria modeloak aurkitu behar duen klase bakoitzeko datu asko bildu eta etiketatzea. Ereduek atazak datu etiketatu gutxirekin edo batere daturik gabe

burutzeko ahalmenak, testuak etiketatzeari dagozkion gastuak murrizten lagun dezake [24].

4. Materialak eta metodoak

Kapitulu honetan, Gradu Amaierako Lana burutu ahal izateko erabili diren corpusak eta metodoak aurkezten dira. Corpusak nondik datozen eta nola anotatuta dauden aurkezten da. Horretaz gain, definituriko helburuak betetzeko zein metodo konkretu bete diren aurkezten da.

4.1 Corpusak

Proiektua garatzeko bi corpus erabili dira. Lehenbizikoa, ingelesezko *MIMIC-III* corpuseko testuz osatuta dago [6]. Izatez, testu hutsak dira, baina, ataza hau egiteko etiketatuak izan dira BioDonostiako bi medikuren bitartez. Bigarrena, *RareDis* corpusa da. Domeinu bereko antzeko corpus bat. Ikerketa sakonagoak egiteko balio izan duena.

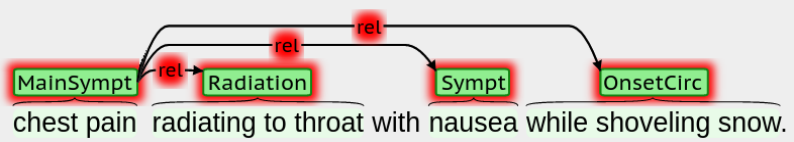
4.1.1 *MIMIC-III* corpusa

MIMIC-III ospitale tertziario handi bateko (*Beth Israel Deaconess Medical Center*, Boston) zainketa intentsiboetako unitateetan ospitaleratutako pazienteei buruzko informazioa biltzen duen ingelesezko corpus publiko potolo bat da. *MIMIC* siglak *Medical Information Mart for Intensive Care* espresioari egiten dio erreferentzia, euskaraz *Zainketa Intentsiboetarako Informazio Medikoaren Merkatua* izango litzatekeena.

Bertan informazio ugari biltzen da, beste datu askoren artean, bizi-zeinuak, medikamentuak, laborategietako neurketak, osasun-profesionalek erregistraturiko oharrak, fluido-balantzak, prozedura-kodeak, diagnostiko-kodeak, irudi bidezko diagnostiko-txostenak, ospitaleko egonaldien iraupena eta biziraupen datuak aurki daitezke. Corpusak sostengua ematen die ikerketa akademiko zein industrialari, baita kalitatea hobetzeko ekimenei eta goi-mailako irakaskuntzako ikastaroei ere [25].

Honen barnean aurki daitezkeen testuak, hainbat ataletan banaturik daude. Zati horietako bi *Chief Complaint* (CC) eta *History of Present Illness* (HPI) dira. Bere horretan euskaratuta, *Kexa Nagusia* eta *Egungo Gaixotasunaren Historia* izango lirateke. Laburrean azalduta, CC atalean bisitaren arrazoia deskribatzen duen adierazpen labur bat joaten da. Normalean, gaixoaren hitzetan idatzi ohi da. Bestalde, HPI atalean, pazientearen egungo gaixotasunaren bilakaeraren deskribapena idazten da. Deskribapen hori, kronologikoa izan ohi da, eta lehen zeinutik edo sintomatik bisita egunera arteko guztia aipatzen du. Bertan, minaren kokapena, intentsitatea, iraupena, faktore astungarri eta arintzaileak eta bestelako hainbat datu gorde behar izaten dira [26]. Alegia, 1.1 sekzioan azalduriko alarma-sintomen inguruko informazioa biltzen duen testua da.

Proiektu honetarako, bisitaren arrazoi nagusia bularraldeko mina deneko txosten klinikoak nahi ziren. Hori dela eta, *MIMIC* corpusean CC atalean bularraldeko minaren aipamenak (ingelesez *chest pain*) zituzten testuak hartu dira. Testu hauen HPI atalean, bularraldeko minari dagozkion alarma-sintomak agertu behar lukete.



4.1 Irudia: Notazio adibide bat

Behin testuak izanda, etiketatzaileek bi lan izan dituzte: entitateak etiketatzea eta erlazioak etiketatzea. Entitateei dagokienez, hamar klase definitu dira:

- *MainSympt*: Testuan agertzen diren sintoma nagusiaren aipamen guztiak sartzen dira bertan. Kasu honetan, bularraldeko minaren aipamenak. Honetaz gain, bularraldeko minari erreferentzia egiten dioten anaforak ere bai. Adibideak: *chest pain*, *the pain*...
- *Time*: Sintoma nagusia noiz hasi zen adierazten duten espresioak. Adibideak: *one week ago*, *yesterday*...
- *OnsetGrad*: Bularreko minak zein bilakaera izaten duen denboran zehar zehazten duten adierazpideak. Adibideak: *improvement and worsening*...
- *Radiation*: Bularraldeko minaren irradiazioaren inguruan informazioa ematen duten espresioak. Adibideak: *radiating to the left arm*, *nonradiating*...
- *Sympt*: sintoma nagusiaz gain, gaixoak dituen sintoma gehigarri guztiak. Adibideak: *abdominal pain*, *nausea*, *chronic cough*...
- *Duration*: Minaren iraupena zein izan zen. Adibideak: *the last two weeks*, *for one hour*, *lasting 10-15 minutes*...
- *ImprovePosNeg*: Bularraldeko mina hobetzen edo okertzen duten faktoreak. Adibideak: *by 1 minute of rest*, *with inspiration*, *nitroglycerin*...
- *OnsetCirc*: Bularraldeko mina zein momentutan agertzen den zehazten duten espresioak. Adibideak: *while walking*, *while watching TV*...
- *Localization*: Bularraldean minak duen kokapena. Adibideak: *in center of chest*, *substernal*...
- *Type*: Bularraldeko mina zein motatakoa den, mota ezberdinak existitzen baitira. Adibideak: *heaviness*, *unable to characterize*...

Erlazioen kasuan berriz, *rel* erlazioa definitu da. *MainSympt* bat eta bestelako entitate bat elkarrekin baldin badoaz, erlazio honen bidez lotzen dira.

Etiketatze guztia BRAT formatuan egin da. 4.1 irudian notazio adibide txiki bat agertzen da. Argazkian lau entitate ezberdin ikus daitezke etiketaturik: *MainSympt* bat eta hiru ezaugarri. Hiru ezaugarriak sintoma nagusiaren inguruko informazioa ematen dutenez, *rel* motako erlazioak aurki ditzakegu sintoma nagusi eta gainerako hiru datuen artean.

Egia esan, corpusak ez du aldaketarik izan 2022ko abenduaz geroztik. Hori dela eta 2.3 sekzioan azaldu bezala, 20 testu etiketatu bakarrik daude. Testu bakoitzak bi bertsio

	1 etiketatzailea			2 etiketatzailea		
Klasea	Kopurua	Erl. kop.	Dist.	Kopurua	Erl. kop.	Dist.
MainSympt	60	60	-	58	58	-
Time	11	8	1.13	20	9	0.89
OnsetGrad	2	1	0	0	0	0
Radiation	8	8	1.13	6	6	1
Sympt	92	73	5.44	88	58	6.21
Duration	5	3	1.33	7	4	2.75
ImprovePosNeg	18	16	4.44	20	15	4.33
OnsetCirc	11	9	1.11	13	9	0.44
Localization	13	9	0.78	16	11	0.45
Type	5	3	0.67	11	8	0.63
Guztira	225	190	1.78	239	178	2.09

4.1 Taula: *MIMIC-III* corpuseko kontaketak

ezberdin ditu, anotatzaile bakoitzak bere kabuz etiketatu baititu. Ez dago bertsio bateratu bat. Bi etiketatzaileen arteko adostasuna neurtu da eta %70 baino apur bat baxuagoa da.

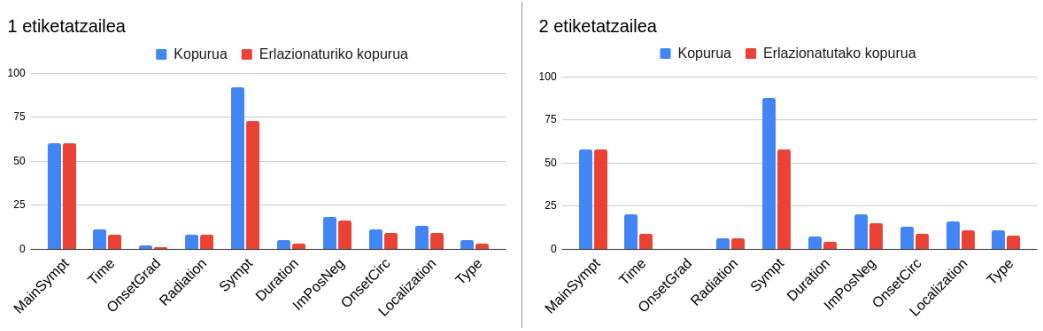
Corpusaren inguruko datu gehiago izatearren, hainbat kontaketa egin dira. Etiketatzaile bakoitzak klase bakoitzeko zenbat entitate identifikatu dituen zenbatu da. Gainera, topaturiko entitate bakoitza beste entitateren batekin erlazionatu duen ere kontatu da. Honela, identifikatu diren entitateen artean bularraldeko minarekin lotuta zeudenak zein diren jakin daiteke. Bestalde, erlazionaturikoen kasuan, bi entitateen artean egon ohi den batezbesteko distantzia esaldika ere neurtu da. Bi entitateak esaldi berean egonez gero, distantzia 0 da. Ondoz ondokoan egonez gero 1 eta bien esaldien artean beste esaldi bat egonez gero 2.

4.1 taulan topa daitezke kontaketa guztiak. Ikus daitekeen moduan topaturiko entitate kopuruaren eta erlazionaturikoen arteko aldea ez da oso handia. Etiketaturiko entitate gehienak bularreko minarekin erlazionaturik daude. Hori bai, klaseko agerpen kopuruen artean desoreka handia dago. Argi eta garbi *MainSympt* eta *Sympt* klaseak besteekiko gailentzen dira. Askoz kasu gehiago dituzte. Hau gutxi balitz, klase askotan ia ez dago kasurik, adibidez OnsetGrad klasean (2. etiketatzaileak ez du kasurik topatu). Desoreka hauek 4.2 irudiko barra diagrametan islatzen dira. Honetaz gain, ikusten da klase guztiak distantzia ezberdinetara egon ohi direla. *OnsetGrad* klasea adibidez, beti esaldi berean aurkitzen da. Baina, *Sympt* klasea distantzia handiagoetara agertzen da.

Beraz, nahiko deskonpentsatuta dagoen eta gainera kasu etiketatu gutxi dituen corpus bat da honako hau.

4.1.2 *RareDis* corpusa

RareDis gaixotasun arraroak eta hauen zeinu eta sintomak etiketaturik dituen corpus bat da. Gaixotasun arraroak populazioaren zati txiki batek bakarrik izaten dituen gaixotasunak dira. Halere, 6000 gaixotasun arraro baino gehiago identifikatu dira eta hauek munduko 300 milioi biztanle baino gehiagori eragiten die. Hauen inguruan oso informazio gutxi egon ohi denez, hauen diagnosiak asko atzeratzen dira. Horregatik, hauen inguruko informazio gehiago eskuratzeko beharra dago. Hizkuntza Naturalaren Prozesamenduko teknikak



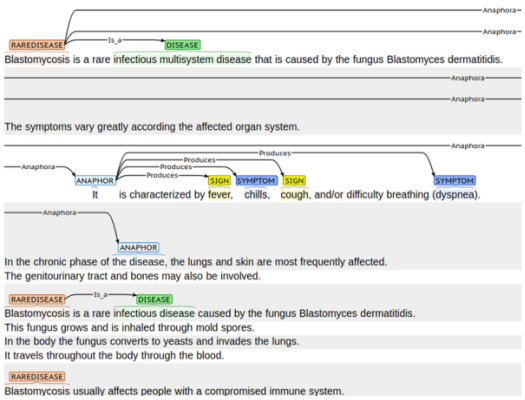
4.2 Irudia: *MIMIC-III* corpuseko kontaketen barra diagramak

erabiliz, ezagutza hau areagotu nahi da. Horretarako, corpus hau etiketatu da. Helburua, bertan eredu ezberdinak entrenatzea da, ondoren testuetatik gaixotasun arraroen inguruko informazioa erauzteko [27].

RareDis corpusean dauden testuak, Gaixotasun Arraroen Erakunde Nazionalak (*National Organisation for Rare Diseases, NORD*) sorturiko gaixotasun arraroen corpusetik daude hartuta. Corpus horrek, 1200 gaixotasun arraro baino gehiagoren inguruko informazioa gordetzen du. Gaixotasun bakoitzeko, honako sekzio hauek dituen testu bat dago bertan: eztabaida orokorra, zeinuak eta sintomak, kausak, kaltetutako populazioak, lotutako nahasmenduak, diagnostikoa, terapia estandarrak, ikerketa-terapiak, NORDen parte diren erakundeak eta beste erakunde batzuk. *RareDis* corpusa sortzeko, lehen zazpi sekzioak hartu zituzten [27].

Aurreko corpusaren antz handia du, izan ere, aipaturiko testuetan entitate mota ezberdinak eta entitate hauek elkarrekin lortzen dituzten erlazioak topa ditzakegu. Entitateei dagokienez, sei klase ezberdin topa daitezke:

- *DISEASE*: Organo, sistema edo organismo baten egoera anormala, hainbat arrazoirengatik sortua, hala nola infekzioagatik, hanturagatik, ingurumen-faktoreengatik edo akats genetikoagatik. Zeinu, sintoma edo bien multzo baten bidez identifikagarria izan ohi dena. Adibideak: *cancer, alzheimer, cardiovascular disease...*
- *RAREDISEASE*: Populazioaren zati txiki batek bakarrik izaten dituen gaixotasunak. Europan gaixotasun bat arrarotzat hartzen da 2000 biztanleko pentsona bat baino gutxiagori eragiten badie. Adibideak: *acquired aplastic anemia, Fryns syndrome, giant cell myocarditis*
- *SKINRAREDISEASE*: Larruazaleko gaixotasun arraroak dira. Adibideak: *cold urticaria, progeria, Werner syndrome...*
- *SYMPTOM*: Pertsona batek jasan dezakeen arazo fisiko edo psikologiko bat, gaixotasun baten presentzia adierazi dezakeena. Ezin dira ikusi eta proba medikoetan ez dira agertzen. Gaixoak subjektiboki adierazten dituen aztarnak dira. Adibideak: *fatigue, dyspnea, pain...*
- *SIGN*: Pertsona batek gaixotasun bat izan dezakeela adieraz dezakeen eta azterketa fisiko edo proba baten emaitza gisa agertu den zerbait. Adibideak: *inflammation, rash, abnormal heart rate, hypothermia...*




```
T1      RAREDISEASE 0 13        Blastomycosis
T2      DISEASE 24 54   infectious multisystem disease
T3      ANAPHOR 173 175 It
T4      SIGN 196 201    fever
T5      SYMPTOM 203 209 chills
T6      SIGN 211 216    cough
T7      SYMPTOM 247 254 dyspnea
T8      ANAPHOR 281 292 the disease
T9      RAREDISEASE 399 412     Blastomycosis
T10     DISEASE 423 441 infectious disease
T11     RAREDISEASE 662 675     Blastomycosis
R1      Is_a Arg1:T1 Arg2:T2
R2      Produces Arg1:T3 Arg2:T4
R3      Produces Arg1:T3 Arg2:T5
R4      Produces Arg1:T3 Arg2:T7
R6      Produces Arg1:T3 Arg2:T6
R7      Is_a Arg1:T9 Arg2:T10
R5      Anaphora Arg1:T1 Arg2:T3
R8      Anaphora Arg1:T1 Arg2:T8
```

4.3 Irudia: BRAT bidez anotatutako testu baten adibidea

- *ANAPHOR*: Aurrerago aipatu den gaixotasun edo gaixotasun arraro bati erreferentzia egiten dioten izenordain, hitz edo esapidea. Adibideak: *this disease, these diseases, it...*

Bestalde, erlazioak definitzeko, beste sei klase ezberdin erabiltzen dira:

- *Produces*: Edozein gaixotasun eta honek eragindako edozein sintoma edo zeinuren arteko erlazioa.
- *Increases_risk_of*: Gaixotasun baten eta nahasmendu baten arteko erlazioa, non gaixotasunak nahasmendu hori izateko probabilitatea handitzen duen.
- *Is_a*: Gaixotasun baten eta hau aipatzeko modu orokorrago baten arteko erlazioa.
- *Is_acron*: Akronimo baten eta bere forma oso edo luzearen arteko erlazioa.
- *Is_synon*: Gaixotasun bera definitzen duten bi izenen arteko erlazioa.
- *Anaphora*: Anaphora baten eta bere aurrekariaren arteko erlazioa. Aurrekariak gaixotasun edo gaixotasun arraro bat izan behar du.

4.3 irudian corpuseko testu etiketatu bat ikus daiteke BRAT formatuan. Bertan, hainbat entitate eta erlazio agertzen dira.

Momentuz, *RareDis* corpuseko train eta dev zatiak bakarrik daude publikaturik. Corpusaren egileak hainbat ikerketa burutzen ari omen dira eta hauek amaitu ostean argitaratuko omen dute test zatia. Bitartean, train zatiko 1459 testu eta dev-eko 209 testu atzitu daitezke.

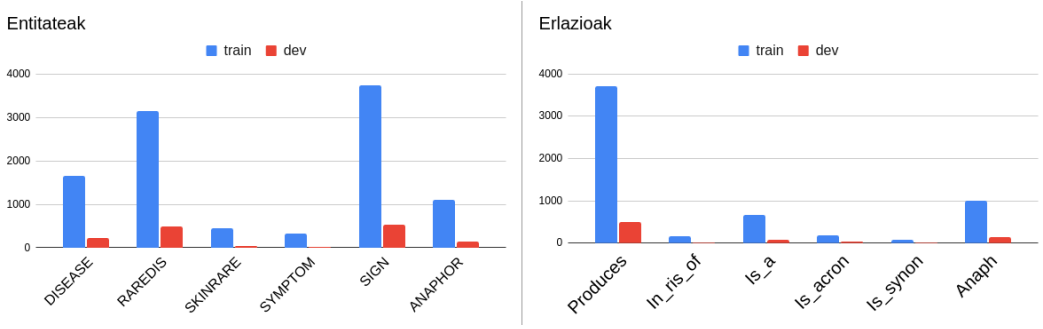
Corpusaren informazio gehiago eduki ahal izateko, entitate eta erlazio klase bakoitzeko zenbat kasu dauden zenbatu da. Kontaketa hauek 4.2 tauletan daude.

Taula horietan bilduriko datuak oso ezberdinak dira aurreko corpuseko zenbaketekiko (4.1 taula). Oraingo taulan ikus daitekeenez, entitate eta erlazio bakoitzeko askoz ere kasu gehiago zenbatu dira. Halere, hemen ere desoreka puntu bat agertzen da. Entitateen kasuan, bistakoa da klase errepikatuenak *RAREDISEASE* eta *SIGN* direla. Baina, gainerako klaseen artean oreka onargarria dago. Aldiz, erlazioen kasuan, gehienetan topa daitekeena *Produces* klasea da, gainerakoak baino askoz gehiagotan errepikatzen dena. Beste klase guztiak nahiko orekaturik daudela esan daiteke.

Entitateak		
Klasea	train	dev
DISEASE	1647	230
RAREDISEASE	3157	480
SKINRAREDISEASE	451	45
SYMPTOM	319	24
SIGN	3744	528
ANAPHOR	1108	151
Guztira	10426	1458

Erlazioak		
Klasea	train	dev
Produces	3717	499
Increases_risk_of	154	21
Is_a	657	81
Is_acron	182	33
Is_synon	77	16
Anaphora	999	138
Guztira	5786	788

4.2 Taula: *RareDis* corpuseko entitateen eta erlazioen kontaketak



4.4 Irudia: *RareDis* corpuseko kontaketen barra diagramak

Aipaturiko desoreka hauek, hobeto hautematen dira 4.4 irudiko barra diagrametan. Barren arteko luzeren ezberdintasunei dagokienean, antzekotasunak nabarmenak dira 4.2 irudikoekin. Baina ardatz bertikaleko balioei erreparatuz gero, oraingo corpusean dauden kasu kopuruak askoz handiagoak direla eztabaidaezina da. Ezaugarri horrek ematen dio aberastasuna corpus honi. Test zatia oraindik plazaratu gabe egonagatik ere, baliabide interesgarria da corpus hau.

4.2 Metodoak

Proiektu honetan, bi azpi-ataza bereizten dira. Alde batetik, testuetan alarma-sintomei dagozkien entitate guztiak topatu behar dira (*MainSympt* eta gainontzeko guztiak). Beste aldetik, *MainSympt* entitateekin zein alarma-sintoma dauden erlazionaturik erabaki behar da.

Hau honela izanik, bi eratara plantea daiteke ebazpena. Batetik, bi lanak burutzen dituen modelo bakar bat garatu daiteke. Hau eginez gero, ereduak gaitasuna eduki beharko luke entitateak identifikatzeaz gain, bularraldeko minarekin erlazionaturik ez dauden alarma-sintomak baztertzeko. BIO notazio bidez, MainSympt-arekin bakarrik lotuta dauden entitateak etiketatu daitezke. Edo denak etiketatu daitezke eta MainSympt-arekin erlazionaturikoei marka berezi bat gehitu dakieke. Baina, hori guztia ikasteko sistemak ahalmen handia eduki beharko luke.

Bestetik, ataza bakoitza burutzeko eredu ezberdin bat garatu daiteke. Lehenbizikoak, testuan agertzen diren *MainSympt* eta alarma-sintoma guztiak topatu beharko lituzke. Alarma-sintoma hauen artean, sintoma nagusiarekin erlazionaturikoak eta erlazionatu

gabeak egongo lirateke. Bigarren ereduak, lehenak aurkitu dituen entitate guztien artean bularraldeko minarekin zein doazen erabaki beharko luke. Kontuan izan behar da, lehen ereduak akatsak egingo dituela, posible izango baita izatez bularraldeko minarekin erlazionatu gabe dauden entitateak ez topatzea. Akats hauek bigarren ereduari pasako dizkio, eta honek, bere aldetik akats gehiago egingo ditu (eredu perfekturik ez baita existitzen). Beraz, aukera honek akatsak metatzeko arriskua du.

Bi estrategia horien artean bigarrena hartzea erabaki da. Erroreak metatzeko arriskua egonagatik ere, bi atazak bereizita ikasteak biak batera ikasteak baino errazagoa baitirudi. Bi eredu sinpleen errore metatua eredu konplexuaren errorea baino txikiagoa izango dela pentsatzen da.

Hori kontuan izanik, lehenbizi entitate-erauzle bat garatu behar izan da. Bi arazo ikusi dira. Batetik, datu etiketatu oso gutxi daude. Bestetik, klase kopurua handia da. Beraz, eredu bat garatzea zaila da. Hori egin beharrean, publikatuta dauden eredu ezberdinak ikertu dira.

Entitate-erauzlearen ostean erlazio-erauzlea garatu behar izan da. Beste behin, kasu etiketatu kopurua murritza da. Hori dela eta, teknika ezagunagoak albo batera utzi eta erlazioak erauzteko inferentzian oinarrituriko sistema bat garatu da.

Behin ereduak edukita, erabilitako metodoak ebaluatu dira. Hau ez da erraza, izan ere datu etiketatu kopuruak txikiak dira. Bai zerbait ikasteko, baina baita metodoak ebaluatzeko ere. Kasu gutxiren gainean emaitza onak bueltatzeak ez du orokortzeko gaitasunik erakusten. Garaturiko eredua benetan ona dela frogatu nahi izan da.

Horren aurrean, *RareDis* corpusa hartu da eta honen gainean entrenatzeko eszenario txikiak simulatu dira. Eszenario hauetan inferentzian oinarrituriko sistema probatu da erlazioak erauzteko. Probak test handi batean eginda, hauen orokortzeko gaitasuna ebaluatu da.

Azkenik, erlazioak erauzteko hizkuntza eredu aurre-entrenatuetan oinarrituriko modeloak probatu dira sorturiko eszenario txikietan. Hauek lorturiko emaitzak aztertu dira eta eredu hauen eta inferentzian oinarriturikoaren artean konparaketa bat egin da.

5. Proiektuaren garapena

Kapitulu honetan proiektuaren garapen prozesu guztia azaltzen da. 2. kapituluan bildutako helburuak lortu ahal izateko jorraturiko bidea alegia. Hartutako erabaki, erabilitako tresna eta garaturiko modelo guztien inguruko informazio guztia topa daiteke bertan. Atal hau da proiektuaren mamia.

5.1 Entitate-erauzlea

Entitate-erauzle bat garatu ahal izateko, datu etiketatu ugari behar dira. Zenbat eta datu gehiago eduki, orduan eta emaitza hobeak lortuko dira. Baina, proiektu honetan, datu etiketatuen falta sumatu da. *MIMIC-III* corpusa oso txikia da. 4.1 taulako datuetara itzuliz, agerpen kopuruak oso txikiak dira eredu batek zerbait ikasi ahal izateko. *MainSympt* eta *Sympt* klaseetan beharbada zerbait ikasteko adina datu badago, baina, gainerakoekin ezer gutxi egin daiteke.

Datu gutxi edukitzeaz gain, klase kopurua handia da. Beraz, eredu bakar bati erlazio guztiak ikastea asko kostako zaiola pentsatu da. Ondorioz, entitate-erauzle bakar bat egin beharrean lana entitate-erauzle ezberdinen artean banatzea erabaki da; bakoitza klase kopuru txikiago batean espezializatu dadin.

Halere, datu gutxi edukitzearen arazoak presente jarraitzen du. Zaila da eredu sendo bat garatzea. Baina, Hizkuntza Naturalaren Prozesamenduan medikuntza dezente dago ikertuta. Arlo honetako entitate-erauzle ezberdinak daude publikatuta. Hauek corpus honetako entitateak ongi erauzten ote dituzten probatu da. Corpusean etiketaturiko kasu guztiek test gisa funtzionatu dute.

Ondorengo azalpenetan, klase bakoitzeko erabili diren tresna ezberdinak aurkezten dira.

5.1.1 MainSympt eta Sympt

Bi klase hauek medikuntzako testuetan ager daitezken sintoma arruntak dira. Sintoma erauzle ezberdinak ikertu dira eta bi probatu dira: *cTAKES* eta *MetaMap*.

5.1.1.1 *cTAKES*

cTAKES Hizkuntza Naturalaren Prozesamenduko kode irekiko sistema bat da historia kliniko elektronikoetako testu ez egituratutik informazio klinikoa erauzteko balio duena. Ohar klinikoetan mota ezberdinetako izendun entitateak identifikatzen ditu: drogak, gaixotasunak/nahasmenduak, zeinuak/sintomak, kokapen anatomikoak eta prozedurak. Izendun entitate bakoitzeko hainbat atributu identifikatzen dira: entitateari dagokion testua, ontologiaren kodea, testuingurua eta ezeztatuta edo ezeztatu gabe dagoen [28].

Honek topatzen dituen klaseetatik, zeinu eta sintomei dagozkienak dira proiektu honetan interesekoak. Badirudi kokapen anatomikoenak ere lagun dezakeela *Localization* klasea

aurkitzen. Baina, *cTAKES*-ek topatzen dituen kokapenak gorputzeko atal ezberdinak dira (burua, bularraldea, abdomena...), eta proiektu honetan bularraldeko kokapen ezberdinak topatu nahi dira.

Egia esan, tresna hau probatzen hainbat saiakera egin dira, baina interfaze nahiko konplexuak ditu. Honetaz gain, ez da lortu kode bidez era erraz batean atzitzea. Hori dela eta, alde batera utzi da.

5.1.1.2 *MetaMap*

UMLS (*Unified Medical Language System*) fitxategi eta software multzo bat da hiztegi biomediko ugari integratzen eta bateratzen dituena [29]. *MetaMap* berriz, UMLSn aurki daitezkeen ontologiak testuetan identifikatzen dituen software erreminta bat da. Honek, teknika sinbolikoak, Hizkuntza Naturalaren Prozesamendukoak eta hizkuntzalaritza konputazionalekoak erabiltzen ditu. Guztira 127 klase ezberdinetako entitateak topatzen ditu. Hauen artean, *sosy* klasea (*Sign or Symptom*) topa daiteke, zeinu eta sintoma ezberdinei dagokiena [30].

Entitateen detekzioa bi fasetan bereizten du. Hasteko, testuan aipatzen diren eta UMLSn agertzen diren termino guztiak identifikatzen ditu. Adibidez, testuan *chest pain* azaltzen bada, hiru termino ezberdin identifikatuko ditu *chest*, *pain* eta *chest pain*. Bigarren fasean, testuinguruaren arabera testuan izatez aipatu diren terminoak zein diren erabakitzen du. Aurreko adibidearekin jarraituz, une honetan *chest pain* hartuko luke eta gainerako biak baztertu egingo lituzke.

MetaMap izatez Java programazio lengoaian dago idatzita. Baina, Anthony Rios-ek Python-en *wrapper* bat sortu zuen: *PyMetaMap* [31]. Honek, *MetaMap* era erraz batean erabiltzea ahalbidetzen du. Horrekin, testuetan agertzen diren entitateen mota, UMLSn duten CUI kodea eta testuan zein posiziotan agertzen diren lor daiteke.

Behin zeinu eta sintomak topatuta erraza da *MainSympt* eta *Sympt* klaseen arteko bereizketa egitea. Ingelesezko testuetan, bularraldeko minaren aipamenak *chest pain* edo *CP* gisa agertzen dira. Beraz, aurkitu diren sintomen artean itxura hori dutenak *MainSympt* bezala sailkatzen dira eta gainerakoak *Sympt* bezala.

5.1.2 Time eta Duration

Klase hauek izatez espresio tenporalak dira, eta espresio tenporalak erauzten dituzten sistema asko daude publikatuta. Horietako bat *SUTime* da. *SUTime* testuetan denborazko espresioak aurkitzen eta normalizatzen dituen liburutegi bat da. Hau da, *next wednesday at 3pm* gisako espresioak topatu eta *2016-02-17T15:00* gisa irudikatzeko gaitasuna du. Erregeletan oinarritutako sistema determinista bat da eta hedagarria izateko diseinatuta dago [32].

Denborazko espresioak lau klase ezberdinetan sailkatzen ditu: *DATE*, *TIME*, *DURATION* eta *SET*. Proiektu honetako *Time* klaseak datak eta denborak adierazten dituzten espresioak (*09-09-2022*, *yesterday*, *one week ago*...) biltzen ditu. Honelako entitateak *DATE* eta *TIME* klaseetan biltzen dira. Bestalde, *MetaMap*-en *DURATION* klasea eta hemen definituriko *Duration* bat datoz.

5.1.3 Radiation

Radiation klasearentzat ez da topatu entitate-erauzle publikaturik. Baina, klaseko kasuak aztertuta guztietan patroi bat topatu da. Beti, *radiate*, *radiation* edo *nonradiating* hitzen hitz eratorriren bat topa daiteke. Hau jakinda, estrategia sinple bat jarraituta, klase honetako entitateak erauzten dituen programa bat garatu da.

Zehazki, lematizazioan (3.1 sekzioa) oinarritzen den programa bat sortu da. Hasteko, testu guztia lematizatzen du. Ondoren, lema guztien artean *radiate*, *radiation* edo *nonradiating* lemak bilatzen ditu. Hauetakoren bat topatuz gero, lemari dagokion berezko testua eta testuan dituen kokapena lortzen eta itzultzen ditu.

Programa honek topatzen dituen entitateak ez datoz guztiz bat etiketatutakoekin. Izan ere, *radiating to both shoulders* erako kasu bat etorriz gero, *radiating* bakarrik topatuko du. Baina, behintzat, testuan irradiazioaren inguruan informazioa eman dela jakiteko balio du. Beraz, ontzat ematen da (Kointzidentzia erlaxatuen ebaluazioa, 3.3.1.2 sekzioa).

Lematizazioa egiteko, *EHRKit* liburutegiko *get_lemmas* funtzioa erabili da. *EHRKit* testu klinikoak aztertzeko *python* liburutegi bat da. Bi zatitan banatzen da. Alde batetik, *MIMIC-III* corpusa atzitzeko interfaze multzo bat eskaintzen du. Beste aldetik, hirugarren pertsonen liburutegi bilduma bat gordetzen du HNPko 12 ataza ezberdin burutzeko: izendun entitateen erauzketa, testu-laburketa, itzulpen automatikoa... Testuen lematizazioa bigarren atalean sailkatzen da [33].

5.1.4 Gainerako klaseak

Gainerako klaseak oso espezifikoak dira. Zaila da klase horiek identifikatzeko gai den eta argitaratuta dagoen entitate-erauzle bat aurkitzea. Honetaz gain, *Radiation* klasekoek ez bezala, klase horietako kasuek ez dute eredu konkretu bat jarraitzen, elkarrengandik nahiko ezberdinak dira. Beraz, ezin daiteke lematizazioan oinarrituriko estrategia berezirik aplikatu. Eta nola ez, eredu tradizional bat entrenatzeko nahiko kasurik ere ez dago.

Gainera, proiektu honetan zailtasun nagusia erlazioak erauztean dago. Hau da, alarma-sintomak izatez sintoma nagusiarekin erlazionaturik dauden edo ez jakitean. Hori dela eta, entitate erauzketa honela uztea eta ikerketa erlazio erauzketara bideratzea erabaki da.

5.2 Erlazio-erauzlea

Erlazio-erauzleak sarrera moduan testuinguru bat eta bertan aipatzen diren bi entitate hartzen ditu. Entitateetako batek *MainSympt* klasekoa izan behar du eta besteak alarma-sintoma bat. Erlazio-erauzleak gai izan behar du entitateak erlazionaturik dauden edo ez esateko (sailkapen bitarra).

Ataza hau ikasteko eskuragarri dauden datu etiketatu kopuru murritzak bultzatuta, erlazio-erauzleak garatzeko teknika tradizionalak albo batera utzi eta *zero-shot* eta *few-shot* teknikak ikertzea erabaki da (3.5 sekzioa). Tresna eta liburutegi ezberdinak daude horrelako teknikak aplikatzeko. Proiektu honetan, *Ask2Transformers* (A2T) liburutegiaren gaineko ikerketa bat egin da esperimentu ugari burutuz.

5.2.1 *Ask2Transformers* liburutegia

Ask2Transformers liburutegia, *zero-shot* teknikak aplikatuz testuetan Informazio Erauzketa egiteko tresna bat da. Hau da, entrenamendurako datu-etiketaturik ez dagoenean izendun entitateen erauzketa eta erlazio erauzketa egiteko bideak eskaintzen ditu. Kasu honetan, erlazio erauzketarako erabili da.

Ask2Transformers liburutegiaren funtsa NLI problema da (3.4 sekzioa). Adibide batekin, ongi ulertzen da atzean duen ideia. Hona hemen testu posible bat:

Gaixoak bularraldeko mina sentitu zuen astelehenean eta atzo eztula hasi zitzaion.

Bertan, *MainSympt* bat (*bularraldeko mina*) eta *Time* klaseko bi entitate (*astelehenean* eta *atzo*) aurki daitezke. Hiru entitate horiek erabilita, bi esaldi posible hauek sor daitezke:

1. *Bularraldeko mina astelehenean hasi zen.*
2. *Bularraldeko mina atzo hasi zen.*

NLI sistema bati p premisa moduan goiko testua eta h hipotesi gisa lehen esaldia pasako balitzaizkio, honek premisarekin hipotesia ondoriozta daitekeela esango luke (*ENTAILMENT*). Ondorioz, *bularraldeko mina* eta *astelehenean* entitateen artean erlazioa dagoela esango litzateke. Aldiz, hipotesitzat bigarren esaldia emango balitzaio, premisaren eta hipotesiaren artean kontraesan bat dagoela esango luke (*CONTRADICTION*). Beraz, *atzo* entitatea ez doala sintoma nagusiarekin ondorioztatuko litzateke.

Funtsean, horixe da A2Tk egiten duena. NLI sistema aurre-entrenatu bat edukitzen du kargaturik. Honetaz gain, erlazio klase posible bakoitzeko txantiloi multzo bat edukitzen du definiturik (programatzaileak eginak). Aurreko adibidera itzuliz, bi txantiloi posible *X Y hasi zen* eta *Xren hasiera-data Y da* dira. Bestalde, erlazio bat baliozkoa izateko baldintzen definizioa du gorderik (hau ere programatzaileak egina). Erlazio klase bakoitzeko, erlazio hori zein motatako entitateen artean eman daitekeen adierazteko balio dute baldintza hauek. Aipatu beharra dago, klase ezberdinetan baldintza berdin bat definitzeko aukera dagoela. Bi baldintza posible *MainSympt:Time* eta *MainSympt:Sympt* dira.

Informazio hori guztia edukita, erlazioak sailkatzeko gai da. Sarrera gisa, testuinguru bat (*string* bat), bertan agertzen diren bi entitate (bi *substring*) eta baldintza bat (entitateen klaseen informazioa) jasotzen ditu. Testuingurua premisatzat hartzen du. Aldiz, adierazitako baldintza betetzen duten klaseetan definitutako txantiloi bakoitzeko, hipotesi bat definitzen du. Konkretuki, X eta Y agertzen diren lekuetan bi entitateak ordezkatzen ditu. NLI sistemari testuingurua eta hipotesiak ematen dizkio. Honek, hipotesi bakoitzeko, *ENTAILMENT* klasearen konfiantza-maila ($[0, 1]$ tarteko balio bat) ematen dio. Hauek jasota, konfiantza-maila handiena duen hipotesiarekin gelditzen da. Maila hau programatzaileak definituriko atalase bat (lehenetsia 0.5) baino txikiagoa bada, bi entitateen artean erlaziorik ez dagoela itzultzen du. Bestela, hipotesiari dagokion klasea esleitzen dio erlazioari.

Beraz, programatzaileak hiru ataza egin behar ditu. Batetik, NLI sistema aurre-entrenatu bat aukeratu behar du. Eredu ugari topa ditzake *Hugging Face* webgunean [13]. Bestetik, klase bakoitzeko erlazioa baliozkoa izateko baldintzak definitu behar ditu. Azkenik, klase bakoitzeko txantiloiak prestatu behar ditu. Esaldi sinpleak jartzea gomendatzen da. Hobe da klase batentzat txantiloi korapilatsu bat jarri baino, hiru sinple jartzea. Erabiltzen diren txantiloiek sinpleak izateaz gain oso esanguratsuak izan behar dute. Benetan erlazio bat

sailkatzeko ezagutza eman behar dute. Hiru pausu sinple horiek jarraituta, erlazio erauzlea garaturik egongo litzateke [34] [35] [36] [37].

5.2.2 Zero-shot *MIMIC-III* corpusean

Ask2Transformers tresna *MIMIC-III* corpusean probatzea pentsatu da. Horretarako, hainbat erabaki hartu behar izan dira.

Lehenbizikoa, erlazioen klaseen definizioa izan da. Alegia, zenbat erlazio klase eta zein ezarriko diren. Bi definizio posible identifikatu dira:

1. **Erlazio klase bat**: 4.1.1 sekzioan ikusi den moduan, *MIMIC-III*-ko testuetan erlazio mota bakar bat agertzen da (*rel*). Beraz, aukeretako bat A2Tn bi klase bakarrik definitzea da: *no_relation* eta *rel*. *Rel* klaseko baldintzetan *MainSympt:X* posible guztiak sartuko lirateke, X edozein alarma-sintoma izanik. Honetaz gain, klase honentzat txantiloi zerrenda bat sortuko litzateke.

2. **Erlazio klase ugari**: Erlazio guztiak klase berean sartu beharrean, alarma-sintomaren arabera banandu daitezke. Hau da, *no_relation* klaseaz gain, bederatzi erlazio klase defini daitezke (alarma-sintoma bakoitzeko bat). Hau honela eginda, klase bakoitzak *MainSympt:X* itxurako baldintza bakar bat izango luke, X klaseari dagokion alarma-sintoma izanik. Bestalde, klase bakoitzeko txantiloi gutxi batzuk idatziko lirateke.

Bigarren aukera egitea erabaki da. Izan ere, alarma-sintoma ezberdinentzat sortu beharreko txantiloiek ezer gutxi dute amankomunean. Klaseak oso ezberdinak dira elkarren artean. Beraz, bereizita kontrolatu nahi izan dira. Honela eginda, badirudi ataza klase anitzeko sailkapen batean bilakatzen dela. Baina, oinarrian sailkapen bitar bat izaten jarraitzen du. Esan bezala, klase bakoitzean baldintza bakar bat dago eta ez da errepikatzen klaseen artean. Beraz, kasu berri bat etortzen denean klase bakar bateko baldintza beteko du. Ondorioz, kasuarentzat bi iragarpen bakarrik egin ahal izango dira: klase hori edo *no_relation*. Hortaz, funtsean bi aukerak berdinak dira, ezberdintasun bakarra dute: *MainSympt* eta alarma-sintoma baten artean erlazioa ote dagoen erabakitzeko, lehen aukerak txantiloi guztiak erabiltzen ditu eta bigarrenak alarma-sintomarentzat espresuki sortuak. Bigarrenean egoera hobeto kontrolatzen denaren irudipena dago.

Bigarren erabakia, sortu beharreko txantiloiena izan da. Klase bakoitzeko, hainbat txantiloi sinple sortu dira. Helburua txantiloi hauen bidez klasea asmatzea denez, ahalik eta bereizgarrienak izateko ahalegina egin da. Adibide batzuk aipatzearren, *Duration* klasean, X minak Y denbora iraun zuela ondorioztatu behar da. Hori esateko, ingelesez bi aditz ditugu: *durate* eta *persist*. Beraz, bi txantiloi hauek sor daitezke:

1. *X durated Y*
2. *X persist for Y*

Sympt klaseari dagokionez berriz, modu asko daude esateko X gaixotasunaren sintoma dela Y. Zuzenean, hori esan daiteke, edo X-k Y barne hartzen duela ere esan daiteke. Hona hemen bi adibide:

1. *X is a symptom associated with Y*

2. *X includes Y*

Klaseen gehiengoari hiruzpalau txantiloi sortu zaizkio. Baina, *ImprovePosNeg* klasea bereziki zaila dela ikusi da eta honetan txantiloi gehiago erabiltzea erabaki da. Erabilitako txantiloi guztiak ezagutu nahi izanez gero, A eranskinean topatu daitezke.

Hartu beharreko hirugarren erabakia, NLI sistemaren aukeraketa izan da. A2Tren egileek lau eredu ezberdin proposatzen dituzte: *roberta-large-mnli*, *joeddav/xlm-roberta-large-xnli*, *facebook/bart-large-mnli* eta *microsoft/deberta-v2-xlarge-mnli*. Laurak *Hugging Face* webgunean aurki daitezke. Esperimentu hau egiteko *microsoft/deberta-v2-xlarge-mnli* eredua erabili da.

Azken erabakia testuinguruari dagokiona da. A2Tri kasu bat pasatzen zaionean testuinguru posible ezberdinak erabil daitezke. Hona hemen topatu diren hiru aukera ezberdinak:

1. **Testu osoa**: Lehen aukera testu osoa pasatzea da. Hau da, bi entitateak agertu baino lehen aipatzen den guztia, bien arteko informazio guztia eta baita bien ondorengoa ere. Egia esan gehiegizkoa dela iruditzen da, bi entitateren artean erlazioa existitzen dela jakiteko informazio gutxiago nahikoa izan daiteke.

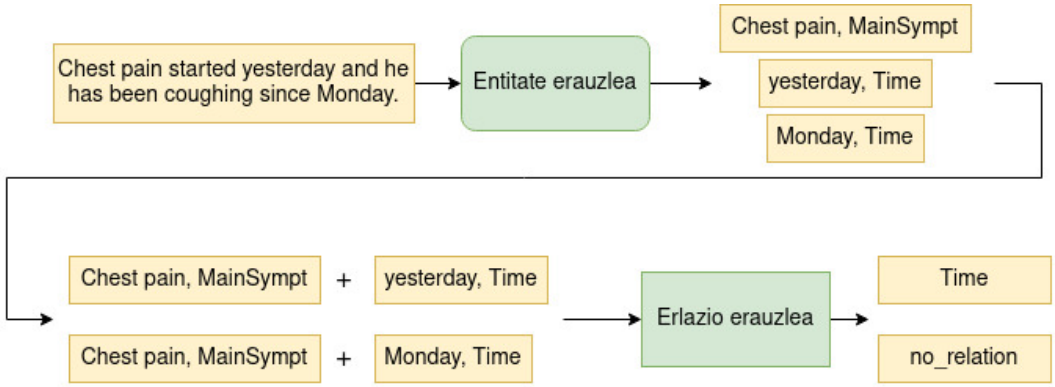
2. **Bi esaldi eta tarteko guztia**: Bigarren aukeran, entitateen aurretik eta ondoren aipaturiko informazioa kendu egiten da. Zehazki, testua esaldika zatitzen da, eta lehen entitatearen esalditik hasita bigarren entitatearen esaldiraino dauden esaldi guztiak hartzen dira. Biak esaldi berean egongo balira, esaldi hori bakarrik hartuko litzateke.

3. **Bi esaldi**: Aurreko aukeraren oso antzekoa da, baina, entitateak agertzen diren bi esaldiak bakarrik hartzen dira. Kasu honetan, tarteko guztiak baztertu egiten dira. Beste behin, biak esaldi berean egonez gero, esaldi bakar bat hartuko litzateke.

Lehen aukera nahiko baztertuta gelditu da, informazio gehiegi pasatzen dela pentsatzen delako. Baina, beste bien artean ez dago baten edo bestearen alde apustu egiteko arrazoi nagusirik. Hori dela eta, esperimentazioan biak erabiltzea erabaki da. Honela, bietako aukera hoberena zein den ikusteko aukera egon da.

Erabaki guztiak hartuta, corpusean lehen etiketatzaileak etiketatutako erlazioak pasa zaizkio A2T ereduari. Lehen etiketatzailearenak aukeratu dira klase guztietako kasuak topatu dituen bakarra delako. Baina hau ez da egin edozein modutan. Kontuan izan behar da, erlazio-erauzlea entitate-erauzlearen ostean datorrela. Honek emandako irteerak erabilita lortu behar ditu erlazioak. Are gehiago, entitate-erauzleak topatu dituen alarma-sintomen artean *MainSympt*-ekin erlaziorik ez duten kasuei *no_relation* klasea esleitzeko gai izan behar du.

Corpusean bi motatako alarma-sintomak topa ditzakegu: *MainSympt*-ekin erlazionatuta daudenak eta erlazionatu gabe daudenak. Erlazionaturikoen kasuan, sintoma nagusiaren entitateak, alarma-sintomak eta dagozkien testuinguruak eman zaizkio A2Tri. Hauek jasota, A2Tren iragarpenak alarma-sintomaren klasea eman behar luke. Erlazionatu gabeetan, alarma-sintomatik gertuen dagoen sintoma nagusia hartzen da. Esaldi berean hasten da bila. Bertan aurkitzen ez bada, aurreko esaldietan bilatzen da. Aurreko esaldietan ere aurkitzen ez bada, ondorengoetan begiratzen da. Kasu hauetan iragarpenak *no_relation* izan behar luke.



5.1 Irudia: Entitate-erauzlearen irteerak erlazio erauzleari pasatzeko eskema.

Entitate-erauzleek topatu dituzten entitateen kasuan, anotatuta badaude, entitate-erauzleek identifikatu bezala pasatzen zaizkio erlazio-erauzleari. Alegia, etiketatzaileak *the cough* jarri badu eta erauzleak *cough* lortu badu, *cough* pasatzen zaio. Honela, erlazio-erauzlea entitate-erauzleen irteerak tratatzeko gai ote den ikusi nahi izan da. Aldiz, anotatu gabe badaude, corpusean anotatuta dauden eta erlaziorik ez duten entitateak bezala tratatu dira (aurreko paragrafoan azaldu bezala). Kasu horietan, *no_relation* itzuli behar luke erlazio-erauzleak.

5.1 irudian, azalduriko prozesu guztia islatzen duen eskema bat aurkitzen da. Sarrerako testuan, entitate erauzleek hiru entitate topatu dituzte. Horietako bat sintoma nagusia da, eta beste biak *Time* motakoak. Denborazko lehen entitatearen kasuan, corpusean sintoma nagusiarekin erlazionaturik agertzen da. Hori dela eta, sintoma nagusia eta denborazko entitatea pasa zaizkio erlazio-erauzleari. Erlazioa existitzen denez, *Time* itzuli beharko luke. Bigarrenari dagokionez, corpusean entitatea ez dago anotatuta. Ondorioz, gertuen duen sintoma nagusia topatu da (*Chest pain*) eta honekin batera pasa zaio erlazio-erauzleari. Anotatu gabe dagoenez, *no_relation* itzuli beharko luke.

Hau guztia esanda, azpimarratu beharra dago *zero-shot* ereduarekin ez direla ebaluatu erlazio etiketatu guztiak. Corpusa bi zatitan banatu da: 5 testu garapenerako eta 15 testerako. 5.2.2 sekzioan azaldu den esperimentua 5 testuen gainean egin da. Banaketa hau egin da, ondoren esperimentua zabaltzeko aukera egoteko (adibidez, *fine-tuning* bat egiteko). Halere, egia da azkenean ez dela esperimentua zabaldu. Izan ere, corpusa oso txikia da, bai eredu bat garatzeko, eta baita eredua ebaluatzeko ere. Ezingo litzateke ziur esan garaturiko eredua ona denik, beraz esperimentu hori albo batera utzi eta proiektua beste helburu batzuetara bideratu da. Bost testu horietan, klase guztietako ahalik eta erlazio gehien hartzeko ahalegina egin da. Testu horietako klase bakoitzeko entitateen agerpenak 5.1 taulan ikus daitezke.

5.2.3 RareDis corpusa

6.1.2.1 sekzioan ikusiko den moduan, *zero-shot* esperimentuak nahiko emaitza onak eman ditu. Beraz, eredua garaturik dagoela esan al daiteke? Esperimentu hori nahikoa al da eredua ontzat emateko? *Zero-shot* eta *few-shot* teknikak probatzen direnean beti galdera berak etortzen dira burura. Kasu gutxi batzuetan ondo funtzionatzeak ez du esan nahi guztietan ondo funtzionatuko duenik, zorte kontua izan daiteke. Beharbada mundu errealean topatuko

Klasea	Kopurua
MainSympt	19
Time	4
OnsetGrad	1
Radiation	3
Sympt	30
Duration	3
ImprovePosNeg	8
OnsetCirc	5
Localization	7
Type	2

5.1 Taula: Garapeneko testuetan agertzen diren entitate kopuruak

dituen kasu askotan ez du ondo funtzionatuko. Hortaz, *zero-shot* eta *few-shot* teknikak benetan gai al dira eredu sendo bat garatzeko? Hori egiaztatu nahi izan da. *Ask2Transformers* liburutegiaren gainean esperimentu sakonagoak egin dira, entrenatzeko kasurik erabili gabe (edo gutxi erabilita) kasu askoren gainean probatuz.

Horretarako *RareDis* corpusa erabili da (4.1.2 sekzioa), proiektu honetarako interesgarria dena. Bertan, askoz erlazio etiketatu gehiago daudenez, analisi sakonagoak egin daitezke eta ondorio garbiagoak atera. Corpus honen gainean, *Ask2Transformers* liburutegia probatu da *zero-shot* eta *few-shot* teknikak aplikatuz. Honetaz gain, erreferentzia gisa *random baseline* bat sortu da. Azkenik, egoera berdinean egonez gero (datu etiketatu gutxi) teknika tradizionalak erabilita zein emaitza lortuko genituzkeen egiaztatu nahi izan da.

5.2.3.1 *Zero-shot*

Beste behin, A2T liburutegia erabiltzeko hainbat erabaki hartu behar izan dira.

Kasu honetan, *MIMIC-III* corpusean ez bezala, klase bat baino gehiago daude definituta. Ondorioz, klase guztiak gehitu dira A2Tn (*multiclass* sailkapena). Klase bakoitzak bere baldintzak ditu eta baldintza berdin bat klase ezberdinetan agertu daiteke. Laburrean azalduz, *Produces* klasea gaixotasun baten (*DISEASE*, *RAREDISEASE*, *SKINRAREDISEASE* edo *ANAPHOR*) eta zeinu edo sintoma (*SIGN* edo *SYMPTOM*) baten artean; *Anaphora* klasea gaixotasun eta anafora (*ANAPHOR*) banaren artean; *Increases_risk_of* eta *Is_a* klaseak bi gaixotasunen artean eta *Is_synon* eta *Is_acron* klaseak mota bereko bi gaixotasunen artean ager daitezke. Baldintzen definizio zehatza jakin nahi izanez gero, B eranskinean dago ikusgai.

Klase bakoitzeko kasu gutxi batzuen gainean proba txikiak eginez, klaseko hainbat txantiloi definitu dira. Guztiak aztertu nahi izanez gero, C eranskinean daude idatzirik.

NLI sistema gisa berriz, hasieran *microsoft/deberta-v2-xlarge-mnli* eredua hartu da. Horixe da txantiloiak definitzeko egindako probetan erabilitakoa. Baina, *few-shot* teknikak aplikatzean *fine-tuning* prozesuak azkarragoak izan zitezen, ondoren eredu txikiago bat erabili da. Zehazki *roberta-large-mnli* eredua.

Aldiz, testuinguruari dagokionez, 6.1.2.1 sekzioan ikusiko den moduan bi esaldikoarekin emaitza hobeak lortzen direnez, hemen bi esaldikoa bakarrik erabili da.

p	*Baller-Gerold syndrome* (BGS) is a rare *genetic disorder* that is apparent at birth.	
h	*Baller-Gerold syndrome* is a *genetic disorder*	ENTAILMENT
h	*Baller-Gerold syndrome* may result in *genetic disorder*	NEUTRAL
h	*BGS* is caused by *genetic disorder*	CONTRADICTION

5.2 Taula: Erlazio etiketatu batekin sor daitezkeen hiru NLI kasu

5.2.3.2 *Few-shot*

Orain arte, *zero-shot* teknika bakarrik aipatu da. Ereduari ez zaio kasurik ematen entrenatzeko. Baina, zenbait testuingurutan, klase bakoitzeko hainbat kasu etiketatu edukitzen dira. Adibidez, hori gertatzen da *MIMIC-III* corpusean. Kasu horiek, gutxi izanagatik ere, *zero-shot* eredua hobetzen lagun dezakete.

Ask2Transformers liburutegia erabiltzean, aldez aurretik beste pertsona batek entrenatu duen NLI sistema bat aprobetxatzen da. Sistema horiek izatez beste corpus batzuekin entrenatu dira eta ez dute inoiz ikusi *zero-shot* bidez ebatzi nahi den problemako kasurik. Eredu hauei, etiketatutako adibideak erabiliz *fine-tuning* txiki bat eginez gero, emaitzak hobetzeko aukera dago.

Hori bai, horretarako, datu etiketatuak ereduaren sarreren formatura pasa behar dira. Kasu honetan, NLI sistema baten formatura. Hauek, sarrera gisa p premisa bat eta h hipotesi bat hartzen dute eta irteera moduan *ENTAILMENT*, *NEUTRAL* edo *CONTRADICTION* itzultzen dute. *RareDis* corpusetik, adibide bakoitzeko, testuingurua eta bi entitateak lor daitezke. Informazio hori eta A2T ereduan definituriko txantiloiak erabilita, NLI sistemarentzat kasu ezberdinak sortu behar dira.

Erlazio etiketatu bakoitzeko, NLI sistemarentzat kasu bat baino gehiago sor daitezke. Guztien premisa berdina izango da: erlaziotik lortzen den testuingurua. Hipotesiak aldiz, era ezberdinetara lor daitezke. Sortzen diren moduaren arabera, klase bateko edo besteko (*ENTAILMENT*, *NEUTRAL* edo *CONTRADICTION*) kasuak lortuko dira. Hona hemen proiektu honetan nola lortu diren klase bakoitzerako hipotesiak:

- *ENTAILMENT*: Erlazio etiketatuaren klasearentzat definitu diren txantiloi bakoitzeko hipotesi bat sortzen da. Txantiloian X eta Y bi entitateez ordezkatzen dira. Hau honela egin da, azken finean, sistemak txantiloi horiek edukiko dituelako erlazioa existitzen dela frogatzeko.
- *NEUTRAL*: Erlazio etiketatuaren klasearenak ez diren gainerako txantiloi bakoitzeko, hipotesi bat sortzen da (betiere txantiloiaren klaseko baldintzaren bat betetzen badu). Txantiloian X eta Y bi entitateez ordezkatzen dira. *NEUTRAL* bezala jarri dira, berez hipotesi horiek betetzen ez diren arren, erlazioa badagoelako bi entitateen artean.
- *CONTRADICTION*: Erlazio etiketatuaren testuinguruan agertzen diren entitateen artean erlazionaturik ez dauden bikoteak lortzen dira. Bikote bakoitzeko, definituta dagoen txantiloi bakoitzeko hipotesi bat sortzen da X eta Y entitate bikoteaz ordezkatuz (betiere txantiloiaren klaseko baldintzaren bat betetzen badu). Azken finean, hauen artean erlaziorik ez dagoenez, ezingo baita inolako hipotesirik ondorioztatu.

5.2 taulan, *Is_a* klaseko erlazio etiketatu batetik sor daitezkeen hiru NLI kasu aurkezten dira.

Eszenarioa	Zatia	Klaseko kasu kop.	Kasu kop. guztira
1-1	train	36	108
	dev	36	108
8-4	train	288	864
	dev	144	432
16-8	train	576	1728
	dev	288	864
32-16	train	1152	3456
	dev	569	1707

5.3 Taula: Eszenario bakoitzean sorturiko kasu kopuruak

4.2 taulan, *RareDis* corpusean erlazio klase bakoitzeko train eta dev zatietan zenbat kasu etiketatu dauden ikusi daiteke. *Few-shot* teknikarekin esperimentuak egite aldera, corpuseko oso kasu etiketatu gutxi daudela atzigarri simulatu da. Zehazki, lau i-j eszenario definitu dira, ausaz klase bakoitzeko train zatiko i kasu eta dev zatiko j kasu ezberdin hartuz. Definituriko lau eszenarioak hauek dira: 1-1, 8-4, 16-8 eta 32-16. Tamaina ezberdinetako eszenarioak hartu dira datu kopuruak zein eragin duen jakin nahi izan delako. Aipatu beharra dago, eszenario handiagoak, eszenario txikiagoen kasuak hartu eta hauei gehiago gehituta lortzen direla. Hau da 8-4 eszenarioan 1-1 eszenarioan agertzen diren kasu guztiak eta gehiago daude.

Eszenario bakoitzean, aipaturiko estrategiak jarraituz NLI sistemarentzako kasuak sortu dira. Eta, klaseen artean desoreka handia sortzen dela ikusi da. *ENTAILMENT* klaseko kasu gutxi sor daitezke adibide batekin. Aldiz, *NEUTRAL* eta *CONTRADICTION* kasuak normalean askoz gehiago sortzen dira. Adibide bat ematearren, 32-16 eszenarioan, train zatian 1152 *ENTAILMENT*, 2053 *NEUTRAL* eta 14140 *CONTRADICTION* kasu eta dev zatian 569 *ENTAILMENT*, 1102 *NEUTRAL* eta 8750 *CONTRADICTION* kasu sor daitezke.

Desoreka nabarmen horiek ez dira egokiak izaten entrenamenduak egiteko. Ereduek klase bakar bat iragartzeko joera handia hartzen dute. Hori dela eta, kasuak orekatzea erabaki da. *ENTAILMENT* klaseak, kasu kopuru gutxiena edukitzen dituenak, gainerakoek edukiko duten kasu kopurua definitzen du. Beste klaseetan, sor daitezkeen guztietatik, *ENTAILMENT* klaseak dituen adina adibide ausaz hartzen dira. 5.3 taulan eszenario bakoitzeko klase bakoitzeko sortu diren kasu kopuruak daude ikusgai.

Eszenarioak prest izanda, *roberta-large-mnli* ereduari *fine-tuning*-a egin zaio eszenario bakoitzean. Horretarako, *run_glue.py python script* publikoa erabili da [38]. 16 tamainako *batch*-a erabili da eta 50 *epoch* egiten utzi da. *Epoch* bakoitzean, ereduaren egoera gorde da eta entrenamenduaren amaieran, dev zatian *accuracy* altuena eman duen eredua mantendu da.

5.2.3.3 *Random* oinarri-lerroa

Zero-shot eta *few-shot* teknikekin lorturiko emaitzen kalitatea neurtzeko oinarri bat edukitzeko, *Random* oinarri-lerroa sortu da.

Funtsean *random* funtzio bat den arren, ez da guztiz ausazkoa. *RareDis* corpusean sei erlazio mota daude definiturik (zazpi *no_relation* klasearekin). Baina, oinarri-lerroari entitate bikote bat pasatzen zaionean, ez du aukeratzen zazpi horietako edozein. Aurreko

azpiatalean azaldu den moduan, klase bakoitzak bere baldintzak ditu. Hau da, erlazio mota bakoitza klase konkretu batzuetako entitateen artean eman daiteke. Hau jakinda, *random* oinarri-lerroak entitate bikoteak zein klaseetako baldintzak betetzen dituen begiratzen du eta horien artean bakarrik aukeratzen du klasea ausaz (*no_relation* barne).

5.2.3.4 Ikaskuntza gainbegiratuko ereduak

Beti esan izan da ikasketa gainbegiratuko algoritmoek ez dutela ongi ikasten corpus txiki bat edukitzean. Hori dela eta, joera handia dago corpusean datu etiketatu gutxi daudela ikustean, teknika tradizionalak albo batera utzi eta bestelako estrategia berezi batzuk (adibidez, *zero-shot* eta *few-shot*) probatzen hastekoa. Baina, benetan beharrezkoa al da buruhauste horietan sartzea? Benetan emaitza txarrak lortzen al dituzte ikasketa gainbegiratuko teknikek? Proiektu honetan baietz frogatu nahi izan da. Horretarako, ikasketa gainbegiratua erabiliz lau eszenario ezberdinetarako eredu berriak sortu dira. Helburua, hauekin lortutako emaitzak *Ask2Transformers* liburutegiarekin lorturikoak baino okerragoak izatea zen. Honela, frogaturik geldituko litzateke kasu gutxirekin merezi duela *zero-shot* eta *few-shot* teknikak erabiltzeak.

Hain zuzen, 3.3.2.1 sekzioan azaldutako entitate marketan oinarritutako eredua garatu da. Zehazki *Hugging Face* webguneko *xlm-roberta-base* BERT eredua erabilita. Bestalde, entitateen hasierako markei dagozkien tokenen errepresentazioak kateatuta pasa zaizkio *feed forward* geruzari. *Feed forward* geruzak irteera moduan, erlazioen sei klase posibleak ditu.

No_relation klasea bi eratara modelatu daiteke. Batetik, *feed forward* geruzan esplizituki gehitu daiteke. Honek, entrenamenduko kasuetan erlazio gabeko kasuak gehitzea eskatzen du. Aldi berean, hiperparametro berri bat sartzen du: Zenbat *no_relation* kasu hartuko dira? Bestetik, *no_relation* klasea inplizituki sar daiteke. *Feed forward* geruzan ez da klase hori gehitzen. Baina, iragarritako klaseari konfiantza-maila bat eskatzen zaio. Geruzaren irteeran, klase posible bakoitzaren probabilitatea hartzen da eta probabilitate handiena duen klasea iragartzen da. Erlazio gabeko kasuak modelatzeko, iragarritako klasearen probabilitatea atalase konkretu bat baino handiagoa izatea eskatzen da. Ez bada hala, *no_relation* iragartzen da.

Esperimentazio fase honetan, bi aukerak probatu dira. Gainera, erlazio gabeko klasea esplizituki gehitzen den ereduan, bi *no_relation* etiketatu kopuru ezberdin hartu dira. Alde batetik, eszenarioan agertzen diren gainerako klaseen agerpen kopuru bera. Adibidez 8-4 eszenarioan, train zatian 8 kasu eta dev zatian 4 kasu sartu dira. Beste aldetik, corpusean guztira dauden gainerako kasuen kopuru bera. Alegia, 8-4 eszenarioan test zatian 48 eta dev zatian 24 kasu. *No_relation* kasuak, etiketatutako erlazioen testuinguruan agertzen diren erlazio gabeko entitate bikoteekin lortu dira.

Beraz, eszenario bakoitzeko hiru eredu berri garatu dira. Entrenatzeko, *learning rate* gisa $1e-5$ erabili da eta 32ko *batch*-ak erabili dira. Honetaz gain, entrenamenduak oso azkarrak zirenez, 200 $epoch$ egin dira eta amaieran *F-score* onena eman duen *epoch*-eko eredua mantendu da.

Helburua, esan bezala A2T liburutegia hiru modelo berri hauek baino hobea dela frogatzea izan da.

Klasea	Kopurua
Produces	4045
Increases_risk_of	111
Is_a	659
Is_acron	163
Is_synon	34
Anaphora	1073
no_relation	20000

5.4 Taula: Test zatiko klaseko kasu kopurua

5.2.3.5 Test zatia

Eredu guztiak ebaluatzeko, *RareDis* corpusetik 32-16 eszenarioan sartu ez diren kasu guztiak hartu dira. Honela, ereduek entrenatzeko inoiz ikusi ez dituzten hainbat kasu lortu dira. 5.4 taulan, klase bakoitzeko zenbat kasu konkretu dauden ikus daiteke.

Erlazio gabeko kasuak sortzeko, corpusean agertzen diren entitate bikote posible guztiak lortu dira eta hauen artean erlaziorik ez zuten bikoteak hartu dira. Egia esan guztira izugarri kasu pila sor zitezkeen. Testean hauetako 20000 sartu dira.

6. Emaitzen analisia

Kapitulu honetan, aurreko kapituluan aurkezturiko esperimentu guztiekin lortu diren emaitzak aztertzen dira. Batetik, lorturiko emaitzak bere horretan aurkezten dira eta eredu ezberdinen artean konparazioak egiten dira. Bestetik, lorturiko emaitzen inguruan eztabaida bat egiten da. Alegia, emaitzetatik atera daitezkeen ondorioak aztertzen dira.

6.1 Emaitzak

6.1.1 Entitate-erauzlea

Entitate-erauzleen kasuan, klase bakoitza bere kabuz aztertu da. Klase bakoitzeko, bi etiketatzaileen notazioak kontuan hartuz, doitasun, estaldura eta *F-score* balioak lortu dira.

Zehazki, kointzidentzia erlaxatuaren ebaluazioa egin da (3.3.1.2 sekzioa). Izan ere, izatez zuzenak diren eta kointzidentzia zehatzik ez zuten emaitza ugari topatu dira. Adibide batzuk aipatzearren, *Sympt* klasean, etiketatzaileak *a cough* jarri duen tokian sistemak *cough* eman du, etiketatzaileak *tingling and numbness* jartzean sistemak biak bananduta eman ditu, etiketatzaileak *associated with shortness of breath* eta sistemak *shortness of breath*... Honelako kasuak oso errepikatuak dira *Duration* klasean ere, etiketatzaileek *for* hitza jartzeko joera dutelako eta sistemak ez. Adibidez, *for one hour* eta *one hour*.

Ebaluazio metrikekin lortutako balio guztiak 6.1 taulan ikus daitezke. Klase bakoitzeko, mediku bakoitzaren notazioekin zein doitasun, estaldura eta *F-score* balio lortu diren agertzen da. Egia esan, etiketatzaile batetik bestera emaitzak ez dira gehiegi aldatzen.

Bereziki deigarria den ezaugarrietako bat, taulan agertzen diren doitasun balioak orokorrean baxuak direla da. Doitasun ebaluazio metrikak, sistemak identifikatu dituen entitateetatik ondo sailkatu dituenen proportzioa adierazten du. Baxua izatearen arrazoi nagusia, topatu behar zituen entitateak topatzeaz gain, beste asko topatu dituela izan ohi da. Estaldura balioak nahiko altuak direnez (corpusean etiketatuta daudenetatik ondo sailkatutako entitateen proportzioak), ematen du hori gertatu dela hemen ere.

Hori horrela izanda, ez du asko axola doitasun balioak baxuak izateak. Azken finean, proiektu honetan definitu den entitate-erauzleak, bularraldeko minarekin erlazionaturik dauden alarma-sintoma posibleak bilatzeko balio du. Ez ditu zertan bularraldeko-minarekin erlazionatuak bakarrik topatu behar. Gehiago topatzen baditu, erlazio-erauzlea arduratuko da soberakoak baztertzeaz. Halere, egia da zenbat eta doitasun handiagoa lortu erlazio erauzleak kasu gutxiago baztertu beharko dituela. Baina, kasu honetan, interesatzen dena estaldura balio altuak lortzea da. Hala izanez gero, *MainSympt* klasearekin lotuta dauden alarma-sintomak topatu diren seinale. Hortik aurrerako lanak erlazio erauzleak egingo ditu. Gainera, kasu honetan estaldura balioak nahiko altuak dira. Beraz, sistemak egokiak direla esan daiteke.

Honetaz gain, azpimarratzekoa da *Radiation* klaserako sorturiko lematizatzailearen programa sinpleak lortzen dituen emaitzak. Egia da behar bezala ebaluatzeko datuak falta

Klasea	Tresna	Etik.	Doitasuna	Estaldura	F-score
MainSympt / Sympt	Metamap	1	0.50	0.64	0.56
		2	0.46	0.61	0.52
Time	SUTime	1	0.04	0.78	0.07
		2	0.07	0.88	0.13
Duration	SUTime	1	0.11	0.80	0.19
		2	0.11	0.57	0.18
Radiation	Lematizatzailea	1	1.00	1.00	1.00
		2	0.75	1.00	0.86

6.1 Taula: Entitate klase bakoitzeko ebaluazioan lorturiko emaitzak

direla (oso gutxi daude). Baina, etiketaturiko kasuekin behintzat oso ondo funtzionatzen du. Zenbait kasutan Adimen Artifizialeko programa konplexuak albo batera utzi eta ideia sinple bat garatuta emaitza onak lortzen dira.

Kontuan izan behar da emaitza horiek eredurik entrenatu gabe lortu direla. Hau da, ez dute esfortzu handiegirik eskatu. Hori horrela izanda ere, emaitza onak lortu dira. Beraz, datu-etiketatu gutxi edukitzen diren kasuetan, interesgarria izan daiteke publikatuta dauden ereduen inguruko ikerketa bat egitea.

6.1.2 Erlazio-erauzlea

Erlazioak erauzteari dagokionez, emaitza asko daude aztertzeko. Batetik, *MIMIC-III* corpusaren gainean bi testuinguru motarekin (gehienez bi esaldi erabilita eta entitateen arteko esaldi guztiak erabilita) egindako *zero-shot*-ak lorturikoak. Bestetik, *RareDis* corpusaren gainean egindako *zero-shot* eta *few-shot* ezberdinek lortutakoak. Baita teknika tradizionalekin egindako saiakerenak ere. Guztiek eman dute zer esana.

Esperimentu orok, errore-matrize propioak ditu. Asko dira eta hauen arteko konparaketak egitea zaila izaten da. Askoz sinpleagoa izaten da eredu ezberdinen arteko doitasun, estaldura eta *F-score* ezberdinak konparatzea. Halere, esperimentu guztien errore-matrize guztiak atzigarri daude E eranskinean.

6.1.2.1 *Zero-shot MIMIC-III* corpusean

Eredu honetan, aurrez azalduta dagoen moduan, entitate klase bakoitzeko sailkapen bitar bat egiten da (*relation* edo *no_relation*). Hori dela eta, klase bakoitzaren emaitzak bananduta ebaluatu dira. 6.2 taulan, klase bakoitzeko sailkapen bitarrean lorturiko doitasun, estaldura eta *F-score* ezberdinak aurkezten dira. Gainera, probaturiko bi testuinguru ezberdinen arteko konparaketa egiten da. 2E-k gehienez bi esaldi erabiltzen dituenari egiten dio erreferentzia. TO-k berriz, entitateen bi esaldiak eta tarteko guztiak erabiltzen dituenari (Testuinguru Osoa).

Sympt klasean, ikusten da bi esaldiko testuinguruarekin emaitza hobeak lortzen direla. Testuinguru osoa erabiltzen denean, *no_relation* klasea esleitzeko joera handia hartzen du. Zenbait kasutan asmatzen du, eta horregatik, klase horretako estaldura hobea da. Baina, izatez erlazioa duten kasu dezentetan erlaziorik ez dagoela ere esaten du. *No_relation* klaseko doitasun balioaren jaitsiera nabarmenak adierazten du hori. Beraz, erlazio gabeko kasu gehiago topatzen ditu, baina existitzen diren erlazio ugari ez ditu ikusten. Hau da,

Klasea	Iragarpena	Doitasuna		Estaldura		F-score	
		2E	TO	2E	TO	2E	TO
Sympt	relation	**0.75**	0.72	**0.89**	0.67	**0.81**	0.69
	no_relation	**0.50**	0.31	0.27	**0.36**	**0.35**	0.33
	macro avg	**0.62**	0.51	**0.58**	0.52	**0.58**	0.51
Time	relation	0.17	**0.20**	**1.00**	**1.00**	0.29	**0.33**
	no_relation	**1.00**	**1.00**	0.23	**0.38**	0.38	**0.56**
	macro avg	0.58	**0.60**	0.62	**0.69**	0.33	**0.44**
Duration	relation	0.09	**0.14**	**1.00**	**1.00**	0.17	**0.25**
	no_relation	**1.00**	**1.00**	0.09	**0.45**	0.17	**0.62**
	macro avg	0.55	**0.57**	0.55	**0.73**	0.17	**0.44**
OnsetCirc	relation	**1.00**	**1.00**	**0.67**	**0.67**	**0.80**	**0.80**
	no_relation	**0.67**	**0.67**	**1.00**	**1.00**	**0.80**	**0.80**
	macro avg	**0.83**	**0.83**	**0.83**	**0.83**	**0.80**	**0.80**
OnsetGrad	relation	0.00	**1.00**	**1.00**	**1.00**	0.00	**1.00**
	no_relation	**1.00**	**1.00**	0.00	**1.00**	0.00	**1.00**
	macro avg	0.50	**1.00**	0.50	**1.00**	0.00	**1.00**
Radiation	relation	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	no_relation	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	macro avg	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
ImprovePosNeg	relation	**0.75**	0.67	**0.50**	0.33	**0.60**	0.44
	no_relation	**0.25**	0.20	**0.50**	**0.50**	**0.33**	0.29
	macro avg	**0.50**	0.43	**0.50**	0.42	**0.47**	0.37
Type	relation	**0.00**	**0.00**	**1.00**	**1.00**	**0.00**	**0.00**
	no_relation	**1.00**	**1.00**	**0.50**	**0.50**	**0.67**	**0.67**
	macro avg	**0.50**	**0.50**	**0.75**	**0.75**	**0.33**	**0.33**
Localization	relation	**0.33**	0.25	**0.33**	**0.33**	**0.33**	0.29
	no_relation	**0.50**	0.33	**0.50**	0.25	**0.50**	0.29
	macro avg	**0.42**	0.29	**0.42**	0.29	**0.42**	0.29

6.2 Taula: A2T liburutegiarekin *MIMIC-III* corpusean lorturiko emaitzak testuinguru ezberdinekin

testuinguru osokoak, *no_relation* esleitzeko joera handia du. Era berean, bi esaldikoak erlazioa badagoela esateko joera duela esan daiteke. Baina, batezbesteko doitasun, estaldura eta *F-score* balioek zalantzarik gabe bi esaldikoaren alde egiten dute.

Time klaseari dagokionez, testuinguru osoa erabiltzen duen ereduak emaitza hobeak eman ditu, erlazio falta identifikatzeko gaitasun handiagoa erakutsi duelako. Hori, *no_relation* klasearen estaldura balioaren hobekuntzan islatu da. Horrek, gainerako balioetan hobekuntza txikiak eragin ditu eta hortaz testuinguru osoa erabiltzen duenak hobeto funtzionatu du.

Duration klasearen inguruan antzeko zerbait esan daiteke. Ereduak, testuinguru osoa erabiltzean *no_relation* klasea gehiagotan iragartzen du eta honekin klasearen estaldura balioa hazten da. Horrek, zeharka gainerako balioak hobetzea ekarri du.

OnsetGrad klaseari dagokionez, testuinguru osokoak emaitza hobeak eman ditu. Baina, klase honek ez du balio ondorio sendoak ateratzeko. Izan ere, bertan kasu bakar bat probatu da, erlaziorik ez duena. Ondorioz, kasu guztientzat *no_relation* itzultzen duen eredu batek

ere emaitza onak bueltatuko lituzke.

ImprovePosNeg klasearekin, *Sympt* erlazioarekin gertatzen zenaren antzera, ereduak erlazio falta iragartzeko joera handiagoa erakusten du testuinguru osoarekin eta berez badagoen erlazioren bat baztertu egin du. Horren ondorioz, bi esaldiko ereduak metrika guztiak hobeak lortu ditu.

Localization klaseari dagokionez berriz, aurkakoa gertatu da. Erlaziorik ez dagoen kasuren bat bi esaldikoak ongi sailkatu du eta testuinguru osokoak gaizki. Horregatik dira hain ezberdinak *no_relation* klaseko balioak. Horrek, beste metriketan ere ezberdintasun txikiak sortu ditu.

Aldiz, *OnsetCirc*, *Radiation* eta *Type* klaseetan biek emaitza berberak lortu dituzte. Beharbada hau gertatu da kasu hauetan entitateak gehienez esaldi bateko distantziara daudelako eta ondorioz biek testuinguru berak lortu dituztelako.

Beraz, klase guztien arteko konparazio bat eginda, nahiko garbi esan daiteke bi esaldiekin lortutako emaitzak hobeak direla. Zenbaitek pentsa lezake testuinguru osoa erabiltzen duenak *no_relation* hobeto topatzen duela. Gainera, hori topatzea izaten da zailena erlazio-erauzleetan. Baina, hobeto aurkitu baino, besterik gabe *no_relation* klasea gehiagotan iragartzeko joera hartzen duela ematen du.

Are gehiago, izatez zentzua du bi esaldikoak hobeto funtzionatzeak. Azken finean A2Tren atzean NLI sistema bat dago. Funtsean NLI sistema batek itzultzen duen emaitza hartzen da. Hauek, normalean premisa eta hipotesi laburrak erabilita entrenatu ohi dira. Testuinguru osoa hartzen den kasuetan, premisa oso luzea bihur daiteke. 4.1 taulan ikus daitekeen moduan posible da bi entitateen artean lau esaldi baino gehiago egotea. Aldiz, bi esaldikoan beti antzeko luzera duten premisa motzak lortuko dira. NLI sistemak bigarren hauekin emaitza hobeak lortzea espero da.

Konparazioa albo batera utzita, egia esan orokorrean oso emaitza interesgarriak topa daitezke. Klase gehienetan nahiko onak dira. Beraz, A2T liburutegiak ongi funtzionatzen duela esan daiteke. Baina aurreko ataletan aipatzen zen moduan, hemen ongi funtzionatzeak ez du esan nahi A2T berez tresna ona denik. Kasu gutxiren gainean egin da proba eta ezin daiteke esan egoera orokor batean ondo funtzionatuko lukeenik. Baina, hori frogatu nahian *RareDis* corpusean esperimentu sakonagoak egin dira.

6.1.2.2 *Random, zero-shot* eta *few-shot* RareDis corpusean

RareDis corpusean esperimentazioa hasteko, definituriko test zatian *random* oinarri-lerroa, *zero-shot* eredua eta eszenario ezberdinetan entrenatutako *few-shot* ereduak probatu dira. Esperimentu guztietan lorturiko *F-score*-ak 6.3 taulan aurkezten dira.

Bertan, bi *zero-shot* aurki daitezke. Bakoitza, NLI sistema ezberdin batekin egin da. Lehenbizikoa *microsoft/deberta-v2-xlarge-mnli* ereduarekin eta bigarrena *roberta-large-mnli* ereduarekin. 5.2.3.1 sekzioan azaldu bezala A2T ereduarentzat txantiloiak definitzeko esperimentuak *deberta* NLI sistema erabilita egin ziren. Baina, *few-shot* eszenarioak garatzeko modeloa aldatzea erabaki zen (txikiago bat, entrenamenduak azkarrago egiteko): *roberta*. Bigarren eredu honekin *zero-shot* probatzean, ikusi zen txantiloi berdinekin emaitza okerragoak lortzen zituela. Hau bereziki nabarmena da *Is_acron* klasean. Baina, honetaz jabetzerako, NLI sistemari *fine-tuning*-a egiteko kasuak prestatuta zeuden. Txantiloiak aldatzeak hauek aldatzea suposatuko zuenez, bere horretan utzi ziren.

Klasea	Random	Zero-shot (Deberta)	Zero-shot (Roberta)	1-1	8-4	16-8	32-16
Produces	0.35	0.43	0.42	0.24	0.58	0.58	**0.67**
Anaphora	0.29	0.56	0.45	0.61	**0.73**	0.69	0.71
Increases_risk_of	0.03	0.08	0.06	0.04	0.09	**0.13**	0.11
Is_synon	0.01	0.05	0.03	0.02	0.02	0.02	**0.09**
Is_a	0.16	0.46	0.31	0.32	0.40	**0.55**	0.50
Is_acron	0.07	0.1	0.02	0.00	0.17	0.18	**0.19**
no_relation	0.54	0.19	0.03	0.71	0.68	**0.77**	0.75
macro average	0.21	0.27	0.19	0.28	0.38	0.42	**0.43**

6.3 Taula: *Random*, *zero-shot* eta *few-shot* ereduek klase bakoitzeko lorturiko *F-score*-ak

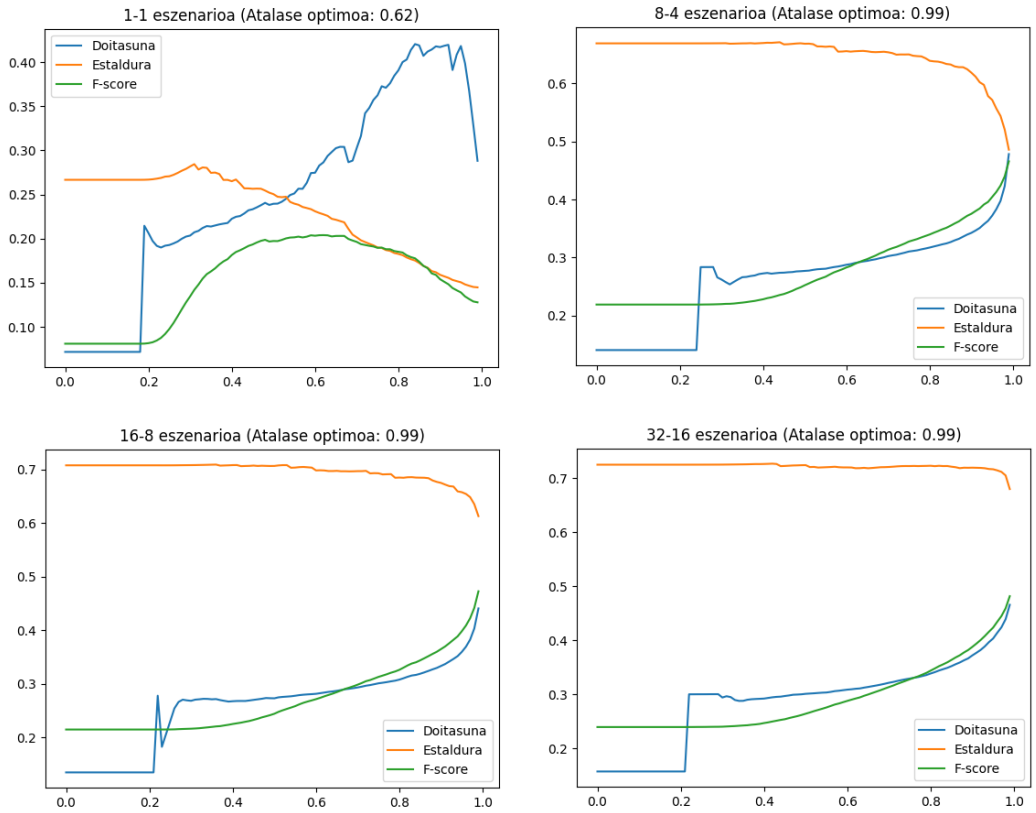
Horrek erakusten du, erabiltzen den NLI ereduak eragin handia duela definituko diren txantiloietan. Eredu desberdinek ezagutza ezberdina baitute. C eranskinean *Is_acron* klaseko txantiloietako bati erreparatuz gero, ikusten da *acronym* hitza esplizituki erabiltzen dela. Bi *zero-shot*-ek klase horretan izan duten aldea ikusita, badirudi *deberta* ereduak badakiela zer den akronimo bat eta *roberta*-k ez.

Bi *zero-shot* esperimentuen arteko konparaketa albo batera utzita, nabarmena da *zero-shot* ereduen eta *random* ereduaren aldea. Orokorrean, klase guztietan askoz emaitza hobeak lortzen dituzte *zero-shot* ereduek. Bi klaserekin lortzen du *random* oinarri-lerroak *zero-shot* ereduren bat gainditzea. Batetik, *Is_acron* klasea *roberta* erabiltzen duenak baino hobeto sailkatzen du. Halere, txantiloiak aldatuz gero *zero-shot*-a hobea izango litzatekeela pentsatzen da. Bestetik, *no_relation* klasea okerrago egiten dute *zero-shot* ereduek. Hau gertatzen da eredu horiek beti erlazioren bat esleitzeko joera handia erakusten duelako. Gutxitan egiten dute *no_relation* klasearen aldeko apustua. Halere, bi klase horiek albo batera utzita *zero-shot* ereduak *random* baino askoz hobeak dira.

Baina, eredua are hobea bilakatzen da entrenatzeko kasu batzuk erabiliz gero. *Zero* eta 1-1 eszenarioen artean, argi eta garbi ikusten da *no_relation* klasearen hobekuntza. Ereduak erlazio gabeko kasuak topatzen ikasten du. Hori bai, *no_relation* gehiegitan esleitzeko joera hartzen du eta ikus daiteke zenbait klasetan $F\text{-}score$-ak bera egiten duela.

Hortik aurrera, entrenatzeko zenbat eta kasu gehiago erabili, orduan eta emaitza hobeak lortzen dira. Beraz, beste behin frogaturik gelditzen da datu-kopuruak handitu ahala Adimen Artifizialeko algoritmoek emaitza hobeak lortuko dituztela. Halere, egia da hobekuntza ez dela beti konstantea, behin 16-8 eszenariora helduta, datuak gehituagatik ere ez dira asko hobetzen emaitzak. Badirudi aurki muga joko duela. Baina, azken hau, egindako laginketarengatik izan daiteke. Eszenario ezberdinak sortzeko, corpuseko erlazio guztien artean gutxi batzuk ausaz hartu dira. Beharbada, harturiko kasu horiek ez dira hoberenak izan. Beste laginketaren batekin joera ezberdinak ikusteko aukera egon daiteke. Berez, laginketa ezberdinak probatzea eta orokorrean zein jokabide hartzen duen aztertzea izango litzateke hoberena.

Aldi berean, egia da, klase askotan emaitza onak lortzen diren arren zenbait klasek nahiko *F-score* baxua dutela. Adibidez, *Is_synon* klaseak. Hau ez da gertatzen klase honetako entitateak aurkitzen ez direlako. Bertakoak ez diren entitate asko (zehazki *no_relation* klasekoak) klase hontakotzat hartzen dituelako baizik. Estaldura altua du, baina doitasun baxua. Oraindik, neurri batean erlazioa ezartzeko joera du.



6.1 Irudia: Atalasea handitu ahala doitasun, estaldura eta *F-score* balioek duten joera

Hala eta guztiz ere, orokorrean oso emaitza onak lortzen ditu. Kontuan izan behar da oso eszenario txikiak direla eta ikerketek esaten dutela honelako sistema batek datu etiketatu asko behar dituela.

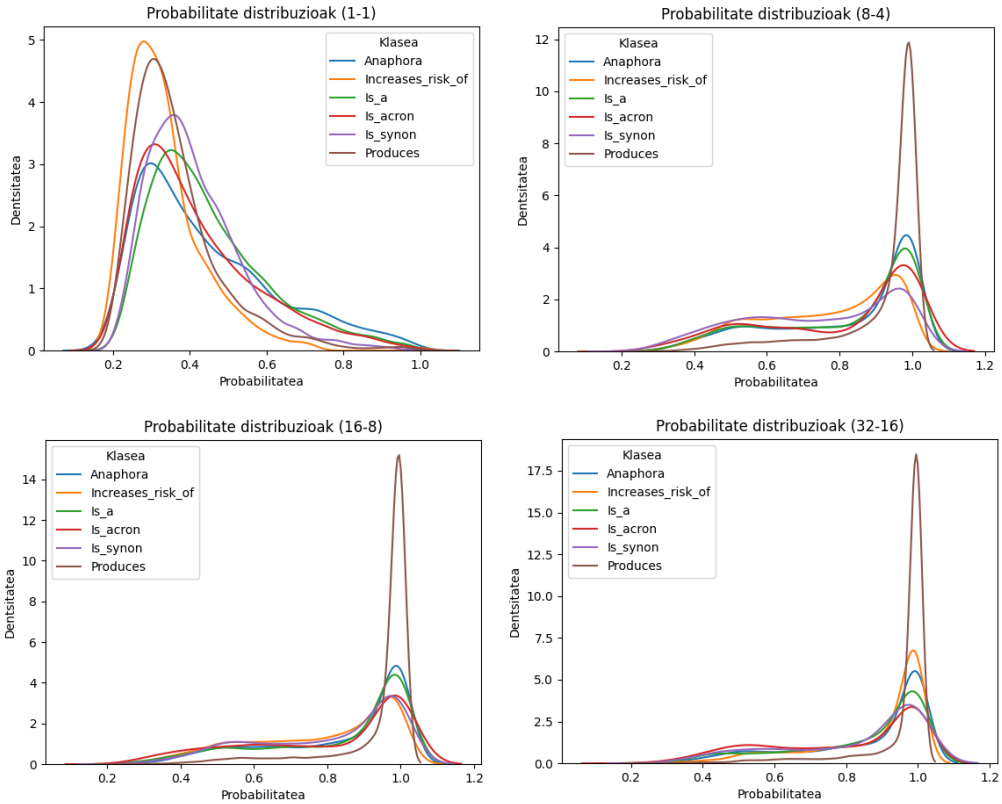
6.1.2.3 Sailkapen gainbegiratuko ereduak

No_relation klasea inplizituki definiturik duen ereduan, atalasea finkatu behar da. Horretarako, atalasea handitu ahala probetako doitasun, estaldura eta *F-score* balioek izan ohi duten joera ikusi da grafiko batzuen bidez. 6.1 irudian agertzen dira grafiko hauek.

Egia esan, nahiko arraroak dira. Normalean, honelako grafikoak irudikatzen direnean, atalasea igo ahala, doitasun balioak gorantz egiten du eta estaldura balioak beherantz. Eta puntu batean ebaki egiten dira. *F-score*-a berriz, aurreko bien ebaki-puntura heldu arte gorantz joaten da eta hau pasa ostean beherantz egiten du.

1-1 eszenarioan bai, definitu den jokabide hori ikus daiteke. 0.6-ko atalasera heltzeko gutxi falta denean doitasun eta estaldura gurutzatu egiten dira. *F-score* optimoa berriz, atalasea 0.62 denean lortzen da.

Baina, gainerako eszenarioetan inoiz ez dira gurutzatzen doitasun eta estaldura balioak. Amaierara arte gerturatuz doazen arren, inoiz ez dute bat egiten. Ondorioz, *F-score*-a gorantz doa amaierara arte. Horrek, hiru eszenarioetan atalase optimoa 0.99 izatea ekarri du. Ez da batere ohikoa hain atalase handiak behar izatea. Beraz, zer gertatzen ari den sakonago aztertu nahi izan da.



6.2 Irudia: Klase bakoitzaren konfiantza-mailaren distribuzioa

Horretarako, dentsitate-grafiko batzuk irudikatu dira. Dentsitate grafikoek, klase bat esleitzen denean honi ematen zaion konfiantza-mailaren distribuzioa erakusten dute. Grafiko horiek 6.2 irudian ikus daitezke.

1-1 eszenarioan, ikusten da iragarritako klasearen konfiantza-mailak nahiko balio baxuetan kontzentratzen direla. Ereduak, klaseen iragarpenak lotsaturik egiten dituela dirudi. Konfiantza handiegirik gabe. Ez da harritzekoa, azken finean entrenamenduan oso kasu gutxi ikusi baititu (klase bakoitzeko erlazio bakar bat). Iragarritako klaseari inoiz ez dio probabilitate handi bat esleituko esperientzia gutxi duelako.

Aldiz, beste eszenarioetan, oso bestelako grafikoak sortzen dira. Balio gehienak 1 inguruan agertzen dira. Bertan mendi handiak ikus daitezke. Klase baten iragarpena konfiantza handiarekin egiten da. Horrek eragiten du hain atalase handia erabili behar izatea. Iragarritako klaseari hain balio altuak ematen badizkio, atalase oso handia erabili beharko da.

Honetaz gain, azken hiru grafikoek badute beste ezaugarri bereizgarri bat. Guztietan, *Produces* klasearen distribuzioa besteekiko oso ezberdina da. Gainerakoek, konfiantza-maila txikietan kasuen kontzentrazio txiki bat erakusten dute eta 1 inguruan kontzentrazio handiago bat. *Produces* klaseari dagokionez, balio txikietako kontzentrazioa txikiagoa da eta 1 ingurukoa askoz handiagoa.

Ezaugarri hau deigarria den arren, atzean egon daitekeen arrazoia ez da hain arraroa. Dentsitate diagramak aztertuz, konfiantza-maila txikietan agertzen diren kasuak *no_relation*

klaseei dagozkienak direla eta eskuineko mendian izatez klase horretakoak direnak direla esan daiteke. Are gehiago, eskuineko mendian erlazio gabeko kasuak ere pilatzen dira, bestela atalaseak ez lirateke 0.99 izatera helduko. Hau esanda, suposatuko da *no_relation* klasea uniformeki banatzen dela gainerako klaseen artean. Hau da, *no_relation* klaseari klase guztiak proportzio berean esleitzen zaizkiola (alegia, ausaz). Beraz, klase guztientzat konfiantza maila txikietan topatuko den erlazio gabeko kasuen kopurua antzekoa da. Ezin daiteke gauza bera esan 1 inguruan topatzen diren kasu kopuruengatik. 4.4 irudiko eskuineko barra-diagraman ikus daitekeen moduan, *Produces* klasea askoz ugariagoa da. Gainerako klaseak parekatuago daude. Beraz, suposatzen bada klase guztiak ongi sailkatu direla, klase bakoitzeko ez da kasu kopuru bera egongo 1 inguruan. *Produces* klaseko askoz gehiago egongo dira. Beraz, dentsitate grafikoan hori islatzen da. Konfiantza-maila txikietan dauden kasuen eta 1 inguruan agertzen direnen arteko proportzioa oso ezberdina da *Produces* klasean. Ondorioz, txikietako marrak beherantz egiten du eta 1 ingurukoak gorantz.

6.1.2.4 *Few-shot* eta eredu gainbegiratuen arteko konparazioa

Proiektu honetan, lau eredu nagusi aurkeztu dira. Batetik, *Ask2Transformers* liburutegiarekin garaturiko *few-shot* eredua (A2T). Bestetik, entitate marketan oinarrituta dagoen eta *no_relation* inplizituki definituta duen eredua (EM). Azkenik, entitate marketan oinarrituta dauden eta erlazio gabeko klasea esplizituki definituta duten bi ereduak. Bata, *no_relation* kasu etiketatu gutxirekin entrenatua (EN1), eta bestea kasu gehiagorekin entrenatua (EN2). Eredu bakoitza lau eszenario ezberdinetan probatu da eta bakoitzarekin lorturiko ebaluazio metrikak 6.4 taulan ikus daitezke.

Hasteko, erlazio gabeko klasea esplizituki definiturik duten bi ereduen arteko konparaketa bat egin daiteke. Klase interesgarriena *no_relation* da. EN1 ereduaren kasuan oso doitasun altuak eta oso estaldura baxuak lortzen dira. Eszenarioa handitu ahala, estaldura balioak apur bat gora egiten du, baina ezer gutxi. Estaldura baxuak erlazio gabekoak diren kasu asko gaizki sailkatzen dituela adierazten du. Aldi berean, doitasun altuak, *no_relation* klasekoak direla esan den kasuen gehiengoa ongi sailkatu dela adierazten du. Beraz, argi eta garbi ikusten da eredu honek, *no_relation* esleitzeko joera oso txikia duela. Erlazio gabeko kasuei gutxitan esleitzen die klase hau, eta bertakoak ez direnei are gutxiagotan. Aldiz, EN2 ereduak, eszenario guztietan doitasun eta estaldura handiak ematen ditu. Alegia, erlazio gabeko kasuen gehiengoa ongi sailkatzen du eta erlazio gabekoak direla esan direnen gehiengoa bertakoa da. Beraz, bi ereduen artean, erlazio falta bigarrengoak hobeto sailkatzen duela ematen du.

Halere, *no_relation* klasean doitasun balio handia lortuta ere, ezin daiteke esan gainerako klaseetako kasuei erlazio falta esleitzeko joerarik ez duenik. Testean klase bakoitzeko dauden kasu kopurua oso ezberdina da, klase nagusia *no_relation* izanik (5.4 taula). Gauzak horrela, kasu guztiak erlaziorik ez dutela esanda ere, balio hori 0.77 izango litzateke. Zehazki 1-1 eszenarioan lortu duen balio bera. Eszenario honetan EN1 eta EN2 ereduen estaldura ezberdinen balioak konparatuz gero, ikusten da EN1 ereduarenak askoz hobeak direla. EN2k kasu gehienak *no_relation* eran sailkatu ditu. EN1ek berriz, hobeto banatu du sailkapena.

Hori bai, eszenarioa handitu ahala, EN2 ereduak gainerako klaseak hobeto ikasten ditu eta EN1 ereduarekin parekatzen da. Ez hori bakarrik, *no_relation* klasea hobeto topatzen du; eta klase hau, identifikatzen zailena izaten da. Beraz, esan daiteke entitate marketan oinarritutako ereduak hobeto funtzionatzen duela erlazio gabeko kasu gehiago ematen

bazaizkio entrenatzeko. Betiere, klase bakoitzeko kasu kopuru jakin bat edukiz gero. Gutxiegi edukita, gehienetan erlaziorik ez dagoela esaten duen eredu bat lortuko da.

Hala eta guztiz ere, EN2 eredua ez da probatu diren lau ereduen arteko hoberena. *F-score* zutabeari begiratu besterik ez dago horretaz jabetzeko. Bi eredu hoberenak A2T eta EM dira. Hauen artean berriz, egoera eta helburu ezberdinetarako bat edo beste da onena.

Nahiko garbi esan daiteke 1-1 eszenarioan A2T eredua hobea dela. Orokorrean klase guztiak askoz hobeto egiten ditu. Are gehiago, gainerako hiru ereduen *F-score*-en *macro average*-ei begiratuz gero, ikus daiteke *random* oinarri-lerroarena baino okerragoak direla (6.3 taula). Alegia, ikasketa gainbegiratuko algoritmo bat klaseko kasu bakar batekin entrenatzea *random* eredu bat garatzea baino okerragoa da. Beraz merezi du *few-shot* teknikak aplikatzeak.

Baina, eszenarioa handitu ahala, EM ereduak lortzen dituen emaitzek nabarmen gora egiten dute. Are gehiago, hasierako usteen kontra, eszenario hauetan EM ereduak A2T eredua gainditzen du. Ez da ahaztu behar esperimentu honen helburu nagusia, eszenarioak txikiak direnean A2T liburutegia teknika tradizionalak baino hobea dela frogatzea zela. Taulan agertzen diren emaitzek ez dute hori adierazten. Argi eta garbi, klaseen gehiengoa hobeto egiten du EM ereduak.

Baina, klaseak banan-banan aztertuz gero, fenomeno bitxi bat azter daiteke: *Produces* klasea eszenario guztietan hobeto egiten du A2T ereduak. Gainera, alde nabarmenarekin. *Anaphora* klasearekin antzekoa gertatzen da eta *Increases_risk_of* klasea nahiko parekatuta dago.

6.2 Eztabaida

Entitate-erauzleei dagokienez, ezer gutxi gehitu daiteke. Ikusita entitate-erauzle sendo bat garatzeko zailtasunak egongo zirela, beste ikerlari batzuk garatutako tresnak probatu dira eta ongi funtzionatu dute. Espero zen moduan, ez dute proiektu honetako ataza %100-ean betetzen. Azken finean, helburu ezberdinetarako garaturiko ereduak dira. Beraz, hirugarren pertsona batek garaturiko tresna bat erabiltzean, bakoitzak bere testuingurura eraman behar ditu honen emaitzak. Proiektu honetan, entitate-erauzleek behar baino entitate gehiago aurkitzen zituzten. Horren aurrean, horiek emandako soberako entitateak erlazio-erauzleak baztertuko zituela erabaki zen.

Bestalde, ikusi da erlazioak eraunzteko inferentzian oinarrituriko sistema batek ongi funtzionatzen duela medikuntzako testuetan. *MIMIC-III*-ko testuetan oso emaitza interesgarriak lortu dira (6.2 taula). Emaitza horiek erakusten dute datu etiketaturik ez dela behar testu klinikoetan erlazio erauzketa txukun bat egiteko. Nahikoa da erlazioaren inguruko ezagutza edukitzea. Oso ondorio aberatsa da hori. Txosten honetan behin baino gehiagotan aipatu den moduan, medikuntzaren domeinuan testu anotatuak aurkitzea zaila izaten baita.

6.2 taulako datuetatik atera daitekeen beste ondorioa inferentzia ereduak testuinguru txikiekin hobeto funtzionatzen duela da. Aurreko atalean esan bezala, zentzuduna da hori gertatzea. Azken finean aurre-entrenatutako NLI eredu bat erabiltzen du sistemak. NLI eredu horiek, orokorrean premisa eta hipotesi laburrak erabilita entrenatzen dira. Beraz, A2T liburutegia erabiltzean premisa (testuinguru) eta hipotesi (txantiloi) laburrak erabiltzea gomendatzen dira. Baina, ez A2T liburutegian bakarrik, oinarrian NLI eredu bat duten edozein sistemetan. NLI ereduek, erabilpen-kasu ugari dituzte. Kasu guztietan premisa eta

hipotesi motzak erabiltzea egokiagoa dela uste da.

RareDis corpusaren gainean egin diren esperimentu ezberdinei esker, inferentzia ereduek ere, zenbat eta datu gehiago eduki entrenatzeko, orduan eta emaitza hobeak bueltatzen dituztela ikusi da. Hori bai, datu gutxirekin ere emaitza oso onak lortzeko gai dira. Beraz, bigarren aldiz, frogatu da inferentzia ereduak egoki funtzionatzen duela testu klinikoekin. Bi corpus ezberdinetan emaitza interesgarriak erdietsi ditu, beraz, segurtasun handiz egin daiteke baieztapen hori.

Baina, benetan zer esana eman duena, A2T liburutegiaren eta teknika tradizionalen arteko konparaketa izan da. Helburu nagusia, eszenario txikietan A2T liburutegia eredu ezagunagoak baino hobea dela frogatzea zen. Ezin izan da hori frogatu ordea. Eszenario oso txikietan A2T hobea da. Halere, eszenarioa handitu ahala, EM ereduak A2T ereduak baino hobeto egiten ditu klaseen gehiengoak. Baina, *Produces* klasea eszenario guztietan hobeto egiten du A2T ereduak. *Anaphora* klasearekin antzekoa gertatzen da eta *Increases_risk_of* klasea nahiko parekatuta dago. Azken fenomeno horren atzean dagoen arrazoia ezagutu nahi izan da eta horretarako, hiru klase hauen eta gainerakoen artean zein ezberdintasun dagoen begiratu da.

Corpusean agertzen diren erlazio ezberdinen kasuak aztertu dira, eta *Produces*, *Anaphora* eta *Increases_risk_of* klaseetan izan ezik, beste guztietan klaseko adibideek patroi bat jarraitzen dutela ikusi da. Hau hobeto ulertuko da adibide batzuekin.

Is_a klaseko adibideei dagokienez, kasu gehienetan, esplizituki agertzen dira *is* eta *a* edo *an* hitzak. Segidan, hiru adibide aurkezten dira (Entitateak letra lodiz agertzen dira):

1. ***Anodontia*** *is a* ***genetic disorder***
2. ***Werner syndrome*** *is a rare* ***progressive disorder***
3. ***Ocular melanoma*** *is an extremely rare form of* ***cancer***

Klase honetako kasu gehienek itxura hori dute, oso errepikakorrak dira. Erregeletan oinarritzen den eredu batek ere topatuko luke erlazioa. Baldintzak bi entitateen artean *is a* edo *is an* agertzea eta euren arteko distantzia gehienez x hitzekoa izatea izango lirateke.

Is_acron klaseko kasuei dagokienez, gaixotasuna eta bere akronimoa elkarren ondoan joan ohi dira. Bigarren hau, parentesi artean agertzen da. Hona hemen hiru adibide:

1. ***Branchio-oculo-facial syndrome (BOFS)*** *is a rare genetic disorder*
2. ***Familial medullary thyroid carcinoma (FMTC)*** *is considered a third subtype*
3. *The exact cause of* ***pure autonomic failure (PAF)*** *is not known*

Beste behin, patroi konstante bat agertzen da, erregela bidez ere erraz aurki daitekeena. Baldintzak, elkarren jarraian egotea eta bigarrena parentesi artean agertzea izango lirateke.

Is_synon klasearen adibideetan berriz, gehienetan bi motatako ereduak aurkitzen dira. Batetik, gaixotasuna eta sinonimoa ingelesezko *also known as* esamoldeaz loturik agertzen diren kasuak daude. Bestetik, gaixotasunaren sinonimoa parentesi artean agertzen deneko kasuak agertzen dira (akronimoen antzera). Jarraian lau adibide aurkezten dira:

1. ***Wildervanck syndrome****, also known as* ***cervicooculoacoustic syndrome****, is a rare genetic disorder*

2. ***Schinzel syndrome**, also known as **ulnar-mammary syndrome**, is a rare inherited disorder*
3. *Kawasaki disease is an **acute multisystem inflammatory disease of blood vessels** (**vasculitis**)*
4. ***I-cell disease** (**mucolipidosis II**) is a rare inherited metabolic disorder*

Klase honen detekzioa apur bat zailtzen da bi patroi ezberdin aurki daitezkeelako, eta bietako bat *Is_acron* klasearekin partekatzen delako. Baina halere, nahiko automatikoa izaten jarraitzen du.

Hiru horiek dira oso egitura markatua duten eta EM ereduak askoz hobeto egiten dituen hiru klaseak. Gainerako hiru klaseei dagokienez, adibide ezberdinagoak topatzen dira klase berean.

Increases_risk_of klaseetako adibidean gaixotasun bat edukiz gero bestea egoteko arriskua dagoela orokorrean esplizituki esaten da. Hona hemen adibide batzuk:

1. *Repeated mild episodes of **acute cholecystitis** may result in **chronic cholecystitis***
2. ***Acquired Pure Red Cell Aplasia** may occur secondary to a **tumor of the thymus gland***
3. *Individuals of any age who have a **chronic, inflammatory condition** can potentially develop **the condition**.*

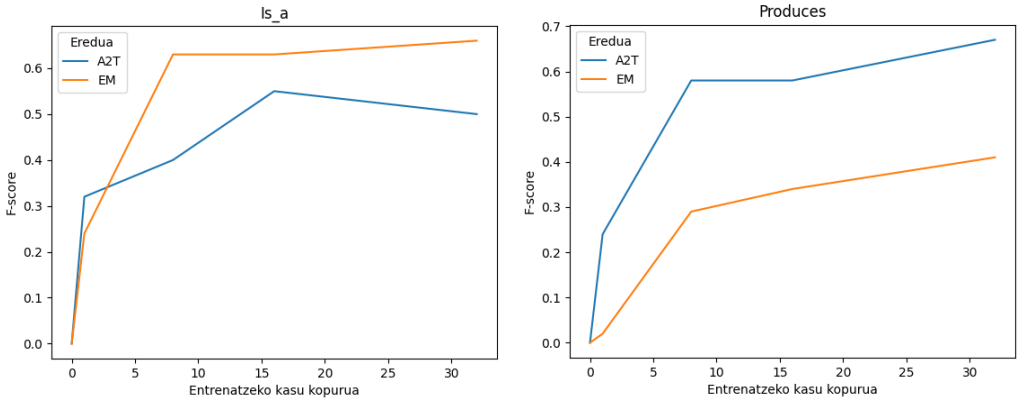
Egia esan, gauza bera esateko era ezberdinak erabiltzen dira. Nahiko egitura errepikakorra dute, baina, nolabait esateko testuingurua ulertzeko aditzen ezagutza handiagoa eduki behar da. Beti ez baita aditz bera erabiltzen.

Antzeko zerbait gertatzen da *Anaphora* klasearekin. Izatez, gaixotasun baten inguruan informazio bat ematen da eta ondoren gaixotasun horrentzat anafora bat erabilita informazio gehiago ematen da. Beraz, nahiko egitura markatua eduki ohi dute. Jarraian, hiru adibide aurkezten dira:

1. ***HS** affects 1 in 2,000 people in North America. **It** also occurs in other regions of the world, although not as well studied.*
2. ***Marden-Walker syndrome** is a rare connective tissue disorder that is inherited as an autosomal recessive trait. Patients with **this disorder** typically have a distinct facial expression*
3. ***Acute posterior multifocal placoid pigment epitheliopathy** (APMPPE) is a rare eye disorder of unknown (idiopathic) cause. **The disorder** is characterized by the impairment of central vision in one eye*

Egia da egitura antzekoa konpartitzen dutela adibide guztiek. Baina, bat bestearen anafora dela jakiteko hizkuntzaren ezagutza minimo bat eduki behar da. Ez da egitura kontua bakarrik.

Azkenik, *Produces* klasea gelditzen da. Guztietan klase aldrebesena. Bertako adibideek ez dute egitura konkretu bat jarraitzen eta gutxitan esaten dute esplizituki gaixotasun batek sintoma bat eragiten duela. Beste askoren artean adibide hauek topa daitezke bertan:



6.3 Irudia: Entrenatzeko klaseko kasu kopuruaren arabera *F-score*-aren joera

1. ***It** is caused by a spiral-shaped bacterium (spirochete). Symptoms include high fever, chills, **muscle aches** and jaundice.*
2. *Eye defects found in patients with **Marshall Syndrome** are nearsightedness, a disease of the eye in which the lens loses its clarity (**cataract**)*
3. *The primary form of **Dengue Fever** is characterized by a skin rash and a high fever with severe pain in the head and muscles. Bouts of **extreme exhaustion** may last for months after the initial symptoms.*

Esan den moduan, adibide hauetan ez da patroirik topatzen. Bakoitzak modu ezberdinean esaten du gaixotasunak sintoma konkretu bat produzitzen duela. Beraz, hauetan ez da nahikoa egitura ezagutzea, hizkuntza ulertzea ere eskatzen du.

Hortaz, hori da bi klase taldeen artean ikusten den ezberdintasun nagusia. EM ereduak hobeto egiten dituen klaseetako adibideek patroi konkretu bat jarraitzen dute. Oso sintaktikoak dira, eta ez dago hizkuntza naturala ezagutu beharrik ondo sailkatzeko. Egituraren inguruko informazioa nahikoa da. Aldiz, gainerako klaseak (bereziki *Produces*) semantikoak dira. Hauek sailkatzen jakiteko hizkuntza menderatu behar da.

Hori esanda, badirudi teknika tradizionalek testuko egitura sintaktikoak oso erraz ikasten dituztela. Baina, ezaugarri semantikoak ikasteko entrenamenduko datu gehiago behar dituztela ematen du. 6.3 irudian bi grafiko aurkezten da. Grafiko bakoitzak, klaseko entrenamenduko kasu kopurua handitu ahala azken bi eredu horiek *Is_a* edo *Produces* klasean lorturiko *F-score*-a nola hazten den erakusten du.

Is_a klaseari dagokionez, ikusten da EM ereduak oso kasu gutxirekin emaitza oso onak lortzen dituela. Eta, behin puntu horretara iritsita, ez du askoz gehiago hobetzen. Kasuak gehitu ahala ezer gutxi igotzen da *F-score*-a. Ematen du nolabaiteko muga batera iristen dela. A2T ereduak berriz, kasu oso gutxirekin EM ereduak baino emaitza hobeak lortzen dituen arren, datu gehiago hartzean hark baino gutxiago ikasten du. Gainera, honek ere, muga bat jotzen du eta ematen duenez kasuak gehituagatik ez du askoz gehiago ikasiko.

Erabat aurkakoa da *Produces* klasearen grafikoa. A2T liburutegiak kasu gutxiagorekin askoz gehiago ikasteko gaitasuna erakusten du. Baina, egia da bi ereduek zenbat eta datu

gehiago eman, orduan eta gehiago ikasten jarraitzen dutela. Gero eta gutxiago ikasten duten arren, itxura du kasu gehiago emanez gero *F-score*-a dezente hobetuko luketela oraindik. Bestalde, azpimarratu behar da *Produces* klasean A2Tk EMri ateratzen dion aldea handiagoa dela EMk A2Tri *Is_a*-n ateratzen diona baino.

Beraz, EMk testuaren egituraren ezaugarriak azkar ikasten ditu. Horregatik egiten ditu hain ondo *Is_a*, *Is_acron* eta *Is_synon* klaseak. Baina, semantika kontuak ikasteko arazoak ditu. Agian, entrenamendurako kasu asko erabilita ikasiko lituzte. Baina, eszenario txikietan ez du oso ondo egiten. Aldiz, A2Tk semantika ulertzeko gaitasuna erakusten du. Ondorioz, *Produces* klasea oso ondo egiten du. *Increases_risk_of* eta *Anaphora* tartean daude. Egitura bat jarraitu ohi dute, baina semantikaren inguruko ezagutzak ere laguntzen du. Horregatik, bi ereduetako inor ez da gailentzen.

Baina, A2T liburutegiak zergatik dauka semantika ulertzeko gaitasun hori? Oinarri teorikoetan 3.4 sekzioan baieztapen hau egiten zen: *NLIk erronka berriak proposatzen ditu: arrazoiketa formaleko kate luzeen ordez, arrazonamendu informala, ezagutza lexiko semantikoa eta hizkuntza-adierazpenaren aldakortasuna azpimarratzen dira.* Hau da, NLI sistema on bat, gai da arrazonamendu informala egiteko, ezagutza lexiko zein semantikoa ditu eta hizkuntza-adierazpenen aldaketetara ohiturik dago. Alegia, *Produces* klasea topatzeko beharrezko ezagutza guztia du. A2T liburutegia NLI sistema batean oinarritzen denez, gai da *Produces* gisako klaseak sailkatzeko.

Hau guztia esanda, gorago aipatzen zen ideia azpimarratzen da: egoera eta helburu ezberdinetarako eredu bat edo beste da onena. Kasu oso gutxi edukiz gero, eskuragarri dauden kasuak nolakoak diren aztertu beharko dira. Sintaktikoak edo semantikoak dira? Sintaktikoak izanez gero entitate marketan oinarritutako eredu batek ondo funtzionatuko du. Bestalde, semantikoak badira, A2T gisako liburutegi batekin emaitza hobeak lortuko dira. Aldiz, kasu asko edukiz gero, ziurrenik entitate marketan oinarritutako sistema batek gaitasuna izango du bi motetako erlazioak erauzten ikasteko.

MIMIC-III corpusera itzuliz (proiektuaren helburu nagusia), bertan bi motatako erlazioak aurki daitezke. Begirada azkar bat eginda, badirudi *Time*, *Duration*, *Localization*, *OnsetCirc* eta *Type* nahiko sintaktikoak direla. Beraz, ez dute beharko A2T liburutegia. Baina, gainerako klaseentzat, printzipioz interesgarriagoa izango litzateke NLI eredu bat garatzea.

Hortaz, ez da frogatu eszenario txikietan A2T liburutegia teknika tradizionalak baino hobea denik (ez da gurpila asmatu). Baina bi datu interesgarri lortu dira. Batetik, A2T liburutegiak erlazio semantikoak oso ondo ikasteko gaitasuna du. Hizkuntza ulertzeko ahalmen handia erakusten du. Bestetik, erlazio sintaktikoak ikasteko oso datu etiketatu gutxi behar direla ikusi da. Erlazio mota horientzat, datu askorekin eta datu gutxirekin ikasitako bi sistemek lortuko lituzketen ematzek ezberdintasun gutxi izango dituztela aurreikusten da.

1-1												
Klasea	Doitasuna				Estaldura				F-score			
	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2
Prod	0.36	**0.41**	0.32	0.35	0.18	0.01	**0.19**	0.00	**0.24**	0.02	0.23	0.01
Ana	**0.49**	0.43	0.17	0.42	**0.82**	0.16	0.68	0.09	**0.61**	0.24	0.28	0.15
In_ris	0.02	**0.15**	0.02	0.03	**0.35**	0.02	0.23	0.01	**0.04**	0.03	**0.04**	0.01
Is_sy	0.01	0.00	0.00	**0.03**	**0.82**	0.06	0.32	0.06	0.02	0.01	0.00	**0.04**
Is_a	**0.47**	0.18	0.05	0.40	0.25	0.37	**0.66**	0.10	**0.32**	0.24	0.09	0.17
Is_ac	0.00	0.03	**0.07**	0.03	0.00	0.05	**0.67**	0.01	0.00	0.04	**0.12**	0.01
no_re	0.78	0.79	**0.95**	0.77	0.65	0.92	0.07	**0.99**	0.71	0.85	0.13	**0.87**
avg	**0.30**	0.28	0.23	0.29	**0.44**	0.23	0.40	0.18	**0.28**	0.20	0.13	0.18
8-4												
Klasea	Doitasuna				Estaldura				F-score			
	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2
Prod	**0.46**	0.24	0.27	**0.46**	**0.79**	0.35	0.75	0.40	**0.58**	0.29	0.39	0.43
Ana	0.61	**0.68**	0.25	0.42	**0.90**	0.53	0.86	0.72	**0.73**	0.59	0.39	0.53
In_ris	0.05	**0.17**	0.02	0.04	0.46	0.24	**0.58**	0.25	0.09	**0.20**	0.05	0.06
Is_sy	0.01	**0.08**	0.02	0.03	0.62	0.32	**0.74**	0.59	0.02	**0.13**	0.04	0.06
Is_a	0.27	**0.75**	0.31	0.36	**0.75**	0.55	0.74	0.70	0.40	**0.63**	0.44	0.47
Is_ac	0.10	**0.62**	0.18	0.28	0.69	0.66	**0.88**	0.76	0.17	**0.64**	0.30	0.41
no_re	0.92	0.81	**0.95**	0.87	0.54	0.75	0.24	**0.76**	0.68	0.78	0.38	**0.81**
avg	0.35	**0.48**	0.28	0.35	**0.68**	0.49	**0.68**	0.60	0.38	**0.47**	0.23	0.39
16-8												
Klasea	Doitasuna				Estaldura				F-score			
	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2
Prod	**0.58**	0.25	0.39	0.51	0.59	0.54	**0.82**	0.60	**0.58**	0.34	0.53	0.55
Ana	0.53	**0.69**	0.25	0.48	**0.96**	0.70	0.91	0.80	**0.69**	**0.69**	0.39	0.60
In_ris	0.08	**0.13**	0.03	0.04	0.32	0.41	**0.63**	0.46	0.13	**0.19**	0.05	0.07
Is_sy	0.01	**0.11**	0.02	0.03	0.62	0.62	**0.79**	0.62	0.02	**0.18**	0.04	0.05
Is_a	0.48	**0.64**	0.41	0.40	0.63	0.62	**0.73**	0.72	0.55	**0.63**	0.53	0.52
Is_ac	0.10	**0.43**	0.16	0.28	0.83	0.79	0.83	**0.84**	0.18	**0.56**	0.27	0.42
no_re	0.88	0.84	**0.97**	0.91	0.69	0.61	0.38	**0.72**	0.77	0.71	0.55	**0.80**
avg	0.38	**0.44**	0.32	0.38	0.66	0.61	**0.72**	0.68	0.41	**0.47**	0.34	0.43
32-16												
Klasea	Doitasuna				Estaldura				F-score			
	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2	A2T	EM	EN1	EN2
Prod	**0.52**	0.30	0.42	0.48	**0.92**	0.69	0.85	0.82	**0.67**	0.41	0.56	0.60
Ana	0.57	**0.65**	0.34	0.27	0.91	0.80	0.92	**0.96**	0.71	**0.72**	0.50	0.43
In_ris	**0.06**	0.04	0.02	0.04	0.58	0.67	**0.76**	0.68	**0.11**	0.08	0.04	0.07
Is_sy	0.05	**0.17**	0.04	0.05	**0.76**	0.65	**0.76**	0.68	0.09	**0.28**	0.07	0.10
Is_a	0.37	**0.83**	0.56	0.44	**0.80**	0.55	0.68	0.79	0.50	**0.66**	0.62	0.57
Is_ac	0.10	**0.40**	0.16	0.36	**0.92**	0.85	0.88	0.89	0.19	**0.55**	0.27	0.52
no_re	**0.97**	0.87	**0.97**	0.96	**0.61**	0.56	0.40	0.54	**0.75**	0.68	0.57	0.70
avg	0.38	**0.46**	0.35	0.37	**0.79**	0.68	0.75	0.77	0.43	**0.48**	0.43	0.43

6.4 Taula: Eredu ezberdinen arteko emaitzen konparaketa

,

7. Plangintzaren desbiderapenak

Proiektu honetan 2.3 sekzioan azaldutako arrisku nagusiena bete da: Ez da lortu corpus etiketatu handirik. Osakidetzako testuak lortu eta etiketatu nahi ziren. Baina, gaur egun, oraindik ez dira lortu. Honetaz gain, *MIMIC-III*-ko ingelesezko testuekin osatutako corpusak bere horretan jarraitzen du.

Arrisku hau gertatzeko probabilitate oso handia aurreikusi zenez, plangintza alternatibo bat zegoen garatuta (2.3 sekzioa). Hori dela eta, horri heldu zaio. Beraz, desbiderapen nagusi bat egon da, baina, aurrez ondo definituta zegoena. Plangintza alternatibo horrentzat ere, *gantt* diagrama eta ataza bakoitzeko igaroko zen denbora biltzen zituen taulak eginda zeuden. Proiektuaren garapenean, benetan zein diagrama eta zein denborak egin diren erregistratu da.

7.1 taulan, benetan jarraitu den *gantt* diagrama aurkezten da. Hasieran definiturikoarekin konparatuz gero (2.1 eta 2.3 bateratuta) oso antzekoak direla ikus daiteke. Ataza guztien denboraldiak aldez aurretik definitu bezala jarraitu dira. Are gehiago, zenbait ataza, espero zena baino azkarrago amaitu dira. Denboraldien informazio zehatzagoa eduki nahi izanez gero, D eranskinean egindako bilera guztien akten laburpen bat aurkezten da.

Ezin daiteke gauza bera esan denborekin. 7.2 taulan agertzen dira berez ataza bakoitzarekin egin diren orduen eta 2.2 eta 2.4 tauletan definitutakoen arteko konparaketa bat. Oso ezberdinak direla ikusten da. Lan batzuekin askoz denbora gehiago behar izan da eta beste batzuekin gutxiago. Deigarrienak ETF.2, TAK.2 eta M.4 atazak dira.

Ez da harritzekoa hori gertatzea. Proiektu hau ikerketa lan bat izan da. Ikerketa lanetan oso zaila izaten da aurrez plangintza zehatz bat definitzea. Lana garatzen doan bitartean esperimentu baten emaitzek bidea erabat alda baitezakete. Ondorengo esperimentua zein izango den defini dezakete. Eta esperimentu ezberdinek denbora ezberdinak behar dituzte. Gutxi gora-beherako bide bat defini daiteke baina ez plangintza osoa zehatz-mehatz.

Hala eta guztiz ere, denbora desbiderapen horiek egonda ere, proiektuan guztira 318 ordu igaro dira. Hasiera batean 300 ordu igaroko zirela definitu zen. Beraz, esperotako denbora baino %6 gehiago igaro da. Hain proiektu handia izanik, ez da hainbeste.

Hortaz, desbiderapen asko egon dira. Baina, hasiera batetik plangintza egoki bat eginda zegoenez, proiektua ongi bukatzeko gaitasuna eduki da.

Ataza	Hilabeteak						
	abu.	...	urt.	ots.	mar.	api.	mai.
EnEDG.1							
EnEDG.2							
ErEDG.1							
ErEDG.2							
ETF.1							
ETF.2							
ETF.3							
TAK.1							
TAK.2							
TAK.3							
M.1							
M.2							
M.3							
M.4							
D.1							
D.2							
D.3							
P.1							
P.2							
P.3							
JK.1							
JK.2							

7.1 Taula: Benetan gauzatu den *gantt* diagrama

Lan-paketea	Ataza	Esperotako denbora (h)	Benetako denbora (h)
Proiektuaren garapena		200	200
Entitate erauzketa datu gutxirekin	EnEDG.1	20	15
	EnEDG.2	30	25
Erlazio erauzketa datu gutxirekin	ErEDG.1	20	10
	ErEDG.2	30	40
Erabilitako teknikak frogatzea	ETF.1	5	2
	ETF.2	35	80
	ETF.3	10	4
Teknika arruntekin konparaketa	TAK.1	5	1
	TAK.2	30	20
	TAK.3	15	3
Dokumentazioa		70	91
Memoria	M.1	1	1.5
	M.2	2	2
	M.3	2	1.5
	M.4	50	75
Defentsa	D.1	3	2
	D.2	5	4
	D.3	7	5
Proiektuaren kudeaketa		30	27
Plangintza	P.1	6	6
	P.2	6	6
	P.3	6	2
Jarraipen eta kontrola	JK.1	2	1
	JK.2	10	12
Guztira		300	318

7.2 Taula: Benetan ataza bakoitzarekin igarotako denbora

8. Ondorioak eta etorkizuneko lanak

Helburu nagusia lortu al da? Neurri batean bai eta bestean ez. Helburu nagusia, Osakidetzako bularraldeko minarekin erlazionaturiko Larrialdi Zerbitzuetako alta-txostenetako sintomen deskribapenetan datu kritikoen presentzia eta absentzia detektatzeko gai izango zen sistema bat garatzea zen. Hasteko, Osakidetzako (gaztelaniazko) testu etiketaturik ez da lortu. Beraz, ezinezkoa izan da hori egitea.

Osakidetzako testuen partez, *MIMIC-III* corpuseko ingelesezko hainbat testu etiketatu batzuk lortu dira. Baina, sistema tradizional bat garatzeko gutxiegi zirenaren irudipena eduki da. Hori dela eta, honelako egoera batean jorratu daitezkeen bide ezberdinak esploratu dira.

Entitateen detekzioa egiteko, dagoeneko publikatuta zeuden hainbat tresna probatu dira: *Metamap* eta *SUTime*. Eta ikusi da emaitza nahiko altuak lortzen dituztela. Hortaz, honelako egoera baten aurrean egonez gero beste ikerlariek garaturiko tresnen inguruko ikerketa bat egitea gomendatzen da.

Aldiz, erlazio erauzketarako *Ask2Transformers* liburutegia sakonki aztertu da. Orain arteko ikerketetan, bestelako domeinuetan ongi funtzionatzen duen tresna bat dela frogatuta zegoen. Proiektu honetan, medikuntzaren domeinura ekarri da. Eta, bertan ere ondo funtzionatzen duela ikusi da. Hain emaitza onak lortzen ditu, non pentsatzen baitzen eszenario txikietan teknika tradizionalek baino askoz emaitza hobeak lortzeko gai izango zela. Esperimentazio prozesu luze baten ostean ikusi da ezetz. Eredu tradizional batzuk baino hobea baden arren, era erraz batean bera baino hobea den eredu bat lor daiteke (EM eredua). Baina, ez %100-ean bera baino hobea den bat.

Ask2Transformers liburutegiaren indargune bat aurkitu da. Datu etiketatu gutxi erabiliz, erlazio semantikoak identifikatzen ongi ikasten du. Alegia, erlazioa existitzen den edo ez erabakitzeko semantika ulertu behar denean emaitza onak ematen ditu. Hori bai, zenbat eta datu etiketatu gehiago lortu, orduan eta emaitza hobeak emango ditu. Era berean, ikasketa gainbegiratuko zenbait algoritmok oso sintaktikoak diren erlazioak kasu etiketatu gutxi erabiliz ondo ikasteko ahalmena dutela ikusi da. Are gehiago, datu-etiketatu asko jarrita ere, ez dirudi emaitzak asko aldatuko direnik. Indargune hori oso interesgarria da. Izan ere, semantika "ulertze" horrek erakusten du benetan badagoela Adimen Artifiziala. Adimen Artifizialean benetan interesekoak diren erlazioak semantikoak dira.

Egia esan, oraindik proiektua irekita dago. Osakidetzako medikuek sistema garatua ikusi nahi dute. Baina, orain beste ikuspuntu batekin begiratu daiteke. Proiektu honek duen arazo nagusia datu falta da. Baina, ikerketa honetan, ikusi da agian ez direla datu etiketatu asko behar.

Etorkizuneko helburua Osakidetzako gaztelaniazko testuak lortzea da. Hauek lortzen direnean, etiketatu egin beharko dira. Anotazio prozesu hori, ikerketa lan hau aintzat hartuta egin daiteke. Erlazioak bi taldetan sailka daitezke: sintaktikoak eta semantikoak. Sintaktikoen kasuan, ikusi da oso kasu etiketatu gutxi beharko liratekeela. Emaitzen eztabaidan esan den moduan, pentsatzen da erlazio sintaktiko horiek *Time*, *Duration*,

Localization, *OnsetCirc* eta *Type* direla. Baina, beharbada gaztelaniazko testuetan aldatu egingo dira. Semantikoen kasuan berriz, gehixeago etiketatuz gero emaitza hobeak lortuko lirateke. Erlazioak sailkatuta, bi eredu garatuko lirateke, A2T eredu bat semantikoentzat eta entitate marketan oinarritutako bat sintaktikoentzat. Proiektu honetako esperientzian oinarrituz, sistema polit bat lortuko litzateke.

Bestalde, proiektu hau bereziki erlazio erauzketaren ikerketara bideratu da. Baina, A2T liburutegiak entitate erauzketa egiteko bidea ere eskaintzen du. Ikerketa honetan aurkitzeko gelditu diren klaseak topatzeko balio dezake. Behin esperimentu txiki bat egin zen eta ongi topatzeko gaitasuna baduela zirudien. Gaztelerazko testuetako entitateak topatzeko gaitasuna ere eduki dezake. Baina, hori egiaztatzeko hemen erlazioekin egin denaren antzeko ikerketa bat egitea eskatzen du horrek.

Laburtuz, proiektu honetatik atera daitekeen ondorio garrantzitsuenetako bat hau da: datu gutxirekin Adimen Artifiziala garatzeko aukera dago.

Eranskinak

A eranskina

Eranskin honetan A2T liburutegia *MIMIC-III* corpusean erabiltzean definituriko txantiloi guztiak ikus daitezke.

```
templates = {
    "ImprovePosNeg": [
        "{X} improved with {Y}",
        "{X} got worse with {Y}",
        "{X} is improved by taking {Y}",
        "{X} doesn't improve taking {Y}",
        "{X} decreased taking {Y}",
        "{X} increased taking {Y}",
        "Patient was given {Y} for the {X}",
        "After {Y} {X} improved",
        "After {Y} {X} got worse",
        "{X} is {Y}",
        "{X} relieved {Y}",
        "{X} got worse {Y}"
    ],
    "Duration": [
        "{X} persist for {Y}",
        "{X} durated {Y}",
        "{X} was present for {Y}"
    ],
    "Radiation": [
        "{X}'s radiation is {Y}",
        "{X} is {Y}"
    ],
    "Sympt": [
        "{Y} is a symptom associated with {X}",
        "{Y} is a symptom related with {X}"
        "{X} includes {Y}",
        "In addition to {X} has {Y}"
    ],
    "Time": [
        "{X} began {Y}",
        "{X} began in {Y}",
        "{X} began on {Y}",
        "{X} appeared {Y}",
        "{X} appeared in {Y}",
        "{X} appeared on {Y}",
```

```
                "{X} occurred {Y}",
                "{X} occurred in {Y}",
                "{X} occurred on {Y}",
            ],
            "Localization": [
                "{X} is localized in {Y}",
                "{X} was localized in {Y}",
                "{X}'s localization is {Y}",
                "{X}'s localization was {Y}"
            ],
            "Type": [
                "{X}'s type is {Y}",
                "{X} is {Y}"
            ],
            "OnsetGrad": [
                "{X}'s onset grad is {Y}",
                "{Y} is related with {X}"
            ],
            "OnsetCirc": [
                "{X}'s onset circ is {Y}",
                "{Y} is related with {X}"
            ]
}
```

B eranskina

Eranskin honetan A2T liburutegia *RareDis* corpusean erabiltzean definituriko baldintza guztiak ikus daitezke.

```
valid_conditions = {
    'Produces': [
        'ANAPHOR:SIGN',
        'ANAPHOR:SYMPTOM',
        'DISEASE:SIGN',
        'DISEASE:SYMPTOM',
        'RAREDISEASE:SIGN',
        'RAREDISEASE:SYMPTOM',
        'SIGN:SIGN',
        'SKINRAREDISEASE:SIGN',
        'SKINRAREDISEASE:SYMPTOM',
    ],
    'Anaphora': [
        'DISEASE:ANAPHOR',
        'RAREDISEASE:ANAPHOR',
        'SIGN:ANAPHOR',
        'SKINRAREDISEASE:ANAPHOR',
    ],
    'Increases_risk_of': [
        'DISEASE:ANAPHOR',
        'DISEASE:DISEASE',
        'DISEASE:RAREDISEASE',
        'DISEASE:SKINRAREDISEASE',
        'RAREDISEASE:ANAPHOR',
        'RAREDISEASE:DISEASE',
        'RAREDISEASE:RAREDISEASE',
        'SKINRAREDISEASE:RAREDISEASE',
    ],
    'Is_synon': [
        'DISEASE:DISEASE',
        'RAREDISEASE:RAREDISEASE',
        'SKINRAREDISEASE:SKINRAREDISEASE',
    ],
    'Is_a': [
        'ANAPHOR:DISEASE',
        'DISEASE:DISEASE',
        'RAREDISEASE:ANAPHOR',
        'RAREDISEASE:DISEASE',
        'RAREDISEASE:RAREDISEASE',
        'SKINRAREDISEASE:DISEASE',
        'SKINRAREDISEASE:SKINRAREDISEASE',
    ],
    'Is_acron': [
```

```
                'DISEASE:DISEASE',
                'RAREDISEASE:RAREDISEASE',
                'SKINRAREDISEASE:SKINRAREDISEASE',
            ],
        }
```

C eranskina

Eranskin honetan A2T liburutegia *RareDis* corpusean erabiltzean definituriko txantiloi guztiak ikus daitezke.

```
templates = {
    'Produces': [
        "{X} produces {Y}",
        "{X} leads to a {Y}",
        "{X} leads to an {Y}",
        "{Y} is a sign or a symptom produced by {X}",
        "{X} typically indicates {Y}",
        "{X} may produce {Y}",
        "People who have {X} contain {Y}",
        "In addition to {X} appears {Y}",
        "{X} results in {Y}"
    ],
    'Anaphora': [
        "{Y} is a reference to {X}",
        "{Y} refers to {X}",
        "{Y} is the disease mentioned above {X}",
        "'{Y}' has been used to refer to {X}"
    ],
    'Increases_risk_of': [
        "{X} increases risk of {Y}",
        "{X} may result in {Y}",
        "{X} can generate {Y}",
        "{X} may suggest {Y}",
        "{X} can be the trigger for {Y}",
        "{Y} is caused by {X}",
        "{X} is a risk factor for {Y}",
        "{X} may also make the patient prone to {Y}",
        "People who have {X} tend to have {Y}"

    ],
    'Is_synon': [
        "{X} is a synonym for {Y}",
        "{X} is also known as {Y}",
        "{X} is also called {Y}",
        "{Y} is also called {X}",
        "The disease is known as {X} or {Y}",
        "{X} and {Y} are synonyms",
        "{X} and {Y} are the same disease"
    ],
    'Is_a': [
        "{X} is {Y} and {Y} is {X}",
        "It has been said that {X} is {Y}",
        "It has been said that {X} is a {Y}",
```

```
                "It has been said that {X} is an {Y}",
                "{X} is a {Y}",
                "{X} is an {Y}",
                "{X} is {Y}",
                "{Y} is a {X}",
                "{Y} is an {X}",
                "{Y} is {X}",
        ],
        'Is_acron': [
                "{X} contains the initials of {Y},
                "{Y}'s acronym is {X}"
        ],
}
```

D eranskina

Eranskin honetan proiektuaren garapenean zehar egindako bilera guztien akten laburpen txiki bat azaltzen da.

2023-01-24

- Problema azaldu da.
- Problemari aurre egiteko modu bat: Ataza guztia sistema bakar batek egin beharrean, entitate ezberdinak topatzeko sistema ezberdinak erabiltzea eta jarraian emaitzak bateratzea.
- Eskuragarri dagoen medikuek etiketaturiko corpusa aztertu behar da: Klase bakoitzeko zenbat entitate? Entitate hauen artean, zenbat daude *MainSympt*-ekin erlazionatuta? *MainSympt* eta alarma-sintomaren artean batez beste zein distantzia dago?

2023-02-03

- Problemari aurre egiteko modu bat: Lehenbizi entitate guztiak identifikatu. Ondoren, enitateak MainSympt-ekin loturik dauden edo ez erabaki sailkatzaile bitar batekin.
- Dagoeneko publikatuta dauden hainbat tresna, corpuseko entitate batzuk detektatzeko gai izango direla iruditzen da (*Metamap*, *Ctakes*, *SUTime*...). Zehazki *MainSympt*, *Sympt*, *Time* eta *Duration*.
- Bestelako entitateak identifikatzeko *Ask2Transformers* eta *ZS4IE* tresnak proba daitezke.
- Antzeko corpusik ba al dago nonbait etiketatuta?

2023-02-10

- *Metamap* eta *SUTime* tresnen emaitzak onak direla ikusi da.
- *Ask2Transformers* tresnarekin esperimentu txiki bat egin da erlazioak bilatzeko eta badirudi ondo dabilela. Proba handiagoak egin behar zaizkio. Corpuseko 20 testuak bi taldetan banatuko dira garapena (5 testu) eta testa (15). Momentuz garapeneko 5 testuak ebaluatuko dira klaseka. Ondoren *fine-tuning* bat egin ahal izango da.

2023-02-17

- Ask2Transformer-ekin lortu diren emaitzak onak dirudite. Baina, erlazio eza klasearen emaitzak tauletan hobeto jarri behar dira. Momentuz klase bakar bezala erakutsi da. Hobea da klase bakoitzeko *no_relation*-ak nola egin dituen adieraztea.
- Atalasearekin jokatu daiteke erlazioa existitzen den edo ez erabakitzeko. Garapenerako zein izango litzateke atalase egokiena?

- Antzeko corpusik ba al dago? *RareDis*?

2023-02-24

- Ask2Transformer-ekin lortu diren emaitzak onak dira. Halere, atalasearen kontuak ez dirudi laguntzen duenik. Aurkitutako TP asko gal daitezke.
- Esperimentuak handitzeko asmoz, orain arte *MIMIC-III* corpusean probaturiko teknikak (entitate erauzleak eta A2T) RareDis corpus-ean probatuko dira.
- Oraindik harrapatu gabeko zenbait klase daude *MIMIC-III* corpusean. Hauek A2Trekin topa daitezkeen begiratu da eta dezente aurkitu ditu (aurkituak/guztiak): *Localization* 7/13, *ImprovePosNeg* 8/18 eta *OnsetCirc* 8/11. Hori bai, honek testua entitate posibletan zatitzeko tresnaren bat eskatzen du.

2023-03-03

- *Metamap*-ek *RareDis* corpuseko *RAREDISEASE*, *DISEASE* eta *SKINRAREDISEASE* klaseetako entitate dezente topatzen ditu eta *SIGN* eta *SYMPTOM*-eko gutxi batzuk.
- A2T liburutegiarekin *RareDis* corpuseko train zatiaren gainean *zero-shot* bat egin da.
- A2Tren emaitzak onak dira, baina, hobeak izango dira *valid condition* gutxiago jarriz gero.
- Esperimentua formalizatzea erabaki da. Lau eszenario ezberdin sortuko dira: 1-1, 8-4, 16-8 eta 32-16. Eszenario bakoitzeko, NLI sistemaren *fine-tuning*-a egiteko kasuak sortu beharko dira.

2023-03-17

- *Valid condition*-ak gutxituta emaitza askoz hobeak lortu dira.
- *fine-tuning*-a egiteko kasuak sortu dira. Orain eszenario bakoitzean entrenamenduak egin behar dira.

2023-03-24

- *fine-tuning*-ak egin dira. Baina, ez da lortu entrenamenduko eredu hoberena hartzea (azkenekoa hartu da).
- Emaitzek onak dirudite. Baina, testean arazo bat egon da. Erabili den testa entrenamenduko kasuez kutsatuta dago. Beraz, eszenario ezberdinen arteko konparaketak ez dira bidezkoak.

2023-03-31

- Lortu diren emaitzak oso interesgarriak dirudite. Kasu gutxi erabilita benetan eredu interesgarriak lortu dira.

- Orain, A2T eredua teknika tradizionalekin konparatu nahi da. Horretarako, entitate marketan oinarrituriko eredu bat garatuko da. Erlazio eza modelatzeko bi modu daude.

2023-04-14

- Entitate marketan oinarrituriko hiru eredu garatu dira. Lehenbizikoarekin, erlazio eza inplizituki duenarekin, atalasearen ikerketa sakon bat egin da.
- Emaitzak aztertu dira eta nahi ez bezala, entitate marketan oinarrituriko ereduak hobea direla dirudi.
- Halere, ikusi da erlazio semantikoak hobeto egiten dituela A2Tk entitate marketan oinarriturikoak baino.

E eranskina

Eranskin honetan, proiektuan zehar egindako esperimentu ezberdinetan lorturiko errore-matrizeak aurki daitezke. Hona hemen esperimentu guztiak:

- *MIMIC-III* corpusean *Ask2Transformers* liburutegiarekin egindako *zero-shot*-ak bi testuinguru motekin.
- *Random* oinarri lerroa *RareDis* corpusean.
- *Zero-shot* eredua *RareDis* corpusean.
- A2T ereduarekin egindako *few-shot* ezberdinak *RareDis* corpusean.
- Entitate marketan oinarrituta dagoen eta erlazio eza inplizituki definiturik duen ereduarekin egindako *few-shot* ezberdinak *RareDis* corpusean.
- Entitate marketan oinarrituta dagoen eta erlazio eza esplizituki definiturik duten bi ereduekin egindako *few-shot* ezberdinak *RareDis* corpusean.

Sympt

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	24	8	
no_rel	3	3	
Doit.	0.75	0.50	0.62
Estal.	0.89	0.27	0.58
F	0.81	0.35	0.58

Time

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	2	10	
no_rel	0	3	
Doit.	0.17	1.00	0.58
Estal.	1.00	0.23	0.62
F	0.29	0.38	0.33

Duration

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	1	10	
no_rel	0	1	
Doit.	0.09	1.00	0.55
Estal.	1.00	0.09	0.55
F	0.17	0.17	0.17

OnsetCirc

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	2	0	
no_rel	1	2	
Doit.	1.00	0.67	0.83
Estal.	0.67	1.00	0.83
F	0.80	0.80	0.80

OnsetGrad

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	0	1	
no_rel	0	0	
Doit.	0.00	1.00	0.50
Estal.	1.00	0.00	0.50
F	0.00	0.00	0.00

Radiation

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	3	0	
no_rel	0	0	
Doit.	1.00	1.00	1.00
Estal.	1.00	1.00	1.00
F	1.00	1.00	1.00

ImprovePosNeg

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	3	1	
no_rel	3	1	
Doit.	0.75	0.25	0.50
Estal.	0.50	0.50	0.50
F	0.60	0.33	0.47

Type

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	0	1	
no_rel	0	1	
Doit.	0.00	1.00	0.50
Estal.	1.00	0.50	0.75
F	0.00	0.67	0.33

Localization

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	1	2	
no_rel	2	2	
Doit.	0.33	0.50	0.42
Estal.	0.33	0.50	0.42
F	0.33	0.50	0.42

1 Taula: *Zero-shot MIMIC-III* corpusean bi esaldiko testuingurua erabilita

Sympt

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	18	7	
no_rel	9	4	
Doit.	0.72	0.31	0.51
Estal.	0.67	0.36	0.52
F	0.69	0.33	0.51

Time

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	2	8	
no_rel	0	5	
Doit.	0.20	1.00	0.60
Estal.	1.00	0.38	0.69
F	0.33	0.56	0.44

Duration

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	1	6	
no_rel	0	5	
Doit.	0.14	1.00	0.57
Estal.	1.00	0.45	0.73
F	0.25	0.62	0.44

OnsetCirc

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	2	0	
no_rel	1	2	
Doit.	1.00	0.67	0.83
Estal.	0.67	1.00	0.83
F	0.80	0.80	0.80

OnsetGrad

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	0	1	
no_rel	0	1	
Doit.	1.00	1.00	1.00
Estal.	1.00	1.00	1.00
F	1.00	1.00	1.00

Radiation

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	3	0	
no_rel	0	0	
Doit.	1.00	1.00	1.00
Estal.	1.00	1.00	1.00
F	1.00	1.00	1.00

ImprovePosNeg

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	2	1	
no_rel	4	1	
Doit.	0.67	0.20	0.43
Estal.	0.33	0.50	0.42
F	0.44	0.29	0.37

Type

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	0	1	
no_rel	0	1	
Doit.	0.00	1.00	0.50
Estal.	1.00	0.50	0.75
F	0.00	0.67	0.33

Localization

Ir. kl. \ Ben. klasea	rel	no_rel	
rel	1	3	
no_rel	2	1	
Doit.	0.25	0.33	0.29
Estal.	0.33	0.25	0.29
F	0.28	0.28	0.29

2 Taula: *Zero-shot MIMIC-III* corpusean testuinguru osoa erabilita

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	2045	0	0	0	0	0	5708	
Ana	0	302	11	0	9	0	677	
In_ris	0	237	34	5	178	24	1997	
Is_sy	0	0	4	5	29	36	822	
Is_a	0	220	8	7	211	31	1539	
Is_ac	0	0	6	10	28	37	838	
no_re	2000	314	48	7	204	35	8419	
Doitasuna	0.26	0.30	0.01	0.01	0.10	0.04	0.76	0.21
Estaldura	0.51	0.28	0.31	0.15	0.32	0.23	0.42	0.32
F	0.35	0.29	0.03	0.01	0.16	0.07	0.54	0.21

3 Taula: *Random* oinarri lerroa *RareDis* corpusean

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	4003	0	0	0	0	0	10505	
Ana	0	948	9	0	17	0	1323	
In_ris	0	40	91	1	60	1	2051	
Is_sy	0	0	1	17	21	11	580	
Is_a	0	72	4	1	455	4	797	
Is_ac	0	0	4	15	92	147	2639	
no_re	42	13	2	0	14	0	2105	
Doitasuna	0.28	0.41	0.04	0.03	0.34	0.05	0.97	0.30
Estaldura	0.99	0.88	0.82	0.50	0.69	0.90	0.11	0.70
F	0.43	0.56	0.08	0.05	0.46	0.10	0.19	0.27

4 Taula: *Zero-shot RareDis* corpusean (*Deberta*)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	4036	0	0	0	0	0	11299	
Ana	0	769	8	0	21	0	1541	
In_ris	0	66	93	2	72	3	2899	
Is_sy	0	0	3	23	34	103	1626	
Is_a	0	232	5	9	525	50	1873	
Is_ac	0	0	1	0	3	7	445	
no_re	9	6	1	0	4	0	317	
Doitasuna	0.26	0.33	0.03	0.01	0.19	0.02	0.94	0.25
Estaldura	0.99	0.72	0.84	0.68	0.80	0.04	0.02	0.58
F	0.42	0.45	0.06	0.03	0.31	0.02	0.03	0.19

5 Taula: *Zero-shot RareDis* corpusean *(Roberta)*

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	721	0	0	0	0	0	1297	
Ana	0	876	15	0	21	0	884	
In_ris	0	181	39	1	119	2	1368	
Is_sy	0	0	14	28	104	149	3196	
Is_a	0	1	2	0	163	1	179	
Is_ac	0	0	0	0	0	0	29	
no_re	3324	15	41	5	252	11	13047	
Doitasuna	0.36	0.49	0.02	0.01	0.47	0.00	0.78	0.33
Estaldura	0.18	0.82	0.35	0.82	0.25	0.00	0.65	0.44
F	0.24	0.61	0.04	0.02	0.32	0.00	0.71	0.28

6 Taula: A2T *RareDis* corpusean (1-1 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	3213	0	0	0	0	0	3786	
Ana	0	970	8	0	12	0	609	
In_ris	0	25	51	0	6	1	903	
Is_sy	0	0	8	21	54	31	1685	
Is_a	0	52	15	1	496	1	1244	
Is_ac	0	0	2	5	29	113	1004	
no_re	832	26	27	7	62	17	10769	
Doitasuna	0.46	0.61	0.05	0.01	0.27	0.10	0.92	0.35
Estaldura	0.79	0.90	0.46	0.62	0.75	0.69	0.54	0.68
F	0.58	0.73	0.09	0.02	0.40	0.17	0.68	0.38

7 Taula: A2T *RareDis* corpusean (8-4 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	2382	0	0	0	0	0	1733	
Ana	0	1032	8	0	19	0	881	
In_ris	0	13	36	0	1	0	376	
Is_sy	0	0	12	21	66	13	1743	
Is_a	0	5	7	0	418	0	432	
Is_ac	0	0	1	6	24	135	1133	
no_re	1663	23	47	7	131	15	13702	
Doitasuna	0.58	0.53	0.08	0.01	0.48	0.10	0.88	0.38
Estaldura	0.59	0.96	0.32	0.62	0.63	0.83	0.69	0.66
F	0.58	0.69	0.13	0.02	0.55	0.18	0.77	0.41

8 Taula: A2T *RareDis* corpusean (16-8 eszenarioa

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	3738	0	0	0	0	0	3406	
Ana	0	980	9	0	21	0	697	
In_ris	0	34	64	0	9	0	986	
Is_sy	0	0	5	26	15	4	501	
Is_a	0	33	12	2	524	1	850	
Is_ac	0	0	0	1	17	150	1290	
no_re	307	26	21	5	73	8	12270	
Doitasuna	0.52	0.57	0.06	0.05	0.37	0.10	0.97	0.38
Estaldura	0.92	0.91	0.58	0.76	0.80	0.92	0.61	0.79
F	0.67	0.71	0.11	0.09	0.50	0.19	0.75	0.43

9 Taula: A2T *RareDis* corpusean (32-16 corpusean)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	39	0	0	0	0	0	55	
Ana	7	175	4	0	1	0	221	
In_ris	0	6	2	0	1	0	4	
Is_sy	159	7	0	2	30	4	394	
Is_a	299	121	14	4	245	13	662	
Is_ac	0	2	1	1	1	8	218	
no_re	3541	762	90	27	381	138	18446	
Doitasuna	0.41	0.43	0.15	0.00	0.18	0.03	0.79	0.28
Estaldura	0.01	0.16	0.02	0.06	0.37	0.05	0.92	0.23
F	0.02	0.24	0.03	0.01	0.24	0.04	0.85	0.20

10 Taula: EM eredua *RareDis* corpusean (1-1 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	1402	1	0	0	0	0	4371	
Ana	12	566	6	0	4	0	242	
In_ris	2	1	27	0	2	0	131	
Is_sy	11	2	1	11	6	0	103	
Is_a	58	0	5	0	364	1	60	
Is_ac	0	0	0	1	0	107	65	
no_re	2560	503	72	22	283	55	15028	
Doitasuna	0.24	0.68	0.17	0.08	0.75	0.62	0.81	0.48
Estaldura	0.35	0.53	0.24	0.32	0.55	0.66	0.75	0.49
F	0.29	0.59	0.20	0.13	0.63	0.64	0.78	0.47

11 Taula: EM eredua *RareDis* corpusean (8-4 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	2194	2	0	0	2	0	6646	
Ana	1	749	6	0	3	0	326	
In_ris	14	7	46	0	4	0	293	
Is_sy	17	5	1	21	4	0	146	
Is_a	62	0	5	0	407	1	159	
Is_ac	0	0	0	3	2	128	162	
no_re	1757	310	53	10	237	34	12268	
Doitasuna	0.25	0.69	0.13	0.11	0.64	0.43	0.84	0.44
Estaldura	0.54	0.70	0.41	0.62	0.62	0.79	0.61	0.61
F	0.34	0.69	0.19	0.18	0.63	0.56	0.71	0.47

12 Taula: EM eredua *RareDis* corpusean (16-8 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	2791	1	1	0	4	0	6629	
Ana	1	856	7	0	1	0	455	
In_ris	84	10	74	1	38	1	1480	
Is_sy	15	5	1	22	2	0	81	
Is_a	26	1	2	1	364	1	46	
Is_ac	2	0	0	0	1	139	204	
no_re	1126	200	26	10	249	22	11105	
Doitasuna	0.30	0.65	0.04	0.17	0.83	0.40	0.87	0.46
Estaldura	0.69	0.80	0.67	0.65	0.55	0.85	0.56	0.68
F	0.41	0.72	0.08	0.28	0.66	0.55	0.68	0.48

13 Taula: EM eredua *RareDis* corpusean (32-16 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	750	41	4	0	37	0	1511	
Ana	413	730	12	4	74	8	2991	
In_ris	118	94	26	5	37	6	1009	
Is_sy	647	79	23	11	60	4	5004	
Is_a	2012	114	30	7	437	22	6586	
Is_ac	61	14	14	4	5	109	1454	
no_re	44	1	2	3	9	14	1445	
Doitasuna	0.32	0.17	0.02	0.00	0.05	0.07	0.95	0.23
Estaldura	0.19	0.68	0.23	0.32	0.66	0.67	0.07	0.40
F	0.23	0.28	0.04	0.00	0.09	0.12	0.13	0.13

14 Taula: EN1 eredua *RareDis* corpusean (1-1 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	3029	6	6	0	14	0	8250	
Ana	142	926	9	0	37	2	2556	
In_ris	263	121	64	0	39	4	2092	
Is_sy	65	10	5	25	41	6	1062	
Is_a	354	6	11	0	486	3	695	
Is_ac	1	0	11	3	10	144	634	
no_re	191	4	5	6	32	4	4711	
Doitasuna	0.27	0.25	0.02	0.02	0.31	0.18	0.95	0.28
Estaldura	0.75	0.86	0.58	0.74	0.74	0.88	0.24	0.68
F	0.39	0.39	0.05	0.04	0.44	0.30	0.38	0.23

15 Taula: EN1 eredua *RareDis* corpusean (8-4 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	3321	16	10	0	7	1	5100	
Ana	103	974	8	0	32	7	2824	
In_ris	182	66	70	0	52	2	2090	
Is_sy	96	13	2	27	34	9	1111	
Is_a	147	2	8	1	484	3	523	
Is_ac	3	0	10	3	17	136	659	
no_re	193	2	3	3	33	5	7693	
Doitasuna	0.39	0.25	0.03	0.02	0.41	0.16	0.97	0.32
Estaldura	0.82	0.91	0.63	0.79	0.73	0.83	0.38	0.72
F	0.53	0.39	0.05	0.04	0.53	0.27	0.55	0.34

16 Taula: EN1 eredua *RareDis* corpusean (16-8 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	3457	7	7	0	35	1	4721	
Ana	55	986	8	0	19	4	1823	
In_ris	158	66	84	2	74	3	3861	
Is_sy	71	4	3	26	41	5	570	
Is_a	59	6	5	0	450	2	277	
Is_ac	13	0	0	2	8	144	729	
no_re	232	4	4	4	32	4	8019	
Doitasuna	0.42	0.34	0.02	0.04	0.56	0.16	0.97	0.35
Estaldura	0.85	0.92	0.76	0.76	0.68	0.88	0.40	0.75
F	0.56	0.50	0.04	0.07	0.62	0.27	0.57	0.43

17 Taula: EN1 eredua *RareDis* corpusean (32-16 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	12	1	0	1	2	3	15	
Ana	11	101	2	0	17	1	110	
In_ris	0	1	1	0	4	3	25	
Is_sy	6	11	1	2	7	1	50	
Is_a	13	32	1	0	69	11	47	
Is_ac	0	0	0	0	0	1	29	
no_re	4003	927	106	31	560	143	19724	
Doitasuna	0.35	0.42	0.03	0.03	0.40	0.03	0.77	0.29
Estaldura	0.00	0.09	0.01	0.06	0.10	0.01	0.99	0.18
F	0.01	0.15	0.01	0.04	0.17	0.01	0.87	0.18

18 Taula: EN2 eredua *RareDis* corpusean (1-1 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	1625	4	2	0	2	0	1881	
Ana	74	770	8	0	15	1	986	
In_ris	40	61	28	1	18	0	623	
Is_sy	49	8	3	20	10	4	583	
Is_a	397	15	6	0	460	2	405	
Is_ac	3	0	5	4	6	124	299	
no_re	1857	215	59	9	148	32	15223	
Doitasuna	0.46	0.42	0.04	0.03	0.36	0.28	0.87	0.35
Estaldura	0.40	0.72	0.25	0.59	0.70	0.76	0.76	0.60
F	0.43	0.53	0.06	0.06	0.47	0.41	0.81	0.39

19 Taula: EN2 eredua *RareDis* corpusean (8-4 eszenarioa)

Iragarritako klasea \ Benetako klasea	Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
Prod	2422	8	2	0	3	0	2283	
Ana	43	860	8	0	12	1	875	
In_ris	79	128	51	0	15	4	1044	
Is_sy	26	4	1	21	12	2	720	
Is_a	298	3	10	0	473	3	383	
Is_ac	15	1	0	3	5	137	332	
no_re	1162	69	39	10	139	16	14363	
Doitasuna	0.51	0.48	0.04	0.03	0.40	0.28	0.91	0.38
Estaldura	0.60	0.80	0.46	0.62	0.72	0.84	0.72	0.68
F	0.55	0.60	0.07	0.05	0.52	0.42	0.80	0.43

20 Taula: EN2 eredua *RareDis* corpusean (16-8 eszenarioa)

Iragarritako klasea	Benetako klasea							
		Prod	Ana	In_ris	Is_sy	Is_a	Is_ac	no_re	
	Prod	3303	1	4	0	7	0	3589	
	Ana	144	1026	11	0	24	4	2530	
	In_ris	64	22	75	0	32	2	1935	
	Is_sy	60	9	1	23	14	2	324	
	Is_a	153	4	14	2	521	5	486	
	Is_ac	3	0	0	2	8	145	240	
	no_re	318	11	6	7	53	5	10896	
	Doitasuna	0.48	0.27	0.04	0.05	0.44	0.36	0.96	0.37
	Estaldura	0.82	0.96	0.68	0.68	0.79	0.89	0.54	0.77
	F	0.60	0.43	0.07	0.10	0.57	0.52	0.70	0.43

21 Taula: EN2 eredua *RareDis* corpusean (32-16 eszenarioa)

Bibliografia

[1] Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379, 2019. Ikusi 1, 2 orrialdeak.

[2] Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319, 2021. Ikusi 1 orrialdea.

[3] RPJC Ramanayake and BMTK Basnayake. Evaluation of red flags minimizes missing serious diseases in primary care. *Journal of family medicine and primary care*, 7(2):315, 2018. Ikusi 1 orrialdea.

[4] Stephen B Johnson, Suzanne Bakken, Daniel Dine, Sookyung Hyun, Eneida Mendonça, Frances Morrison, Tiffani Bright, Tielman Van Vleck, Jesse Wrenn, and Peter Stetson. An electronic health record based on structured narrative. *Journal of the American Medical Informatics Association*, 15(1):54–64, 2008. Ikusi 2 orrialdea.

[5] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29, 2017. Ikusi 2 orrialdea.

[6] Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2, 2016. Ikusi 11, 31 orrialdeak.

[7] Elizabeth D Liddy. Natural language processing. *Encyclopedia of Library and Information Science*, 2001. Ikusi 15 orrialdea.

[8] Ben Lutkevich and Ed Burns. What is natural language processing? an introduction to nlp, Jan 2023. Ikusi 15 orrialdea.

[9] Gregory Grefenstette. Tokenization. *Syntactic Wordclass Tagging*, pages 117–133, 1999. Ikusi 15 orrialdea.

[10] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86, 2004. Ikusi 16 orrialdea.

[11] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. Ikusi 16 orrialdea.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. Ikusi 18 orrialdea.

[13] Hugging face – the ai community building the future. Ikusi 18, 42 orrialdeak.

[14] Shantanu Kumar. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*, 2017. Ikusi 19, 22 orrialdeak.

[15] Rahul Sharnagat. Named entity recognition: A literature survey. *Center For Indian Language Technology*, pages 1–27, 2014. Ikusi 19 orrialdea.

[16] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020. Ikusi 19, 21, and 22 orrialdeak.

[17] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003. Ikusi 21 orrialdea.

[18] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon, 2004. Ikusi 22 orrialdea.

[19] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*, 2019. Ikusi 23 orrialdea.

[20] Baeldung baeldung. F-1 score for multi-class classification, Nov 2022. Ikusi 25 orrialdea.

[21] Bill MacCartney. *Natural language inference*. Stanford University, 2009. Ikusi 26, 27 orrialdeak.

[22] Poulomi Chatterjee. How do zero-shot, one-shot and few-shot learning differ?, Mar 2022. Ikusi 28 orrialdea.

[23] Eram Munawwar. Zero and few shot learning, Jan 2021. Ikusi 28 orrialdea.

[24] Rouse Margaret. Zero-shot, one-shot, few-shot learning, Mar 2023. Ikusi 29 orrialdea.

[25] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. Ikusi 31 orrialdea.

[26] American College of Cardiology. History of present illness. Ikusi 31 orrialdea.

[27] Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. The raredis corpus: a corpus annotated with rare diseases, their signs and symptoms. *Journal of Biomedical Informatics*, 125:103961, 2022. Ikusi 34 orrialdea.

[28] Apache ctakes™ - clinical text analysis knowledge extraction system. Ikusi 39 orrialdea.

[29] Unified medical language system (umls). Ikusi 40 orrialdea.

[30] Alan Aronson. Metamap. Ikusi 40 orrialdea.

[31] Anthony Rios. Pymetamap: Python wraper for metamap. Ikusi 40 orrialdea.

[32] Angel X. Chang and Christopher Manning. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). Ikusi 40 orrialdea.

[33] Irene Li, Keen You, Xiangru Tang, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, and Dragomir Radev. Ehrkit: A python natural language processing toolkit for electronic health record texts. *arXiv preprint arXiv:2204.06604*, 2022. Ikusi 41 orrialdea.

[34] Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. Ikusi 43 orrialdea.

[35] Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, July 2022. Association for Computational Linguistics. Ikusi 43 orrialdea.

[36] Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212,

Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. Ikusi 43 orrialdea.

[37] Oscar Sainz and German Rigau. Ask2Transformers: Zero-shot domain labelling with pre-trained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52, University of South Africa (UNISA), January 2021. Global Wordnet Association. Ikusi 43 orrialdea.

[38] Gugger Sylvain. run_glue.py. https://github.com/huggingface/transformers/blob/main/examples/pytorch/text-classification/run_glue.py, 2023. Ikusi 48 orrialdea.