Hizkuntza Anitzeko Erlazio Semantikoen Erauzketa Medikuntzaren Domeinuan

Egilea: Oscar Sainz
Tutoreak: Oier Lopez de Lacalle eta Gorka Labaka

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2020eko iraila

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

Laburpena

Aro digital honentan datu kopuru handiena textu gordin formatuan aurkitzen da. Datu horiekin lan egiteko Informazio Erauzketa (IE) bihurtzen da oinarri gaur egungo aplikazioetan. Hizkuntzaren prozesaketa automatikoko ataza gehientxuenetan gertatu den bezala ikasketa sakonak artearen egoera ezarri du, baita IEn ere. Jakina da teknika hauek datu kopuru handiak behar dituztela errendimendu ona lortzeko. Badira hainbat domeinu eta testuinguru, datu anotatu gutxikoak, zailtasunak dituztenak ikasketa sakoneko tekniken aurrerapenak modu eraginkorrean erabiltzeko. Anotazio berriak egitea garestia izaten da orokorrean, batez ere eredu berri hauek behar duten kopuruetara iristeko. Lan honen helburu nagusia domeinu eta testuinguru hauentzako modu merke batean ikasketa sakoneko sistemen errendimendua hobetzeko teknikak esploratzea da. Zehatzago esanda, ezagutza-transferentzia eta datuen-gehikuntza automatikoa paradigmetan ikertuko dugu helburua lortzeko. Azkenik, teknika hauek baliabide urrikoa den medikuntzako domeinuko eHealth-KD 2020 ataza-partekatuan aplikatuko eta ebalutako dira, uneko artearen egoera hobetzeko helburuarekin.

Abstract

In this digital age the greatest amount of data is found in raw text format. Information Extraction (IE) to work with this data becomes the basis in today's applications. As has happened in most tasks of automatic language processing, deep learning has established the state of the art in IE as well. It is well known that these techniques require a large amount of data to achieve good performance. There are a number of domains and contexts, with little annotated data, that have difficulties making effective the use of advances in deep learning techniques. Making new annotations is generally expensive, especially to reach the numbers needed for these new models. The main goal of this work is to explore techniques to improve the performance of deep learning systems in a cost-effective way for these domains and contexts. More specifically, we will investigate transfer-learning and automatic data augmentation paradigms to achieve the goal. Finally, these techniques will be applied and evaluated in the shared task eHealth-KD 2020 in the low-resource medical domain, with the goal of improving the state of the art.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1 Sarrera

Informazioaren Erauzketa (IE) adimen artifizialaren zereginetako bat da, eta irakurmen gaitasuna eman dio makinari. Gaur egun, makinak gai dira testua irakurtzeko edo ahozko hizkuntza entzuteko, prozesatzeko eta ulertzeko, informazio-elementu garrantzitsuenak lortuz. IE aplikazio industrial berrien atzean dago, hala nola Google Knowledge Graph (GKG), Googlek bere bilaketa-motorraren emaitzak hobetzeko erabiltzen duen teknologia. IE-k hizkuntza teknologiako hainbat aplikazioetan eragin zuzena du, adibidez, ezagutza baseen osatzea (*knowledge base poppulation* ingelesez) (Ji eta Grishman, 2011) eta galde-erantzun sistemetan (*question answering* ingelesez) (Yu et al., 2017), besteak beste. Berriki, informazio-erauzketako sistemak COVID-19ak eragindako pandemiaren kontra lagungarri direla erakutsi da[1]. Azken hilabetetan sortu diren milaka artikuluetatik ikertzaileri lagungarri zaion informazioa modu egituratu eta zehatzean aurkezteko erabili dira.

Testu gordinetik abiatuta Wikipediako *infobox* baten sorkuntza (edo betetze) automatikoaren ataza IEren adibide argi bat da, hau da, testu gordinetik informazioa erauztea eta errepresentazio egituratu batean aurkeztea. Orokorrean, IEk **entitateen**, **erlazioen** eta **gertaeren** erauzketa barneratzen ditu. Gertaeren kasuan IEk rol semantikoaren etiketatzearekin (*Semantic Rol Labeling* ingelesez) antza handia dauka non gertaerak predikatuak bihurtzen diren eta bete beharreko hutsuneak (*slots* ingelesez) argumentuak. Baina, IEk helburu desberdina du, esaldi bateko informazioa guztiz ondo etiketatu beharrean (mikro irakurketa), dokumentu edo dokumentu multzo bateko informazioa erauztea (makro irakurketa) da. Hori dela eta IE ataza galde-erantzun atazekin parekatzen da kasu askotan. Esan bezala, erlazio-erauzketa IE ataza orokorraren barne dagoen ataza bat da eta bi entitate eta testuinguru bat emanda bi entitateek erakusten duten erlazioa, baldin badago, erauzketari deritzo. Adibidez, "*El asma es una enfermedad que afecta alas vías respiratorias.*"esaldian (`is-a`, *asma*, *enfermedad*) erlazioa ikus dezakegu.

Lengoaia Naturalaren Prozesamenduaren (LNP) beste alorretan bezalaxe informazio-erauzketako atazetan paradigma aldaketa sakona gertatu da azken urteetan. Eskuz sortutako erregelatan oinarritzen ziren sistemak (Chiticariu et al., 2013) ezaugarrietan oinarritutako ikasketa automatikoko ereduez (Lafferty et al., 2001) ordezkatuak izan ziren bezalaxe, gaur egun ikasketa sakoneko teknikak (*deep learning* ingelesez) baliatzen dituzten ereduak dira erabilienak. Horiek dira LNP eta IEri lotutako hainbat atazetan artearen egoera definitzen duten hurbilpenak (Akbik et al., 2019a; Devlin et al., 2019). Paradigma berriak aurrerapen handiak eskaini dituen arren, badira hainbat domeinu eta testuinguru, datu anotatu gutxikoak, zailtasunak dituztenak ikasketa sakoneko tekniken aurrerapenak modu eraginkorrean erabiltzeko. Jakina da, adibidez, testu klinikoen LNPn ikasketa sakoneko metodoek domeinuaren berariazko erronkak arindu baditzakete (domeinuko tresna eta datu-multzoen falta), erronka horiek ez dutela arreta handirik jaso ikerlarien artean (Weegar et al., 2020). Motzean esanda, lan honen helburua datu anotatu gutxiko domeinutan, zehatzago bio-medikuntzako domeinuetan, erlazio-erauzketa modu eraginkorrean egitea da.

[1] `www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge`

Datuen gehikuntza automatikoko teknikak oso erabiliak izan dira ikasketa sakoneko algoritmoekin batera. Teknika horiek batez ere konputagailu bidezko irudien prozesamenduan izan dute arrakasta. Irudien gaineko transformazioen bitartez (Shorten eta Khoshgoftaar, 2019; Caron et al., 2020) entrenamendurako datu-multzoa handitzea ikasketa sakoneko ereduak entrenatzeko ohiko praktika bat da sistema trinko bat lortu nahi denean eta datu gutxi daudenean. LNPn berriz ez da ataza erreza, hain zuzen ere esaldi edo paragrafo batetik baliokide izan daitezkeen beste antzeko adibideak sortzeko Hizkuntza Naturalaren Ulermenaren (HNU) gaitasuna beharrezkoa delako. Hori dela eta, erlazio-erauzketan behintzat, existitzen diren datu-gehikuntza automatikoko teknikek adibide zaratatsuak sortzen dituzte eta ez dira beti eraginkorrak izaten. Urruneko-gainbegiraketa (Mintz et al., 2009; Sainz et al., 2020) edo *Matching the Blanks* (Baldini Soares et al., 2019) dira adibide aipagarrienak erlazio-erauzketari dagokionez. Zarata hori automatikoki gutxitzeko estrategiak (Intxaurrondo et al., 2013) garatu diren arren oraindik hobetzeko aldea badago.

Bestetik, ezagutza-transferentziako teknikak (Devlin et al., 2019) oso arrakastatsuak izan dira LNPko hainbat atazetan. Bereziki ere datu anotatu gutxiko domeinu berezituetan erakutsi dute eraginkortasuna. Lan honen beste helburuetako bat transferentzia eta datuen gehikuntza tekniken azterketa egitea da, domeinu konkretuan erlazio-erauzketako emaitzak hobetzeko asmoz. Zehatzago esanda, MTB datu-gehikuntzarako teknika zerotik berrinplementatuko da eta MLM ezagutza-transferentziako teknika aplikatuko da existitzen diren inplementazioekin.

Egindako esperimentazioa hiru fase nagusietan banatzen da: domeinu orokorreko erlazio-erauzketarako artearen egoerako sistemaren inplementazioa, garatutako sistema domeinu zehatz batera egokitzea eta datu-gehikuntza eta ezagutza-transferentzien tekniken bitartez sistema hobetzen saiatzea. Lehenengo fasean hainbat sistema aztertu dira, beraien artean AGGCN (Guo et al., 2019a), TRE (Alt et al., 2019) eta BERT_{EM} (Baldini Soares et al., 2019), eta horietatik BERT_{EM}ren berrinplementazio bat egin da. Azterketaren ondorioz ikusi da Transformer eta ezagutza-transferentzian oinarritutako sistemek emaitza hobeagoak lortzen dituztela sistema konplexuagoek baino. Bigarren faserako, garatu dugun BERT_{EM} sistema gaztelaniazko medikuntza domeinura egokitu dugu eHealth-KD 2020[2] ataza-partekatuan ebaluatu ahal izateko. Garatu ditugun sistemaren hiru bertsioek beste partaideen sistemak gainditu dituzte erlazio-erauzketara bideratua dagoen atazean. Beraz ondoriozta dezakegu bai BERT_{EM}ek baita XLM_{EM}ek ere jarraitzen duten entitate-marken estrategia egokia dela erlazioen errepresentazio on bat lortzeko. Bestetik, MTB datu-gehikuntzarako estrategiak sistemaren errendimendu orokorra hobetu ez arren doitasun hobea lortzeko balio izan digu estaldura pixkat galduz. Azkeneko fasean dagoeneko geneuzkan emaitzak hobetu ditugu gure sistemaren elementu garrantzitsuena hobetuz, hau da, sistemaren hizkuntza-eredua hobetuz. Horretarako aurrentrenatuta dauden hizkuntza-eredu eleanitzen artean esplorazio sakonago bat garatu dugu baita medikuntza domeinuko testuen bitartez gure MLM hizkuntza-eredua entrenatu ere. Lortutako emaitzekin hiru ondorio nagusi lortu ditugu: batetik hizkuntza-eredu batek entrenamendu orduan ikusita-

[2] `www.knowledge-learning.github.io/ehealthkd-2020/`

ko testu kopurua garrantzitsuagoa dela hizkuntza-ereduaren tamaina bera baino, bestetik, hizkuntza-eredu handiek orokortasun ahalmen handiagoa erakusten dutela entrenamendu, garapen eta testeko emaitzen artean desberdintasuna gutxituz. Azkenik eHealth-KD 2020ko testuek medikuntzaren domeinukoak izan arren ez dute hizkuntza teknikoa erabiltzen, eta beraz, mota horietako testuak erabiltzea hizkuntza-eredua birdoitzeko fabore baino kalte gehiago egiten dute. Hain zuzen ere, MTB bezala MLM birdoiketak sistemaren doitasuna igo du, seguraski medikuntzako terminologia hobeto dagoelako errepresentatuta, baina estaldura jaitsi du, seguruenik eHealth-KDen agertzen diren terminologia arrunta okerrago dagoelako errepresentatuta.

Dokumentu honek lan honetan egindako garapena, esperimentuak eta emaitzak jasotzen ditu. Horrela dago antolatua: lehendabizi sarrera bat aurkezten da lan honen helburuekin batera. Gero, erlazio-erauzketa eta ezagutza-transferentziari buruzko artearen egoera deskribatzen da 2. atalean. Hurrengo atalak ingurune esperimentala azaltzen du, hau da, ataza, erabilitako datu-multzoak eta ebaluazio-neurriak. 4. atalean domeinu orokorreko erlazio-erauzketako artearen egoerako sistemak deskribatu eta aztertu egiten dira, baita egindako berrinplementazioaren azalpenak eman ere. Ondoren, medikuntza domeinura egokitzeko pausuak azaltzen dira 5. atalean egindako esperimentuakin batera. Azkenik 6. atalean sistemaren emaitzak azaltzen dira eta 7. atalean berriz ondorioak eta etorkizuneko lana.

1.1 Helburuak

Laburbilduz, proiektu honek hiru helburu nagusi izango ditu: alde batetik artearen egoerako erlazio-erauzketako sistema baten inplementazioa; bestetik, garatutako sistema domeinu zehatz batera egokitzea; eta azkenik, sistema hobetzen saiatzea datu-gehikuntza eta ezagutza-transferentzia tekniken bitartez. Hiru helburu hauekin lortu nahi duguna domeinu zehatzeko artearen egoera hobetzea da.

Artearen egoerako erlazio-erauzketako sistema baten inplementazioa. Lehendabiziko helburua erlazio-erauzketako sistema baten inplementazioa da. Garatutako sistema artearen egoera definitzen duten sistemen parean egon beharko du. Horretarako, TACRED (Zhang et al., 2017) erlazio-erauzketarako datu-multzo estandarra erabiliko dugu ebaluaketa egiteko. Gure sistema garatu aurretik dagoeneko existitzen diren sistemen azterketa sakon bat egingo da, eta behar izanez gero baten bat berrinplementatu ere.

Garatutako sistema domeinu zehatz batera egokitzea. Proiektu honen helburu garrantzitsuena datu gutxiko domeinuen gaineko erlazio-erauzketa hobetzea da. Helburuaren erdiespena neurtzeko eHealth-KD 2020 (Piad-Morffis et al., 2020) ataza partekatuan parte hartu dugu. Ataza partekatuak gaztelaniazko medikuntza domeinuko testuen gaineko erlazio-erauzketa planteatzen du, zeinetan datu-multzo oso txikia dugun ereduak entrenatzeko.

HAP masterra

Sistema hobetzea datu-gehikuntza eta ezagutza-transferentzien tekniken bitartez. Azkenik, sistema hobetzeko helburuarekin datu-gehikuntzako *Matching the Blanks* eta ezagutza-transferentziako *Masking Language Modeling* teknikak aztertu eta inplementatu ditugu. Teknika horiek aplikatzeko beharrezkoak diren corpus eta datu-multzoak erauzi eta prozesatuko dira LNPko hainbat tresnak erabiliz.

2 Artearen egoera

Lan honetan ezagutza-transferentzia eta datu-gehikuntza teknikak erlazio-erauzketa atazan eduki duten eragina aztertu nahi dugu, batez ere baliabide urriko inguruneetan lor ditzaketen hobekuntzak. Hori dela eta, atal honetan LNP ezagutza-transferentziaren eta erlazio-erauzketaren artearen egoera errepasatuko dugu.

2.1 Erlazio-erauzketa

Erlazio-erauzketa ez da ataza berri bat (Chinchor, 1998) baina beste atazetan bezala ikasketa sakonak bultzada handi bat eman du, hori dela eta azken urteetako lanetan zentratuko gara. Hasiera batean neurona sare konboluzional (CNN, *Convolutional Neural Networks* ingelesez) (Zeng et al., 2015) eta LSTM (*Long Short Term Memory* ingelesezko siglen arabera) (Zhang et al., 2017) neurona sareak erabili izan dira esaldiko erlazio errepresentazioa lortzeko. Geroago, atentzio-mekanismoek (Yu et al., 2019) emaitzak hobetzea lortu zuten. Kanpoko ezagutza, adibidez informazio sintaktikoa, barneratzeko saiakeran agertu ziren grafoen gaineko neurona sare konboluzionalak (GCN, *Graph Convolutional Network* ingelesez) (Kipf eta Welling, 2017) baita atentzio bidez zuzendutako grafoen gaineko neurona sare konboluzionalak (AGGCN, *Attention Guided Graph Convolutional Network* ingelesez) (Guo et al., 2019a) ere. Berriki ezagutza-transferentziak (*transfer-learning* ingelesez) erakutsi du aukera arrakastatsua izatea (ia) etiketaturiko daturik ez daukaten domeinu eta hizkuntzetan (Devlin et al., 2019; Baldini Soares et al., 2019). Transformerretan oinarritutako sekuentzia ereduek (Vaswani et al., 2017) erlazio-erauzketa bezalako informazio-erauzketa ataza askotan artearen egoera hobetu dute (Baldini Soares et al., 2019; Peters et al., 2019; Joshi et al., 2019). Badaude lan batzuk kanpoko ezagutza, adibidez ezagutza-baseetatik aterata, Transformers ereduetan barneratzen saiatu direnak (Peters et al., 2019). Hala eta guztiz ere, entitate-marketan oinarritutako hurbilpen sinpleagoak (ikusi detaile gehiago 4. atalean) antzeko errendimendu konpetitiboa erakusten dute inplementazio azkarrago batekin Baldini Soares et al. (2019). Modu berean, hizkuntza-eredu eleanitzek (Lample eta Conneau, 2019; Conneau et al., 2019) hizkuntza batean ikasitakoa beste hizkuntzetara hedatzeko gaitasuna erakutsi dute. Helburuko hizkuntzan anotaturiko datu oso gutxi dauzkaten erlazio-erauzketako atazentzat etorkizun handiko eredu mota dirudite hizkuntza-eredu eleanitz horiek.

Erlazio-erauzketara zuzendutako datu-gehikuntza Informazio-erauzketan hainbat teknika proposatu dira datuak automatikoki sortzeko. Horien artean esanguratsuenak urruneko gainbegiratzean (*distant-supervision* ingelesez) (Mintz et al., 2009; Sainz et al., 2020) oinarritzen direnak dira, non ezagutza-baseetako erlazioak anotazio gabeko testuarekin lerrokatzen dira zenbait heuristikotan oinarrituta automatikoki entrenamendurako datu (zaratatsuak) bilduz. Berrikiago, Baldini Soares et al. (2019) erlazio-etiketarik behar ez duen datu-gehikuntza proposatzen du eredua domeinura doitzeko. Bestalde, itzulpen automatikoan oinarritutako lanek iradokitu dute entrenamendu datu-multzoa hainbat hizkuntzetara itzuliz onuragarria izan daitekeela LNPko hainbat atazetan emaitzak hobetzeko

(Artetxe et al., 2020). Oraindik ere informazio-erauzketako atazetan ikusteko dago itzulpengintzan oinarritutako teknikak onuragarriak diren.

2.2 Ezagutza-transferentzia

Duela urte batzuk LNParen ikasketa sakoneko teknikek jasan zuten lehendabiziko bultzada gertatu zen hitz-bektoreen (Mikolov et al., 2013) agerpenarekin. Hitz-bektore horiek izan ziren ezagutza-transferentziaren bitartez LNParen alorreko ia ataza guztietan artearen egoera hobetu zuten baliabidea. Geroztik, hitz-bektoreen errepresentazioak ikasteko teknika gehiago (Pennington et al., 2014; Grave et al., 2018) agertu dira, baita karaktereetan oinarrituta (Akbik et al., 2018) ere, azken hauek gaur egun ere artearen egoerako emaitzak lortzen dituzte izendun entitateen erauzketa (*Named Entity Recognition, NER* ingelesez) atazetan (Akbik et al., 2019b; López-Ubeda et al., 2020). Bektore horiek oso onuragarriak izan arren ez dira gai beraiek baitan testuinguruko informazioa jasotzeko, horren ondorioz laister agertu ziren lehendabiziko hizkuntza-ereduen aurrentrenatuen ideia (Howard eta Ruder, 2018; Peters et al., 2018).

LNParen ikasketa sakoneko teknikek jasan duten bigarren bultzada esanguratsua Transformerrekin (Vaswani et al., 2017) eta hauetan oinarritutako hizkuntza-eredu aurrentrenatuekin (Radford et al., 2018; Devlin et al., 2019; Conneau et al., 2019) gertatu zen. Hizkuntza-eredu aurrentrenatuak baino lehen ikasketa sakoneko ereduak zerotik entrenatzea ohikoena zen, gaur egun berriz ia ataza guztietan hizkuntza-eredu aurrentrenatu hauek atazara birdoitzea da jarraitzen den estrategia. Hain zuzen ere, anotatutako datu gutxiko egoeratan aurrentrenamenduan ikasitako ezagutza atazara transferitzeak emaitzak eta ereduaren orokortze ahalmena hobetzen ditu (Hendrycks et al., 2020). Oraindik ere estrategia hau jarraitzen ez dituzten atazak badaude, bereziki datu asko dauzkatenak, adibidez itzulpen neuronal automatikoa baina soilik datu kopuru handiko hizkuntzen artean (Tiedemann eta Thottingal, 2020).

Teorikoki oraindik oso ondo ulertu ez arren badaude hizkuntza-eredu hauek zer ikasten duten aztertzen saiatu diren lanak (Tenney et al., 2019; Clark et al., 2019). Enpirikoki ere ez dago argi hizkuntza-eredu hauek jaso dezaketen ezagutzaren muga, honen adibide *few-shot* edota *zero-shot* atazak ebazteko abilezia erakutsi duten GPT-2 (Radford et al., 2019) eta berrikiago GPT-3 (Brown et al., 2020) hizkuntza-ereduak dira.

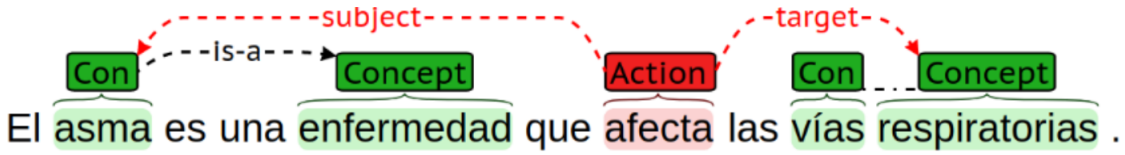
Hizkuntza-ereduen ezagutza muga non dagoen aztertzeko helburuarekin eredu hauen entrenamendurako testu kopurua eta ereduaren parametro kopurua esponentzialki hazi dira 600×10^9 parametroetara (Lepikhin et al., 2020) helduz. Hala ere, ezagutza-transferentziaren ondorioz hizkuntzen arteko (Lample eta Conneau, 2019; Conneau et al., 2019) edo atazen arteko (Yin et al., 2019) *zero-shot* edo *few-shot* egiteko abilezia ez da hizkuntza-eredu handietara mugatzen.

3 Ingurune esperimentala

Atal honetan lan honen ingurune esperimentala definituko dugu. Lehendabizi erlazio-erauzketa atazara sarrera motz bat egingo dugu. Gero, lan honetan zehar erabili ditugun datu-multzoak deskribatuko ditugu eta azkenik inplementatutako sistemak ebaluatzeko erabili ditugun neurriak azalduko ditugu.

3.1 Erlazio-erauzketa ataza

Bi entitate eta testuinguru bat emanda, testuinguru horretan bi entitateak lotzen dituen erlazioa-mota iragartzean datza erlazio-erauzketa. Iragarri beharreko erlazioak mota desberdinetakoak izan daitezke. Mota horiek sintaxiarekin zerikusi handia duten erlazio semantikoetatik hasi eta ezagutza base batean agertzen diren kontzeptuen arteko erlazioetara zabaltzen dira. 1. irudian ikus daiteke adibide bat bi motatako erlazio semantikoak erakusten dituena. Sintaxiarekin edo RSErekin (rol semantikoaren etiketatzea) zerikusi handiagoa duten erlazioak, `subject` eta `target` kasuak batetik, eta kontzeptuen arteko erlazio ontologikoak, `is-a` kasua bestetik.

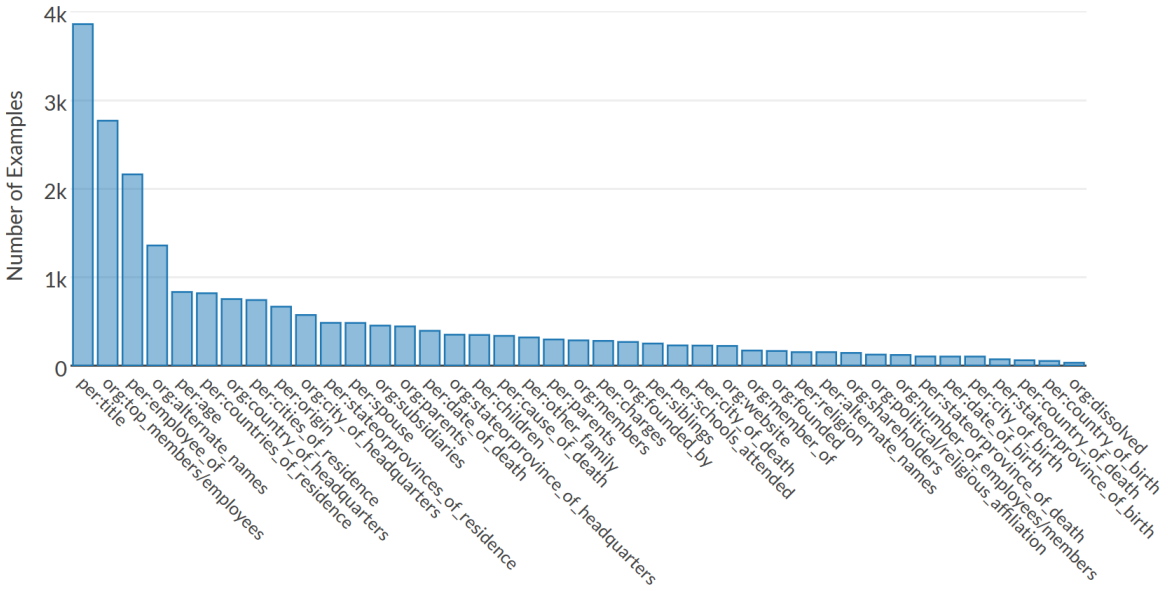


Irudia 1: Mota desberdinetako erlazio semantikoak jasotzen dituen adibide bat. eHealth-KD 2020 ataza-partekatuaren webgunetik aterata.

Sarrera atalean aipatu dugun bezala informazio-erauzketa orokorrean makro ikuspegi bat dauka. Hau da, dokumentu multzo batean informazio zehatz (entitate, erlazio edo gertaera) bat behin baino gehiagotan agertu arren, behin bakarrik erauztearekin balio du. Baina, helburua hori izan arren gaur egungo ataza askok mikro ikuspegi baten bitartez ebaluatzen dira, hau da, esaldi edo dokumentu batetik anotatuta dagoen zenbateko informazioa erauzten den, errepikatuak barne. Bestetik, erlazioak esaldi berdineko edo esaldi desberdinetako (*inter-sentence* edo *intra-sentence* ingelesez) bi entitateen artean izan daitezke. Lan honetan erabiliko ditugun datu-multzoak **esaldi mailako** erlazio-erauzketara, hau da, esaldi berdinean dauden entitateen arteko erlazio-erauzketara daude bideratuta eta **mikro** ikuspegi baten bitartez ebaluatzen dira.

3.2 Datu-multzoak

Lan honetan bi ataza garrantzitsu burutu dira, alde batetik TACRED (Zhang et al., 2017) web eta berrien gaineko erlazio-erauzketarako datu-multzoa, eta, bestetik, eHealth-KD 2020 (Piad-Morffis et al., 2020) medikuntza alorreko entitateen identifikaziorako eta erlazio-erauzketarako datu-multzoa. Azken honek bi azpi-atazaz osatuta egon arren, erlazio-erauzketan soilik zentratu gara.



Irudia 2: TACRED datu-multzoko erlazio-distribuzioa. Irudia TACRED datu-multzoaren webgune ofizialetik aterata.

3.2.1 TACRED

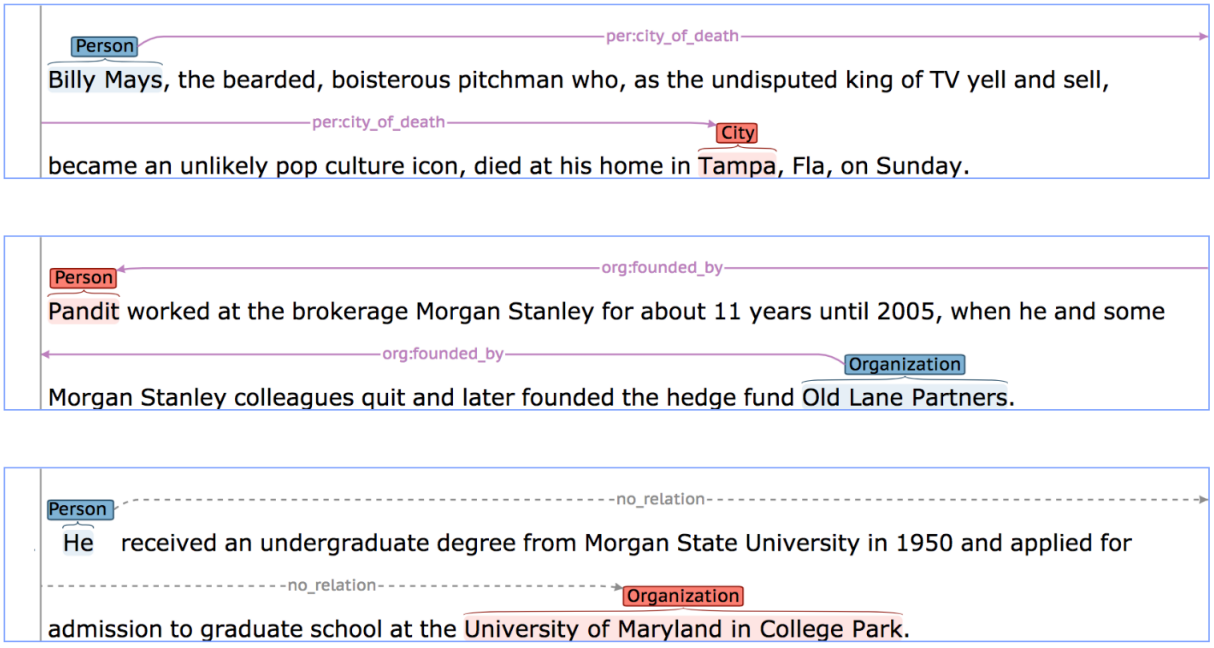
TACRED(Zhang et al., 2017) erlazio-erauzketarako tamaina handiko ingelesezko datu-multzo bat da. Corpusa osatzen duten 106.264 adibideek urtero TAC ezagutza-base osaketa (TAC KBP) (Ji et al., 2015b; Ji eta Nothman, 2016; Ji et al., 2017) erronkan erabiltzen diren corpusetatik daude aterata. Atazak, 41 erlazio-mota definitzen ditu (ikusi 2. irudia) TAC KBP erronkek bezala, baita *no_relation* erlazioa ere, azken honek bi entitateen artean erlaziorik ez dagoela adierazteko erabiltzen da. Hasiera bateko eta gaur egungo TAC KBP erronkek erlazio-multzo desberdina erabiltzen dutenez, TACRED datu-multzoko corpus zaharrek erlazio berrien anotazioekin aberastuak izan dira.

TACRED datu-multzoa bere tamaina eta domeinuarengatik erlazio-erauzketarako datu-multzo estandar bezala dago kontsideratuta. Hori dela eta, artearen egoerako erlazio-erauzketarako sistemak datu-multzo honetan probatu ohi dira. Lan honetan, inplementatutako sistema artearen egoerarekin konparatzeko erabiliko dugu datu-multzo hau.

Datu-multzoa hiru azpimultzotan dago banatuta: entrenamendu multzoa (TAC KBP 2009–2012 (Li et al., 2011; Ellis et al., 2012) corpuseko 68,124 adibide), garapen multzoa (TAC KBP 2013 (Ellis et al., 2013) corpuseko 22,631 adibide) eta testeko multzoa (TAC KBP 2014 (Ji et al., 2015a) corpuseko 15,509 adibide). Aipatutako adibideek hurrengo informazioarekin daude etiketatuta: erlazioko bi entitateen esaldiko kokalekua, entitateen mota[3] eta erlazio-mota.

3. irudian TACRED datu-multzoko hiru adibide erakusten dira. Lehendabizikoa per-

[3]Stanford NER sistemak erabiltzen dituen etiketa finak.



Irudia 3: TACRED datu-multzoko erlazioen hiru adibide. Irudia TACRED datu-multzoaren webgune ofizialetik aterata.

tsona eta hiri baten arteko `pertsona:hiltze_hiria` erlazioa erakusten du. Bigarrenak pertsona eta erakunde baten arteko `erakunde:erakunde_sortzaile` erlazioa erakusten du. Azkenik pertsona bat erreferentziatzen duen erakusle eta erakunde baten artean erlaziorik ez dagoen adibide bat erakusten du hirugarrenak.

3.2.2 eHealth-KD 2020

Medikuntza domienuko testuen gaineko informazio-erauzketa sustatzen du eHealthKD ataza partekatuak. Hirugarren urtea da txapelketa ospatzen dela eta, aurreko edizioetan bezala, izendun entitateen erauzketa eta erlazio-erauzketa dira ataza partekatu honen helburu. Ataza burutzeko Medline[4] webgunetik lortutako medikuntzari buruzko gaztelerazko testuak erabili dira. Testuak medikuntzaren domeinukoak izan arren, erauzi beharreko entitate eta erlazio motak nahiko orokorrak dira.

Entitateen artean hurrengo sailkapena bereizten da:

- **Kontzeptu (Concept):** Termino, kontzeptu edo ideia garrantzitsua esaldiko ezagutza domeinuan. Adibidez: *asma*, *enfermedad*, *vias respiratorias*.
- **Ekintza (Action):** Beste entitateen eraldaketa edo prozesua identifikatzen du. Adibidez: *afecta*, *exposicion*, *(provoca) daños*.

[4]MedlinePlus (Internet). Bethesda (MD): National Library of Medicine (US). Available from: `https://medlineplus.gov/`.

- **Predikatua (Predicate):** Testuko beste elementu batzuen iragazketa edo funtzioa adierazten du. Adibidez: *mayores* predikatua *personas* entitatea iragazten du *60 años* argumentua kontutan hartuz.

- **Erreferentzia (Reference):** Beste entitate bati erreferentzia egiten dion elementua. Adibidez: *esta*.

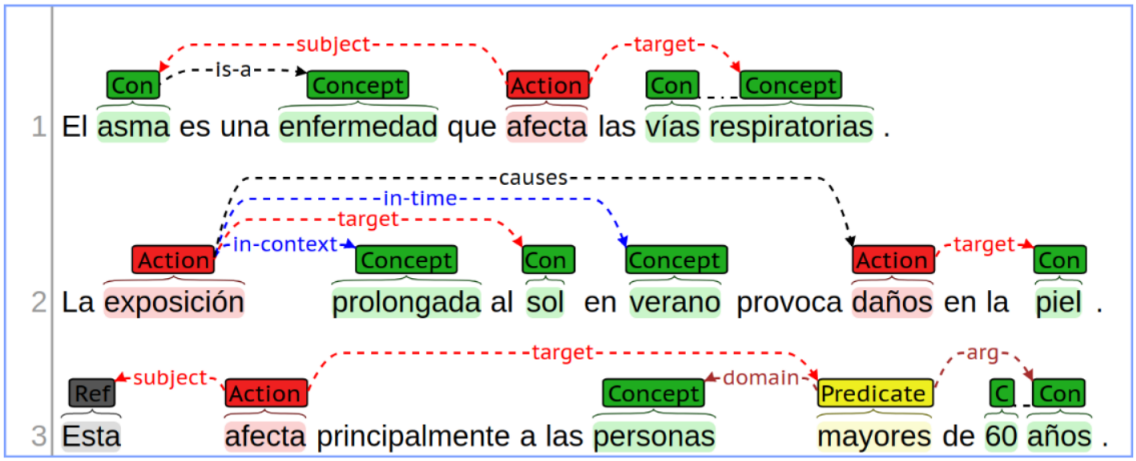
4. irudian ikus daitezke aipatutako entitate moten adibideak testuinguruan, baita beraien arteko erlazioak ere. Zehaztu beharreko hamahiru erlazioak lau multzo nagusitan banatzen dira:

- **Erlazio orokorrak (General relations):** Informazio orokorra jasotzen duten erlazio semantikoak. Erlazio hauetan edozein motatako entitateak hartu dezakete parte.

  - **is-a:** Erlazio honek bi entitateen arteko hiperonimia erlazioa adierazten du. Adibidez: (*asma*, `is-a`, *enfermedad*).
  - **same-as:** Bi entitateak semantikoki berdinak direla adierazten du. Adibidez: "... entre las vértebras o huesos de la columna."(*vértebras*, `same-as`, *huesos de la columna*).
  - **has-property:** Bigarren entitatea lehenengoaren ezaugarri bat dela adierazten du. Adibidez: "La histoplasmosis es a menudo leve y sin síntomas."(*histoplasmosis*, `has-property`, *leve*).
  - **part-of:** Erlazio honek bi entitateen arteko meronimia erlazioa adierazten du. Adibidez: "La fenilalanina se encuentra en casi todos los alimentos."(*fenilalanina*, `part-of`, *todos*).
  - **causes:** Lehendabiziko entitatea bigarrenaren agerpena edo existentzia eragiten duela adierazten du erlazio honek. Adibidez: (*exposición*, `causes`, *daños*).
  - **entails:** Lehendabiziko entitatetik bigarrenaren agerpena edo existentzia ondoriozta daitekeela adierazten du erlazio honek. Adibidez: "Los médicos usan pruebas que examinan el ano para diagnosticarlo."(*examinan*, `entails`, *diagnosticarlo*).

- **Testuinguru erlazioak (Contextual relations):** Erlazio hauek esaldiko entitateen testuingurua adierazteko erabiltzen dira.

  - **in-time:** Esaldiko elementu bat denbora-tarte jakin batean gertatzen, agertzen edo existitzen dela adierazteko erabiltzen da. Adibidez: (*exposición*, `in-time`, *verano*).
  - **in-place:** Esaldiko elementu bat leku edo posizio jakin batean gertatzen, agertzen edo existitzen dela adierazteko erabiltzen da. Adibidez: "Se han producido brotes del chikungunya en África"(*brotes*, `in-place`, *Africa*).



Irudia 4: eHealth-KD 2020 datu-multzoko erlazioen hiru adibide. Irudia eHealth-KD 2020 datu-multzoaren webgune ofizialetik aterata.

  - **in-context:** Esaldi bateko elementu baten modu edo egoera jakin batean agertzen dela adierazteko erabiltzen da. Adibidez: (*exposición*, `in-context`, *prolongada*).

- **Ekintza-rolak (Action roles):** Ekintza batekiko elementuen rola adierazteko erabiltzen dira.
  - **subject:** Ekintza eragiten duen elementua. Adibidez: (*asma*, `subject`, *afecta*).
  - **target:** Ekintza jasaten duen elementua. Adibidez: (*afecta*, `target`, *vias respiratorias*).

- **Predikatu-rolak (Predicate roles):** Predikatu batekiko elementuen rola adierazteko erabiltzen dira.
  - **domain:** Predikatua aplikatzen zaion entitate nagusia zein den adierazten du. Adibidez: (*mayores*, `domain`, *personas*)
  - **arg:** Predikatuari buruz informazio gehigarria ematen duten entitateen eta predikatuaren arteko erlazioa adierazten du. Adibidez: (*mayores*, `arg`, *60 años*).

Erlazioen definizioari eta adibideei begira erlazio orokorren multzoa ezagutza-base batean egon daitezkeen erlazioak biltzen ditu, eta beraz testuinguruaren menpekotasun txikiagoa erakusten dute. Beste guztiak berriz, sintaxi mailako ezagutzatik gertuago daude, eta beraz, testuinguruarekiko menpeagoak dira.

Erlazio-erauzketa ataza Esan bezala, eHealth-KD 2020 bi azpiataza barneratzen ditu: entitate-erauzketa eta erlazio-erauzketa. Bi atazak batera edo jarraian betetzeko daude pentsatuta. Hori dela eta, lehendabizikoan egindako akatsak bigarrenera hedatzea ekiditea

ezinezkoa izaten da. Beraz, zaila izaten da sistemak bat bi atazetan batera ebaluatzea, horregatik ataza-partekatuak 4 azpiataza aurkezten ditu:

- **Ataza nagusia (Main Task):** Txapelketako ataza nagusia, bertan bi azpiatazak jarraian edo batera egiten dira.
- **A azpiataza (Subtask A):** Entitate-erauzketara soilik zuzendutako azpiataza.
- **B azpiataza (Subtask B):** Erlazio-erauzketara soilik zuzendutako azpiataza.
- **Domeinu aldaketa azpiataza (Alternative Domain Evaluation):** Ataza nagusian bezala, bi azpiatazak jarraian egiten dira, baina kasu honetan beste domeinu bateko testuak erabiliz. Ataza hau ezagutza-transferentzia ebaluatzeko dago pentsatuta.

Lan honetan **B azpiatazan** bakarrik jardungo dugu. eHealth-KD 2020 ataza-partekatuaren datu-multzoa 3 azpimultzotan zatitzen da: entrenamendurako azpimultzoa, garapenerako azpimultzoa eta testerako azpimultzoa. Bai entrenamenduko baita garapeneko azpimultzoak berdinak dira azpiataz guztientzat. Hurrenez hurren 800 eta 199 adibidez daude osatuta. Testeko azpimultzoa berriz 100 adibide jasotzen ditu. Azkenik, *ensemble* deituriko beste azpimultzo bat dago, azken honek beste 3000 adibide automatikoki etiketatuta jasotzen ditu, beste urteetako sistemak erabiliz. *Ensemble* azpimultzoa entrenamendurako datu-gehikuntza gisa erabiltzeko dago pentsatuta.

3.3 Ebaluaketa neurriak

Erlazio-erauzketan beste sailkapen ataza askotan bezala hiru dira erabiltzen diren ebaluaketa neurri nagusiak: doitasuna, estaldura eta F1-neurria. Sailkapen bitar baten aurrean neurri hauek kalkulatzeko klase positiboa izaten da kontutan, sailkapena bitarra ez denean berriz neurri hauen batezbesteko desberdinak erabiltzen dira.

Doitasuna Neurri honek sailkatzaileak klase positibo bezala iragarritakoen artean zenbat asmatu dituen neurtzen du.

$$Doitasuna = \frac{\#Asmatutakoak}{\#Iragarritakoak} \tag{1}$$

Estaldura Neurri honek klase positiboko adibideen artean zenbat asmatu dituen adierazten du.

$$Estaldura = \frac{\#Asmatutakoak}{\#Adibide\_positiboak} \tag{2}$$

HAP masterra

F1-Neurria Azkenik, neurri honek **doitasuna** eta **estalduraren** arteko bataz-beste harmonikoa da.

$$F1 = 2 \times \frac{(Doitasuna \times Estaldura)}{(Doitasuna + Estaldura)} \tag{3}$$

Erlazio-erauzketan orokorrean klase anitzeko sailkapen bat izaten da, non erlazio bezainbeste klase positibo dauden eta erlazio negatibo bat[5] dagoen. Kasu hauetan **mikro** batezbestekoa erabili ohi da. Batezbestekoa kalkulatzeko demagun P erlazio positiboen multzoa dela, doitasuna horrela definituta egongo litzateke:

$$Doitasuna_{mikro} = \frac{\sum_{i \in P} \#Asmatutakoak_i}{\sum_{i \in P} \#Iragarritakoak_i}$$

[5]Erlazio negatiboa bi elementuen artean erlaziorik ez dagoela adierazten du.

4 Erlazio-erauzketa domeinu orokorrean

Atal honetan garatutako sistemaren inplementazioari buruz arituko gara. Lehenik eta behin artearen egoerako sistemen azterketa bat aurkeztuko dugu. Gero, beharrezkoak izan diren berinplementazioei buruz hitz egingo dugu. Azkenik, hautatutako sistemen arteko TACRED datu-multzoaren gainean egindako konparaketa erakutsiko dugu.

4.1 Sistemen deskribapena

Azterketa garatzeko Erlazio-Erauzketako artearen egoera hartu dugu kontutan. Zehatzago, TACRED datu-multzoaren gainean emaitza hoberenak lortzen dituzten sistemak. 2. atalean aipatu dugun bezala, sistema hauek bi motatan bereiz daitezke: atazari zuzendutako arkitekturadun sistemak edo arkitektura orokorreko sistemak. Atazari zuzendutako arkitekturadun sistemen artean AGGCN[6] (Guo et al., 2019a) aukeratu dugu. Arkitektura orokorra duten sistemen artean berriz TRE[7] (Alt et al., 2019) eta BERT_{EM} (Baldini Soares et al., 2019) aukeratu ditugu.

4.1.1 AGGCN

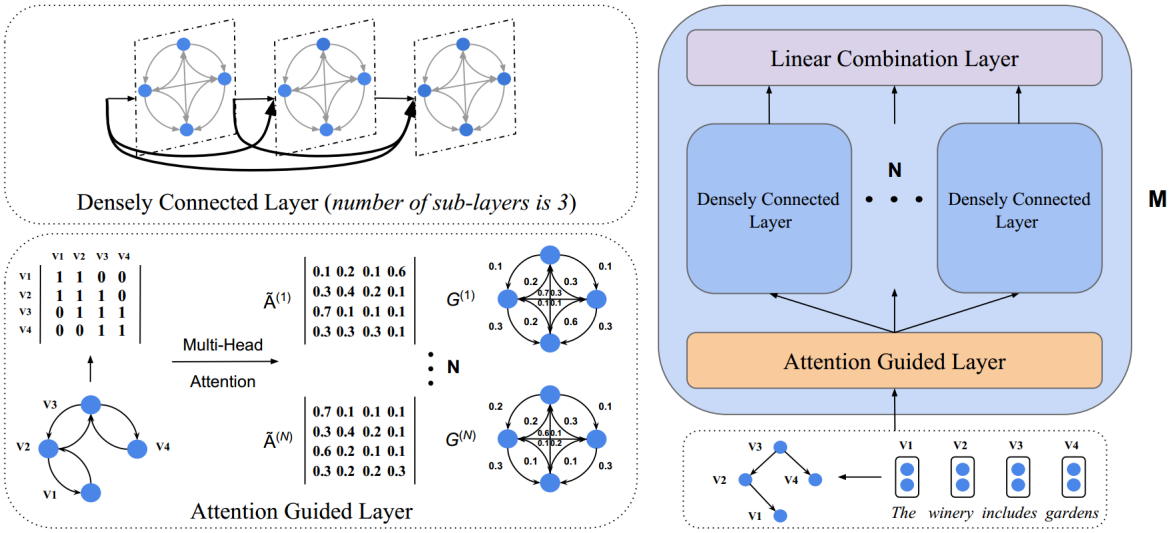
Atentzio bidez zuzendutako grafoen gaineko sare konboluzionalak aurkezten dituzte Guo et al. (2019a). Sare hauek dependentzi zuhaitzetan eta autoatentzioan (*self-attention* ingelesez) oinarritzen dira erlazio-erauzketarako egokiak diren hitzen errepresentazioak lortzeko. Horretarako, Grafoen gaineko Sare Konboluzionalak (GCN, *Graph Convolutional Networks* ingelesez) eta Atentzio bidez zuzendutako Grafoen gaineko Sare Konboluzionalak (AGGCN, *Attention Guided Graph Convolutional Networks* ingelesez) geruzak konbinatzen dituzte.

Grafoen gaineko Sare Konboluzionalak (GCN) Lehen aldiz erpin sailkapen erdigainbegiratuan (Kipf eta Welling, 2017) agertu ziren. Hasieran zuzendugabeko grafoetara mugatuta egon arren, ez zen denbora asko pasa Marcheggiani eta Titov (2017) grafo zuzenduen gainean erabili zuten arte. Azken hauek, dependentzi-zuhaitzen gaineko informazioa kodetzeko erabili zuten rol-semantiko etiketazio atazean. Formalki, GCNa horrela definituta dago:

$$h_i^{(l)} = \sigma(\sum_{j=1}^{n} A_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)}) \tag{4}$$

non, A dependentzi-zuhaitza errepresentatuko duen auzokidetasun-matrizea, $h_j^{(l-1)}$ aurreko uneko hitzaren errepresentazioa, $W \in \mathbb{R}^{d \times d}$ pisu-matrizea, $b \in \mathbb{R}^d$ alboratze-bektorea, d ezkutuko-egoeraren taimaina eta σ aktibazio-funtzioa diren.

[6] `github.com/Cartus/AGGCN`
[7] `github.com/DFKI-NLP/TRE`



Irudia 5: AGGCN bloke bat osatzen duten geruzen deskribapena erakusten du irudiak. Irudian atentzio bidez zuzendutako geruza (Atention Guided Layer ingelesez), konexio dentsodun geruza (Densely Connected Layer ingelesez) eta konbinaketa linearreko geruza (Linear Combination Layer ingelesez) agertzen dira. Irudia Guo et al. (2019a) autoreen artikulutik aterata.

AGGCN geruza Atentzio bidez zuzendutako grafoen gaineko konboluzio geruza bat hainbat elementuz dago osatuta 5. irudiak erakusten duen bezala. Lehendabizi, atentzio-mekanismoan oinarritutako auzokidetasun-matrizeak kalkulatuko dira 5. ekuazioa jarraituz (5. irudiaren ezker-behealdean dago adibide bat). Behin N auzokide-matrizeak edukita, N konexio dentsodun geruza paralelo aplikatzen dira (irudiaren ezker-goikaldean), azkenik, konbinaketa linearreko geruzaren bitartez bat egiteko (irudiaren eskubialdean). Guo et al. (2019a)ek proposatutako eredua sortzeko AGGCN M bloke konkatenatu egiten dira.

Atentzio bidez zuzendutako geruza Ohiko GCNak ez bezala, atentzio-mekanismoetan oinarritutako auzokidetasun-matrizeak erabiltzen dituzte AGGCN geruzetan. 5. irudiaren ezker-beheko aldean ikus daiteke adibide bat. Hain zuzen ere, A auzokidetasun-matrizea erabili beharrean, atentzio-mekanismo baten bitartez kalkulatutako $\widetilde{A}$ erabiltzen dute. Zehazki, AGGCNek (Guo et al., 2019a) buru-anitzeko atentzioa (Vaswani et al., 2017) erabiltzen dute. Beraz,

$$\widetilde{A}^{(t)} = softmax(\frac{h^{(l-1)}W_t^Q \times (h^{(l-1)}W_t^K)^T}{\sqrt{d}}) \quad (5)$$

non, $W_t^Q \in \mathbb{R}^{d\times d}$ eta $W_t^K \in \mathbb{R}^{d\times d}$ t. buruko kontsulta eta gako proiekzio matrizeak diren. Modu honetan N auzokide-matrize sortzen dituzte.

HAP masterra

Konexio dentsodun geruza Guo et al. (2019b)-ren ideia jarraituta, azpigeruzen arteko konexio dentsuak barneratzen dituzte. Konexio hauek uneko egoera ezkutua baino aurreko egoera ezkutu guztien informazioa uneko geruzan edukitzean datza. Beste modu batera esanda, l. geruzaren sarrera bezala $g^{(l)}$ izango da, non $g^{(l)} = [x; h^{(1)}; ...; h^{(l-1)}]$ aurreko egoera ezkutuen konkatenazio bat izango den. Praktikan, konexio dentsodun geruza hauek L azpigeruza dituzte, ikusi 5. irudiko goi-ezkerreko adibidea. Azpigeruza bakoitzaren tamaina $d_{azpi} = d/L$ izango da, non d egoera ezkutuaren tamaina den. Konexio dentsodun geruza hauen irteera L azpigeruza guztien irteeren konkatenazioa izango da, eta, beraz, d tamaina berreskuratuko da. Formalki, azpigeruza bakoitzak hurrengo itxura izango du:

$$h_{ti}^{(l)} = \sigma(\sum_{j=1}^{n} A_{ij}^{(t)} W^{(l)} g_j^{(l)} + b^{(l)}) \tag{6}$$

non, A auzokide-matrizea $\widetilde{A}^{(t)}$ atentzio-mekanismoan oinarritutako auzokide-matrizearengatik ordezkatzen den eta $h^{(l-1)}$ egoera ezkutua $g^{(l)}$ egoeren konkatenazioarengatik ordezkatzen den.

Konbinaketa linearreko geruza Lehen esan bezala, buru-anitzeko atentzioa erabiltzen dute N auzokide-matrize sortzeko. Beraz, N konexio dentsodun geruzen irteera konbinatu ahal izateko hauen konbinaketa linear bat proposatzen dute:

$$h_{konb} = W_{konb} h_{konkat} + b_{konb} \tag{7}$$

non, h_{konkat} konexio dentsodun geruzen irteeren konkatenazioa den, hau da, $h_{konkat} = [h^{(1)}; ...; h^{(N)}]$. $W_{konb} \in \mathbb{R}^{(d \times N) \times d}$ pisu-matriza eta $b_{konb} \in \mathbb{R}^d$ alboratze bektora diren.

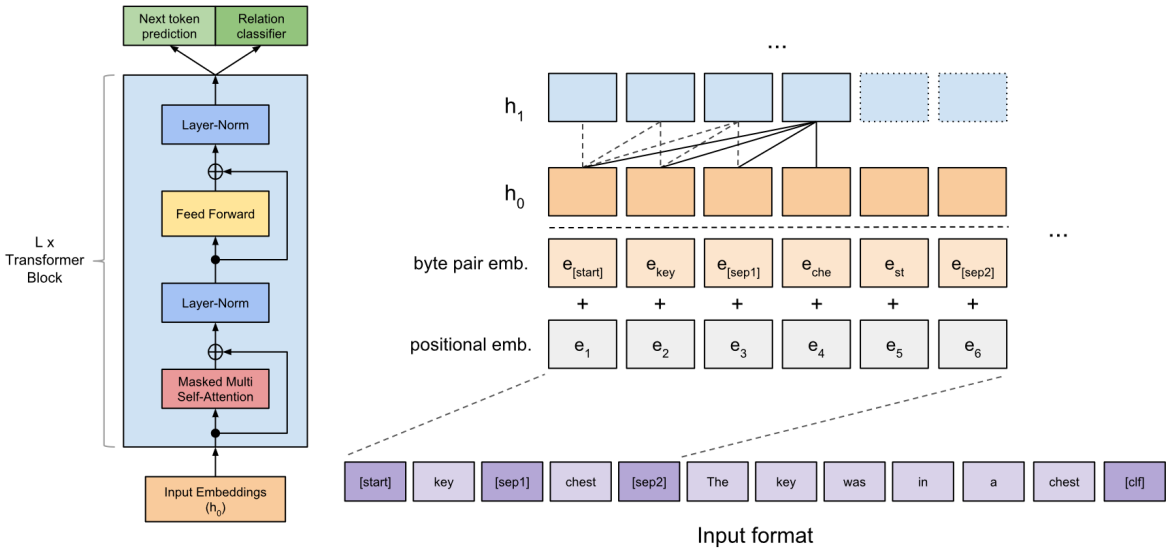
4.1.2 TRE

Gaur egun arrakasta handiko aurrentrenamendu-birdoitze (*pretraining-finetuning* ingelesez) estrategia jarraitzen dute Alt et al. (2019)-ek. Proposatzen duten eredua 6. irudian agertzen da. TRE Transformer-deskodetzaile (Liu et al., 2018) arkitektura jarraitzen du, Transformer (Vaswani et al., 2017) originalaren aldaera bat dekodetzaile zatia bakarrik erabiltzen duena. Oinarrizko egitura horri, erlazio-erauzketara egokitzeko bi hobekuntza proposatzen dituzte: alde batetik, erlazio sailkapenerako geruza bat, eta, bestetik, entitateei buruzko errepresentazio egokia lortzeko sarrera errepresentazio berri bat.

Aurrentrenatutako eredua Esan bezala, TRE aurrentrenamendu-birdoitze estrategia jarraitzen du. Zehatzago, OpenAI taldeak aurkeztutako GPT (Radford et al., 2018) aurrentrenatutako hizkuntza-ereduan oinarritzen da. Hizkuntza-eredu honek hurrengo token iragarketa (*Next Token Prediction* ingelesez) ataza jarraituz entrenatua izan da. Ataza, hurrengo egiantza helburu funtzioa optimizatzean datza:

$$\mathcal{L}_1(\mathcal{C}) = -\sum_i logP(c_i | c_{i-k}, ..., c_{i-1}) \tag{8}$$

HAP masterra



Irudia 6: TRE arkitektura. GPT ereduaren aldaera bat erlazio-erauzketara zuzendutako geruza eta sarrera errepresentazio berri batekin.

non, c_i uneko tokena adierazten duen eta k leiho luzeera. Helburu-funtzio horrek erakutsi izan ditu ezagutza-transferentzia ahalmen handiak, batez ere testu-sorkuntza atazetan (Radford et al., 2019; Brown et al., 2020).

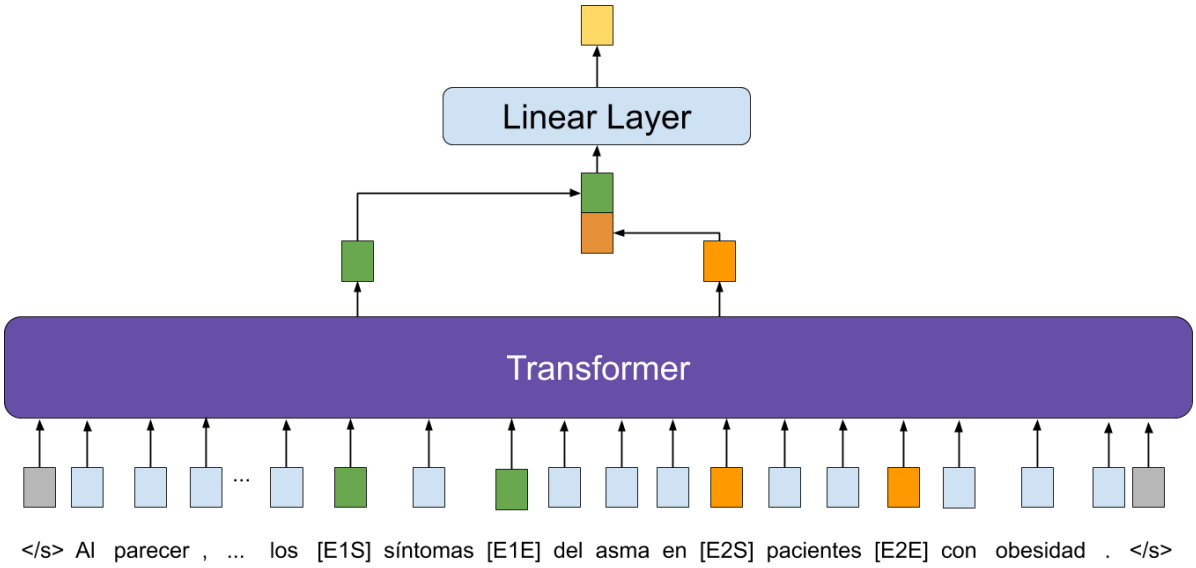
Erlazio-erauzketara egokitzea Erlazio-erauzketa sailkapen ataza bat bezala planteatzen da, baina GPT aurrentrenatutako hizkuntza-eredua testu-sorkuntzara dago bideratuta. Hori dela eta, aipatutako arkitekturan aldaketa batzuk proposatzen dituzte Alt et al. (2019) autoreek. Lehendabizi, sarrera-errepresentazio berri bat aurkezten dute, non, sarrera-sekuentziaren hasieran erlazioan parte hartzen duten entitateak aipatzen diren. Modu honetan, hizkuntza-ereduak esaldia irakurtzera doanean helburu entitateak zeintzuk diren badakizki. Bestetik, sailkapen geruza bat gehitzen diote oinarrizko arkitekturari, geruza hori hurrengo itxura du:

$$P(r|e_1, e_2, x) = softmax(W^r h_m^L + b^r) \tag{9}$$

non, r iragarri beharreko erlazioa den, e_1 eta e_2 erlazioan parte hartzen duten entitateak diren, x testuinguruko token-sekuentzia den, W^r sailkapen geruzako pisu matrizea den, b^r sailkapen geruzako alborapen bektorea den eta h_m^L azkeneko geruzako azkeneko tokenaren egoera ezkutuko errepresentazioa den. Azkenik, ataza ikasteko hurrengo entropia gurutzatua proposatzen dute:

$$\mathcal{L}_2(\mathcal{D}) = -\sum_{i}^{|D|}\sum_{j}^{|r|} r_{ij} log P(r_{ij}|e_{i1}, e_{i2}, x_i) \tag{10}$$

HAP masterra



Irudia 7: BERT_{EM} arkitektura. BERT (edo beste aurrentrenatutako hizkuntza-eredu bat) aldaera bat non EM (Entitate Markak) gehitzen diren erlazioan parte hartzen duten entitateen limiteak adierazteko.

$$\mathcal{L}(\mathcal{D}) = (1 - \lambda)\mathcal{L}_1(\mathcal{D}) + \lambda\mathcal{L}_2(\mathcal{D}) \tag{11}$$

non, aurrentrenamendurako erabili den $\mathcal{L}_1$ eta galera-funtzio berria $\mathcal{L}_2$ konbinatzen duten λ koefiziente bat erabiliz.

4.1.3 BERT_{EM}

Aurreko metodoaren antzera, BERT_{EM}-ek ere aurrentrenamendu-birdoitze estrategia jarraitzen du, baina TREk ez bezala, Baldini Soares et al. (2019) autoreek aurrentrenatutako hizkuntza-eredu eta sarrera-errepresentazio berriak proposatzen dituzte. Kasu honetan, proposatutako arkitektura ez dago hizkuntza-eredu zehatz baten menpe.

Sarrera-errepresentazioari dagokionez, lau token berri gehitzen dira: [E1S], [E1E], [E2S] eta [E2E]. Token horiek, hau da, Entitate Markak, esaldi batean erlazioan parte hartzen duten entitateen mugak zehazteko erabiltzen dira (ikusi 7. irudia). Formalki, $r = (x, e_1, e_2)$ erlazio adibidea da non $x = [x_0, x_1, ..., x_n]$ esaldi bateko token sekuentzia den eta e_1 eta e_2 erlazioko bi entitateak diren. Lehendabizi, x token sekuentzia eraldatzen da EM-k barneratuz:

$$\tilde{x} = [x_0, ..., [E1S], e_1, [E1E], ..., [E2S], e_2, [E2E], ..., x_n]$$

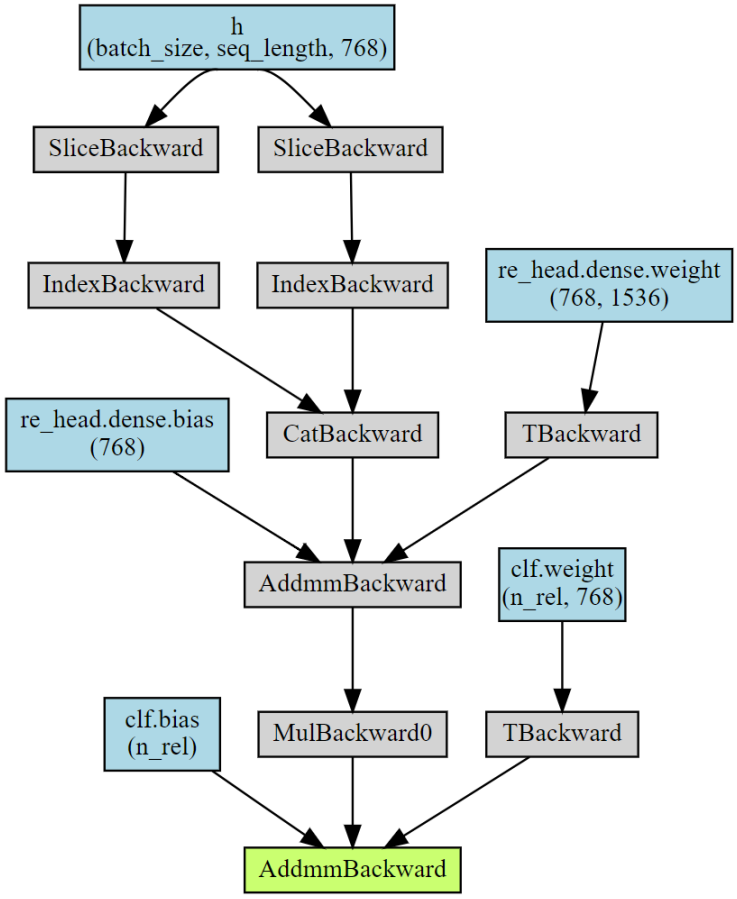
ondoren, $h = Transformer(\tilde{x})$ testuinguru-mailako hitzen errepresentazioak lortzen dira hizkuntza-eredua erabiliz. Jarraian, f_θ erlazio-errepresentazioa kalkulatzen da:

$$f_\theta(h) = \sigma(W_{re}[h_{E1S}; h_{E2S}] + b_{re}) \tag{12}$$

HAP masterra

non, $[h_{E1S}; h_{E2S}]$ erlazioko bi entitateen errepresentazioen konkatenazioa den, $W_{re} \in \mathbb{R}^{2d \times d}$ geruza linearreko pisu matrizea eta $b_{re} \in \mathbb{R}^{d}$ alborapen bektorea diren. Azkenik, aurreko ereduan bezala sailkapen geruza bat (ikusi 9. ekuazioa) gehitzen zaio ereduari.

4.2 Sistemen inplementazioa



Irudia 8: Erlazio-erauzketarako EM barneratzearen estrategia jarraitzen duen geruzaren konputazio grafoa.

Gaur egun ohiko praktika bat da publikatutako artikuluekin batera erabilitako kodea ere publikatzea. Hori da kasua bai AGGCN baita TRE sistementzat, baina ez BERT_{EM}-entzat. Beraz, azterketa egin ahal izateko BERT_{EM} berinplementatzea izan da lan honen lehendabiziko helburuetako bat. Inplementaziorako PyTorch[8] erabili dugu diferentziazio automatikorako ingurune bezala. Inplementatu beharreko eredua bi zatitan banatzen da: kodetzailea edo hizkuntza-eredua eta erlazio-erauzketarako geruza. Kodetzailearen kasuan

[8]PyTorch web orria: `https://pytorch.org/`

HuggingFace taldearen Transformers (Wolf et al., 2019) liburutegia erabili dugu. Liburutegi honek Transformer arkitekturan oinarritutako ereduen inplementazioak eta jadanik entrenaturiko ereduak eskaintzen ditu. Erlazio-erauzketarako burukoaren kasuan guk berinplementatu dugu Baldini Soares et al. (2019) autoreen gidak jarraituz. Geruzaren konputazio grafoa 8. irudian ikus dezakegu.

4.3 Ereduen arteko konparaketa

Aurreko atalean aurkeztu diren hiru sistemak konparatzea da atal honen helburua. Horretarako TACRED datu-multzoa erabili da, erlazio-erauzketarako datu-multzo estandarra kontsideratua dagoena. Sistemak konparatzeko hauen inplementazioak erabili dira artikuluetan erakutsitako emaitzekin batera.

	Doitasuna	Estaldura	F1
AGGCN (Guo et al., 2019a)	**73.1**	64.2	68.2
TRE (Alt et al., 2019)	70.1	65.0	67.4
‡BERT_{EM}	64.0	**71.0**	67.0
BERT_{EM} (Baldini Soares et al., 2019)	-	-	**70.1**

Taula 1: TACRED datu-multzoaren gaineko emaitzak. Guk berinplementatutako sistema ‡ batekin dago irudikatuta.

Lortu ditugun emaitzak 1. taulan daude irudikatuta. Bai AGGCNek baita TREk lortzen dituzten emaitzak beraien artikuluetan aipatutakoak dira. BERT_{EM}ren kasuan desberdintasun nabari bat dago gure inplementazio eta artikuluan erreportaturiko emaitzen artean. Desberdintasun hori eredua ikasteko garaian erabili diren hiperparametroengatik izan daiteke. Baldini Soares et al. (2019) erabilitako parametroak gure ingurune mugatuan ezin izan ditugu erabili, memoria arazoengatik. Bestalde, sistema honen beste inplementazio batek[9] gure emaitzen antzeko balioak lortzen ditu (65.0eko eta 67.0ko F1 *bert-base* eta *bert-large* erabiliz). Beraz, azterketarako guk lorturiko emaitzak hartuko ditugu bakarrik kontutan.

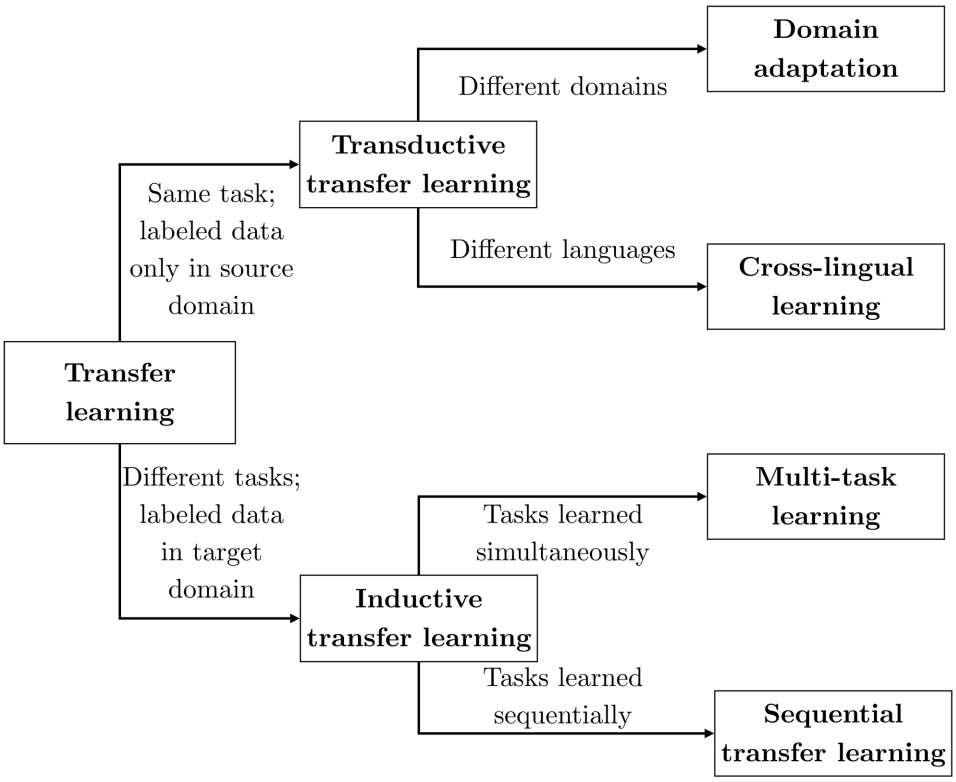
Hiru sistemek emaitza oso antzekoak lortzen dituzte, beraien artean AGGCNk lortutakoak hoberenak izanik. Egitura orokorreko bi sistemek F1 neurria soilik kontuan hartuta emaitza berdinak lortzen dituzte. Hori horrela izanda, BERT_{EM} eta TRE ereduen artean BERT_{EM}ekin gelditzea erabaki dugu eskaintzen duen hizkuntza-eredu malgutasunarengatik. Bestetik, AGGCN eta BERT_{EM} ereduen artean bigarrena aukeratzea erabaki dugu, arrazoi nagusia informazio sintaktikoaren beharra izan da. Gure helburua aukeratutako sistema domeinu zehatz batean aplikatzea izanda, eta informazio sintaktikoa eskura ez izatea arazo bat izan daiteke. Bi sistemen arteko diferentzia txikia dela kontutan hartuta, informazio sintaktiko zaratatsuak ekarri dezakeen errendimendu jaitsiera handiegia izan daiteke. Eta beraz, BERT_{EM} aukera egokiena dela pentsatzen dugu.

[9] `https://github.com/zhpmatrix/BERTem`

5 Medikuntza domeinuko erlazio-erauzketa

Aurreko atalean dagoeneko deskribatu dugu inplementatu dugun sistema. Sistema hori ingeleserako eta domeinu orokorrerako testuentzat artearen egoerako errendimendua lortzen duela ziurtatu dugu. Gure erronka orain domeinu eta hizkuntza berri batera egokitzea da, hain zuzen ere gaztelerazko medikuntza domeinuko testuetara doitzea. Horretarako 3.2.2 atalean aipatutako eHealth-KD 2020 ataza-partekatuan parte hartu dugu.

Ohiko ikasketa-sakoneko edo ikasketa-automatikoko algoritmoek ez bezala, gure sistema aurreikasitako eredu batean oinarritzen da. Azken eredu hau baldintzatuko du gure sistema osoaren hizkuntza eta oinarrizko ezagutza. Hori dela eta, domeinu eta hizkuntz berri batera egokitu nahi badugu, zati hori izango da egokitu beharko duguna.



Irudia 9: LNPrako ezagutza-transferentziaren taxonomia (Ruder, 2019)

Gaur egun arrakasta handia daukan ezagutza-transferentzia estrategia erabiliko dugu domeinura egokitzeko. Egoera eta transferentzia moten arabera ezagutza-transferentzia mota bat edo beste aplikatzen da (ikusi 9. irudia). Gure kasuan anotatutako datumultzoa **helburuko domeinuko** testuak erabiltzen ditu. Beraz, jarraituko dugun estrategia **ezagutza-transferentzia sekuentziala** izango da. Baina ez gara besterik gabe bukaerako atazara birdoitzera mugatu, hurrengo ataletan azalduko ditugun MLM eta MTB doikuntzak sekuentzialki aplikatzen saiatu gara ahalik eta azken atazara zuzendutako ezagutza handiena izan dadin.

5.1 Hizkuntzara egokitzea

Lehendabiziko eta oinarrizko pausua gure sistema gaztelerara egokitzea da. Hori lortzeko, beste hizkuntzen artean, gaztelerarekin entrenatua izan diren bi hizkuntza-eredu aukeratu ditugu: XLM (Lample eta Conneau, 2019) eta XLM-RoBeRTa (Conneau et al., 2019) (XLMR, motzean). Bi ereduek Transformer arkitekturan daude oinarrituta eta hizkuntza anitzeko hizkuntza-eredu maskaratuak (*Masked Language Models* ingelesez) dira. Beraien arteko desberdintasun nagusienak erabilitako datu, hizkuntza eta parametro kopuruak dira, baina entrenatzeko estrategia eta oinarrizko arkitektura berdinak dira.

Hizkuntza-eredua	**XLM-17**	**XLMR-base**	**XLMR-large**
Hizkuntza kopurua	17	100	100
Tokenizazioa	BPE	SPM	SPM
Hiztegi tamaina	200k	250k	250k
Transformer bloke kopurua	16	12	24
Atentzio-buruko kopurua	16	12	16
Egoera-ezkutuko tamaina	1280	768	1024
Parametro kopurua	570M	280M	560M
Gaztelerazko testu kopurua	-	53.3 GiB	53.3 GiB
Guztira testu kopurua	-	2.5TB	2.5TB
Aurrentrenamendu estrategia	MLM	MLM	MLM

Taula 2: Aztertutako hizkuntza-ereduen informazio orokorra.

2. taulan ikus dezakegu XLM, XLMR-base eta XLMR-large ereduen arteko konparaketa bat. Eredu bat deskribatzen duten eremuak 3 multzo handitan bana daitezke: tokenizazioarekin erlazionatuta dauden eremuak, hau da, hizkuntza kopurua, tokenizazio mota eta hiztegi tamaina; ereduaren arkitekturarekin erlazionatuta dauden eremuak, hau da, transformer bloke kopurua, atentzio-buruko kopurua, egora-ezkutuko tamaina eta parametro kopurua; eta azkenik, aurre-entrenamenduarekin erlazionatuta dauden eremuak, hau da, testu kopurua eta estrategia. Multzo horiek kontutan izanik, XLMR-base eta XLMR-large ereduen artean arkitekturan tamainan soilik desberdintzen dira, izenek adierazten duten bezala *large* bertsioa handiago da. XLM eta XLMR ereduen artean berriz aldaketa nahiko daude, baina orokorrean, XLMek hizkuntza gutxiago onartzen ditu tamainan handiagoa izan arren.

5.2 Domeinura egokitzea

Bigarren pausua hizkuntza-eredua domeinura egokitzea da. Horretarako, ezagutza-transferentzian eta datu-gehikuntzan oinarritutako estrategiak sekuentzialki (ikusi 9. irudia) aplikatzea erabaki dugu. Gaztelerazko hizkuntza-eredua domeinura egokitzeko ideiarekin estrategia hauek domeinuko testuekin batera aplikatuko ditugu. Hori dela eta, domeinuko corpusak ere erauzi behar izan ditugu. Atal honetan erauzitako corpusak eta aplikatutako estrategiak azalduko ditugu.

HAP masterra

5.2.1 Medikuntzako corpusen erauzketa

Domeinura egokitzeko medikuntzari lotutako bi corpus erabili ditugu. Batetik Medlineko Laburpenak deitu dugun corpusa eta bestetik Medikuntzako Corpus Bilduma deitu dugun bigarren corpus bat.

- **Medline Laburpenak (MA)** Corpus hau medikuntzako artikulu zientifikoen laburpenak ingelesez eta gazteleraz jasotzen dituen corpus paralelo bat da. 17.254 laburpen jasotzen ditu (8627 hizkuntza bakoitzeko) eta guztira 3.5M tokenez dago osatuta garbitu eta esaldi errepikatuak kendu ondoren. Corpus honen abantaila nagusiena UMLS[10] (Unified Medical Language System) entitateekin anotatuta dagoela da, honek ahalbidetuko digu geroago aipatuko dugun *Matching The Blanks* (MTB) doikuntza egiteko. Corpus honen adibide bat erakusten du 10. adibidea.

  Se observó la [religiosidad $_{C0687003}$] aislada [no $_{C1298908}$] presenta un factor protector [eficaz $_{C1704419}$], pero presenta un [papel $_{C0030351}$] importante en el desarrollo de resiliencia ante la [enfermedad $_{C0012634}$] y una [fuente $_{C0449416}$] [constructora $_{C0403065}$] de red de [apoyo $_{C0344211}$] al [anciano $_{C0001792}$].

  Irudia 10: MA corpuseko adibide bat UMLS erreferentziekin anotatuta.

- **Medikuntzako Corpus Bilduma (MCB)** Corpus hau hainbat iturrietatik jasotako anotatugabeko testuz dago osatuta. Corpus honen bilketaren arrazoiak bi dira: medikuntzaren domeinuko gaztelerazko corpus handiago bat biltzea eta artikulu zientifikoetako hizkuntza teknikoaz gain hizkuntza errazago bat erabiltzen duten testuen bilketa dira. Corpusa osatzen dituzten testu iturriak hiru dira: Medline Laburpenak corpusa, Medical Web Crawl[11] eta MedlinePlus Health Topics[12]. Jasotako testuak ingelesez eta gazteleraz daude, azken honen testu kopuruari lehentasuna emanda. Garbitu eta adibide errepikatuak kendu ondoren 7.4M tokenez osatutako **testu hutsezko** corpusa bildu dugu. MCB corpusak anotaziorik ez dituenez MTB doikuntzarako ez digu balio baina bai orain azalduko dugun MLM doikuntzarako. 11. irudiak MCB corpuseko 3 adibide erakusten ditu.

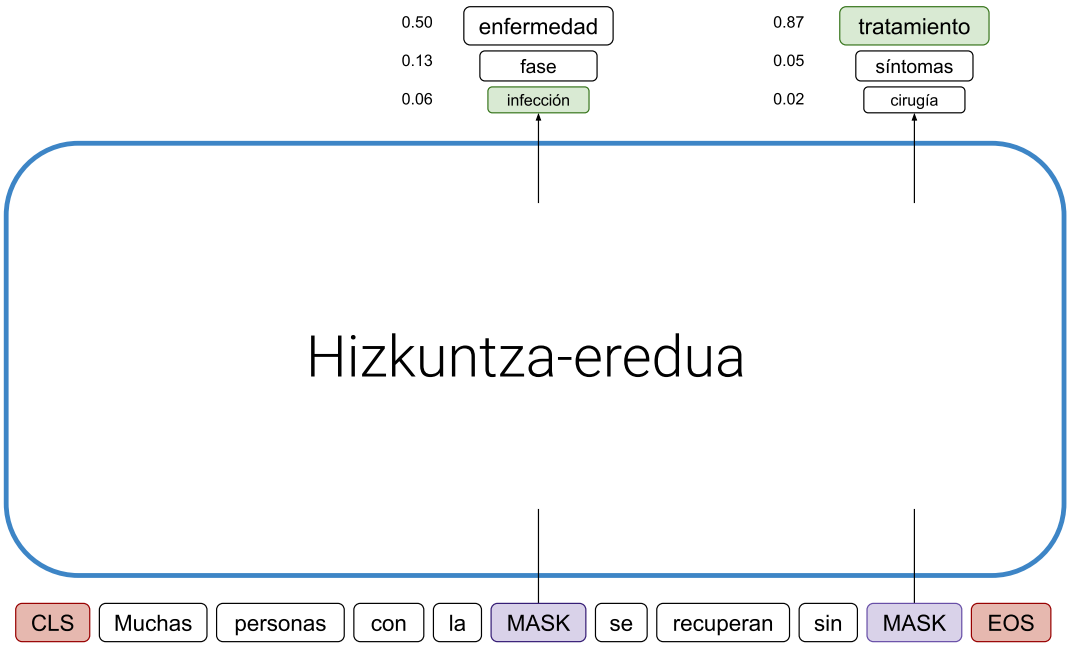
  Ademas, la calidad de vida muestra una mayor relacion con la adaptacion a la enfermedad que con sus sintomas.

  After cancer treatment, many survivors want to find ways to reduce the chances of their cancer coming back.

  Dormir sobre el lado izquierdo también mejora el flujo de sangre entre el corazón, el feto, el útero y los riñones e igualmente quita la presión sobre el hígado.

  Irudia 11: MCB corpuseko 3 adibide.

[10] `www.nlm.nih.gov/research/umls/index.html`
[11] `www.ufal.mff.cuni.cz/ufal_medical_corpus`
[12] `www.medlineplus.gov/spanish/healthtopics.html`



Irudia 12: MLM doikuntzan oinarritutako hizkuntza-eredu baten funtzionamendua. Erabilitako adibidea guk medikuntza domeinura birdoitutako $\text{XLMR}_{\text{base}}$ baten irteera da.

5.2.2 MLM doikuntza

Hizkuntza-eredu maskaratuak (MLM, *Masked Language Model* ingelesez), modu autogainbegiratu batez maskaren bidez ezkutatutako hitzak berreskuratuz entrenatu diren hizkuntza ereduak dira. Lehen aldiz BERT (Devlin et al., 2019) baina baita ere XLM eta XLM-RoBeRTa erabiltzen dute aurrentrenamendu estrategia berdina. Ataza horren bitartez, hitz batzuk ezkutatuta dituen esaldi bat emanda hasierako esaldia berreskuratzen ikasi behar du ereduak (ikusi 12. irudia). Formalki, esaldia errepresentatzen duen token sekuentzia $x = [t_0, t_1, ..., t_n]$ batean p probabilitate baten bitartez t tokenak [MASK] tokenekin ordezkatzen dira. Ondoren, $\hat{x} = f(x)$ hasierako esaldia estimatzen da f funtzio baten bitartez, hizkuntza-eredua alegia. Azkenik $\mathcal{L}(x, \hat{x})$ entropia gurutzatu galera kalkulatzen da.

$$\mathcal{L}(x, \hat{x}) = -\sum_{i}^{n} \sum_{j}^{|V|} x_{ij} \log \hat{x}_{ij} \tag{13}$$

Domeinura egokitzeko, dagoeneko MLM bitartez ikasi duten XLM eta XLM-RoBeRTa, estrategia bera jarraituta domeinuko testuak erabiliz ereduak birdoitzea da ideia. Modu honetan, domeinuko terminologia hobeto errepresentatuta egotea espero dugu.

HAP masterra

5.2.3 MTB doikuntza

Matching The Blanks (MTB) Baldini Soares et al. (2019)-ek lehenbiziz proposatutako entrenamendurako erdi-gainbegiratutako estrategia bat da. Estrategia honek bi entitateen arteko erlazio-errepresentazioa ikasteko aukera aurkezten du eskuzko anotaziorik gabe. Urruneko-gainbegiraketak (Mintz et al., 2009) ezagutza-base batez baliatzen da bi entitateen arteko erlazioa esleitzeko. Etiketak modu zaratatsuan esleitzeari ekiditeko saiakeran, MTB erdi-gainbegiraketak konparaketa bidezko ikasketa (*contrastive learning* ingelesez) proposatzen du, hain zuzen ere, ausazko entitate pareak erabiltzen dira beraien arteko erlazioa zehaztu gabe.

Edozein bi esaldi emanik, ezkutatutako entitate pareak berdinak badira, erlazio bera adierazten dutela onartuko dugu. Aldiz, pare desberdinak ezkutatzen badira, erlazio desberdin bat erakusten dutela esango dugu. Ataza sailkapen bitar bat bezala planteatu dugu eta atazaren intuizioa honakoa da: semantikoki antzekoak diren esaldiak modu antzekoan irudikatuko dira eta ikasitako parametroak erabilgarri izango dira erlazio erauzketako atazan. Sistemak erlazioa iragarri beharrean entitate pareak iragartzen ez ikasteko [BLANK] token bereziengatik ordezkatzen dira esaldiko entitateak probabilitate baten bitartez. Adibidez, 13. irudian MTB datu-multzo baten hiru sarrera ikus daitezke, (1) eta (2) sarrerek entitate-pare berdina partekatzen dute, beraz, adibide positibo bat sortuko lukete. Aldiz, (3) entitate bakar bat besterik ez du partekatzen beste bi sarrerekin, beraz, (1) eta (3) edo (2) eta (3) sarrera-pareek adibide negatiboak sortuko lituzkete.

> (1) Se observó actividad de CK en [BLANK] con dengue con presencia de [BLANK] como vómito, hematemesis y dolor abdominal.
>
> (2) Al parecer, existen mecanismos comunes a ambas patologías que pueden influir en la exacerbación de los [BLANK] del asma en [BLANK] con obesidad.
>
> (3) El [BLANK] promedio para el inicio de ENT fue de 30 (23,5) horas, y el 88,7% de los [BLANK] alcanzaron el objetivo nutricional en 48 horas.

Irudia 13: MTB datu-multzoaren hiru sarrera. Lehendabiziko bi adibideak entitate pare berdina partekatzen dute: *paciente* eta *sintomas*. Hirugarrenak berriz *paciente* eta *tiempo* entitate parea dauka, eta beraz, besteekin soilik entitate bakarra partekatzen du: *pacientes*

Formalki, $r_1 = f(x, e_{11}, e_{12})$ eta $r_2 = f(y, e_{21}, e_{22})$ erlazio errepresentazioak emanda, non x eta y esaldi bakoitzaren token sekuentzia diren, e_{ij} $i.$ esaldiko $j.$ entitatea den eta

$$\delta_{e_{11}e_{21}} = \begin{cases} 1 & \text{baldin} \quad e_{11} = e_{21} \\ 0 & \text{bestela} \end{cases}$$

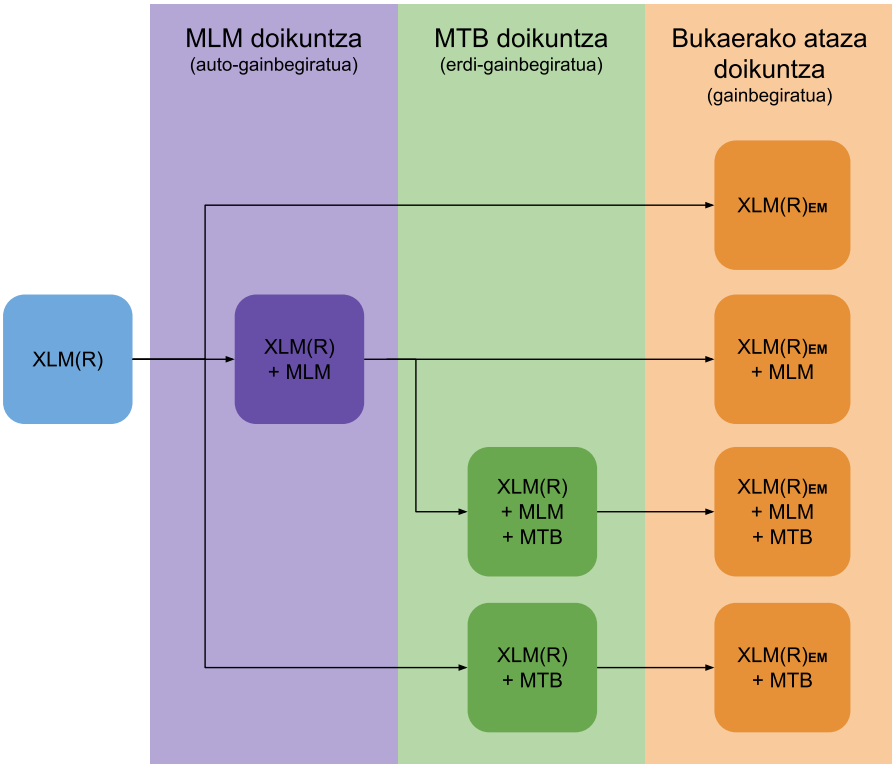
δ Kronecker delta izanda, aurrentrenamendu estrategia honekin ikasteko optimizatu beharreko galera funtzioa hurrengo entropia-gurutzatua da:

HAP masterra

$$\mathcal{L}_{\text{MTB}} = \sum_{i}^{|\mathcal{D}|} \delta_{e_{11}e_{21}}\delta_{e_{12}e_{22}} \log \sigma(r_{i1} \cdot r_{i2}^{\top}) + (1 - \delta_{e_{11}e_{21}}\delta_{e_{12}e_{22}}) \log(1 - \sigma(r_{i1} \cdot r_{i2}^{\top})) \quad (14)$$

5.3 Esperimentuen garapena

Lan honetan sistema eta teknika asko esploratu ditugu. Esploratutako aukerak 14. irudian ikus daitezke. Aukera hauek aurreko atalean aipatu ditugun estrategiekin daude osatuta. Ezkerretik eskubira, bai atazarekiko menpekotasuna, baita behar den gainbegiraketa maila handitzen dira. Azterketa guzti horiek denbora tartea handiak behar izan dituzte bai inplementaziorako baita exekuziorako ere. Hori dela eta, garapena bi fase nagusitan banatu dugu proiektuko epemugetan oinarrituta. Epemuga garrantzitsuena eHealth-KD 2020 ataza partekatuarena izan da, horren arabera bi fase definitu ditugu: oinarrizko sistemaren garapena eta sistemaren hobekuntza.



Irudia 14: Domeinu eta ataza zehatz batera doitzeko proposatutako aukerak. Hiru maila desberdintzen dira, ezkerretik eskubira bukaerako atazara dagoen menpekotasuna handitzen da.

Oinarrizko sistemaren garapena Fase honetan gure oinarri-lerroa izango den sistema garatu dugu. Horretarako, hiperparametro bilaketa mugatzeko helburuarekin erabiliko

dugun aurrentrenatutako hizkuntza-eredua XLM izatea erabaki dugu. Zehazki XLM familiako 17 hizkuntzadun hizkuntza-eredua, Lample eta Conneau (2019) autoreek erreportatutako emaitzen arabera gaztelaniarekin hobeto funtzionatzen duelako aukeratu dugu eredua. Erabiliko dugun eredua finkatuta daukagula hurrengo esperimentuak gauzatu ditugu:

1. **eHealth-KD 2020 atazeko garapen partizioaren gainean birdoitze hiperparametro bilaketa gauzatu:** Gure lehendabiziko pausua XLM eredua erabiliz garapen datu-multzora hobekien doitzen den hiperparametro multzoa aurkitzea izan da. Eginkizun honen arrazoia oinarri-lerro egoki bat lortzea da. Pausu honen ondorioz 14. irudiko XLM_{EM} sistema garatu dugu.

2. **MTB doikuntza inplementatzea:** Bigarren pausurako MTB doikuntza ikusi dugu interesgarri. Pausu honetan MTB bitartez sistema doitzeko kodea garatu eta MA corpusa lortu dugu. Baita ere, MTB doiketaren hiperparametro probak egin ditugu, eta XLM+MTB sistema entrenatu dugu.

3. **MTBra doitutako eredua atazara birdoitzea:** Fase honen azkeneko eginkizuna aurreko pausoan lortutako MTB eredua azken atazara doitzea da. Horretarako, lehen pausua errepikatu dugu baina orain XLM+MTB erabilita XLM ordez. Ondorioz XLM_{EM}+MTB sistema entrenatu dugu.

Sistemaren hobekuntza Bigarren fase honetan orain arte garatu dugun sistema hobetzea da helburua. Horretarako, oinarrizko hizkuntza-eredua gure helburu domeinura ekarrez gero emaitzak hobetuko direla hipotesiaren gainean egin dugu lan. Beraz hipotesia betetzen den edo ez ikusteko hurrengo atazak definitu ditugu:

1. **Hizkuntza-eredu berrien esplorazioa:** Pauso honetan XLMez aparte beste hizkuntza-eredu eleanitzak probatu ditugu, horien artean XLMR (Conneau et al., 2019) familiakoak. Hizkuntza-ereduekin batera hiperparametro bilaketa sakonago bat ere garatu dugu. Pausu honetan $\text{XLMR}_{\text{base-EM}}$ eta $\text{XLMR}_{\text{large-EM}}$ sistemak entrenatu ditugu.

2. **MLM doikuntza inplementatzea:** Pausu honetan bi izan dira egin ditugun eginkizunak, alde batetik MCB corpusa biltzea eta bestetik aipatutako corpusa eta aurreko pausuko hizkuntza-eredu onena erabiliz domeinuko hizkuntza-eredu batzuk entrenatzea hiperparametro desberdinekin. MLM doiketa egiteko ez dugu kode berririk inplementatu behar izan, horren ordez Transformers (Wolf et al., 2019) liburutegiak eskaintzen duen script-a[13] erabili dugu. Horren ondorioz medikuntza domeinuko $\text{XLMR}_{\text{base}}$+MLM sistema lortu dugu.

3. **MLMra doitutako eredua atazara birdoitzea:** Pausu honetan aurreko pausuan lortutako hizkuntza-eredua atazera doitu dugu. Lehenengo pausuan bezala hiperparametro bilaketa bat egin da ere. 14. irudian agertzen den $\text{XLMR}_{\text{base-EM}}$+MLM sistema garatu dugu hemen.

[13] `www.github.com/huggingface/transformers/blob/master/examples/language-modeling/run_language_modeling.py`

4. **MLM eta MTB konbinatzea:** Pausu hau orain arte egin diren hobekuntza guztiak aplikatzean datza. Emaitzen atalean (6. atala) ikusiko dugun bezala, pauso honi ekidin diogu aurreko pausuetan lortutako emaitzengatik.

5.4 Hiperparametroak

Hiperparametroen doiketa egoki bat egitea garrantzi handikoa da. Gaur egungo eredu handiekin errendimendua asko aldatzen da hiperparametroen balioen arabera, ez hori bakarrik, batzutan ezer ez ikastera ere eramaten du parametro konbinaketa oker batek. Hori dela eta, hiperparametro-bilaketa kontu handiz egin beharreko prozesu bat da, gaur egun bereziki denbora asko hartzen duelako.

Orokorrean hauek izan dira esploratu ditugun hiperparametroak:

- **Ikasketa tasa:** parametro optimizazio momentuan erroreak zenbaterainoko eragina edukiko duen adierazten duen pisua. Probatu ditugun balioak 10^{-6} eta 10^{-3} tarteko balioak izan dira, sistemen portaeraren arabera hauek aldatuz.
- **Batch tamaina:** urrats bakoitzean erabiliko diren adibide kopurua. Erabili diren balioak GPUak duen memoria eta ereduaren tamainaren araberakoak izan dira.
- **FP16:** Doitasun erdiko koma mugikorra erabiltzea kalkulua azkartzeko.
- **Gradiente-pilatze kopurua:** parametroak optimizatzeko urrats kopurua. Batch tamaina oso handia izan ezin denean, batch handiago batek ekar dezakeen onurak lortzeko erabiltzen den teknika bat da. 64ko edo 128ko batch tamainen onurak lortzeko konbinazioak probatu ditugu.
- **Epoka kopurua:** datu-multzoa ikasketa prozesuan errepikatuko den kopurua. Hemen beti 100 erabili dugu eta gelditze goiztiarraren arabera ikasketa bukatu.
- **Dropout probabilitatea:** dropout-a aplikatzeko probabilitatea. Beti 0.2 balioa erabili dugu.
- **Erregularizazio indarra:** L2 erregularizazio pisua. Parametro honetan 0.1 eta 0.01 balioak probatu ditugu.
- **Optimizatzailea:** ereduko parametroak eguneratzeko erabiliko den algoritmoa. Lan honetan lau algoritmo aztertu ditugu: gradiente jaitsiera estokastikoa (SGD), momentudun gradiente jaitsiera estokastikoa (Qian, 1999) (SGD+Momentum), Adam (Kingma eta Ba, 2014) eta AdamW (Loshchilov eta Hutter, 2017).
- **Antolatzaile linearra erabiltzea:** linearki ikasketa tasa txikitzen duen antolatzailea erabiltzea edo ez.
- **Berotze urrats kopurua:** zenbatetan berreskuratuko den hasierako ikasketa tasa. Aukera hau antolatzaile linearra erabiltzen bada aztertu dugu soilik. Antolatzaile linearra erabili dugunean 0 eta 3 aukerak probatu ditugu.

Hiperparametroa	**BERT(TACRED)**	**XLM-17**	**XLMR-base**	**XLMR-large**
Ikasketa-tasa	3×10^{-4}	3×10^{-4}	1×10^{-5}	1×10^{-3}
Batch tamaina	16	8	64	16
FP16		✓	✓	✓
Gradiente-pilatze kopurua	4	8	1	8
Epoka (maximo) kopurua	100	100	100	100
Dropout	0.2	0.2	0.2	0.2
Erregularizazio indarra	0.01	0.01	0.01	0.01
Optimizatzailea	SGD	SGD	Adam	SGD
Antolatzaile linearra erabili			✓	
Berotze urrats kopurua			0	
gelditze goiztiarra erabili	✓	✓	✓	✓
Patzientzia	3	3	3	3
Errore atalasea	0	0	0	0

Taula 3: Hiperparametro bilaketaren ondorioz lortutako konbinazio onenak. BERT eredua TACRED datu-multzoan erabili da, besteak berriz eHealth-KD 2020 atazan.

- **gelditze goiztiarra**[14] **erabili:** ikasketa momentuan errorea jaisteari uzten dionean gelditzea edo ez. Beti erabili dugu.
- **Pazientzia:** gelditze goiztiarra aplikatzeko errorea jaitsi ez duten beharrezko urrats kopurua. Hiru epoka erabili ditugu beti.
- **Errore atalasea:** errorea ez dela jaitsi adierazteko atalasea. Datu-multzoaren arabera balio desberdinak probatu ditugu.

Erabili ditugun hiperparametroak (ikusi 3. taula) orokorrean ereduen artean oso antzekoak dira. Hauek bilatzeko estandar diren hiperparametro batzuetatik abiatu gara eta ereduaren arabera konbinazio hobeak bilatu ditugu, gehienetan ereduaren tamainan oinarrituz. Enpirikoki ikusi dugu geroz eta eredua handiagoa izan orduan eta batch (batch tamaina bider gradiente-pilatze kopurua) eta ikasketa-tasa handiagoek errendimendu hobea lortzen dutela. Baita ere, Adam optimizatzailea antolatzaile linear batekin erabiltzea SGD optimizatzailea baino emaitza hobeagoak, eta azkarrago, ematen dituela ikasi dugu. Hala ere, Adam optimizatzailea erabiltzeko GPU memoria askoz handiago bat behar denez eredu txikiekin probatu ahal izan dugu bakarrik. Azkenik, FP16 entrenamendurako erabiltzea edo ez emaitzetan ez da ia nabaritzen eta ikasketako prozesua gutxi gora-behera bi bider azkartzen du.

[14] *Early Stoppping* ingelesez

6 Emaitzak

Atal honetan gure helburu izan den eHealth-KD 2020 atazan lortutako emaitzak aurkeztuko ditugu. Emaitzak, esperimentuen garapena bezala, bi zatitan banatuko ditugu: alde batetik ataza-partekatuan lortu genituen emaitzak aurkeztuko ditugu, eta ondoren aurreko atalean deskribatutako hobekuntzekin lortutako emaitzak erakutsiko ditugu.

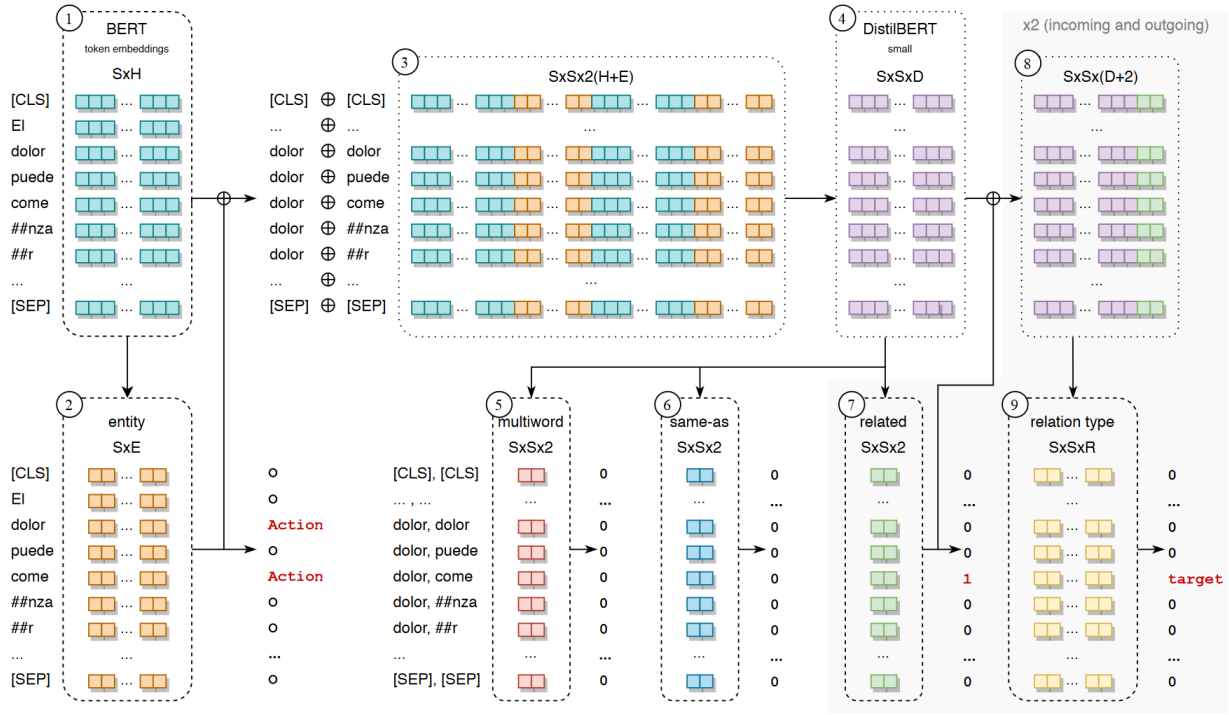
6.1 eHealth-KD 2020 ataza-partekatua

Lehen aipatu dugun bezala, ataza honetan ez da soilik erlazio-erauzketa ebaluatzen. Hori dela eta, IXA-NER-RE (Andrés et al., 2020) taldearen partaide bezala aurkeztu ginen. Bakoitza bere aldetik izendun entitateen erauzketa (NER) eta erlazio-erauzketa (RE) sistemak garatu genituen eta gero atazara batera aurkeztu ginen. Lan honetan erlazio-erauzketako atazean (B azpiataza) lortutako emaitzak aztertuko ditugu.

	Entrenamendua			Garapena			Testa		
Eredua	Doitasuna	Estaldura	F1	Doitasuna	Estaldura	F1	Doitasuna	Estaldura	F1
Vicomtech	-	-	-	-	-	-	0.672	0.515	0.583
UH-MAJA-KD	-	-	-	-	-	-	0.629	0.571	0.599
XLM_{EM}	**0.861**	**0.849**	**0.855**	0.708	0.642	0.674	0.690	**0.625**	**0.656**
XLM_{EM}*	0.767	0.795	0.781	0.707	**0.672**	**0.689**	0.649	0.619	0.633
XLM_{EM}*+MTB	0.788	0.709	0.746	**0.755**	0.616	0.678	**0.707**	0.584	0.640

Taula 4: eHealth-KD 2020 atazeko erlazio-erauzketa azpiatazan lortutako emaitzak. Alde batetik azpiatazako test datu-multzoko beste bi talde onenen emaitzak: Vicomtech eta UH-MAJA-KD. Eta bestetik entrenamendurako, garapenerako eta test datu-multzotan lortutako gure hiru sistemen emaitzak.

Beste partaideen sistemak Gure sistemen eta beste partaideen sistemen arteko konparaketa on bat egiteko hauek proposatutako sistemak ezagutzea beharrezkoa da. Horregatik modu labur batean aurkeztuko ditugu. Alde batetik, Vicomtech (García-Pablos et al., 2020) taldeak izendun entitateen erauzketa eta erlazio-erauzketa batera egiten ikasten duen sistema bat proposatzen dute (ikusi 15. irudia). Sistema horrek BERT erabiltzen du oinarri gisa. BERT gainean sailkapen geruza bat erabiltzen dute NER ataza burutzeko. Gero, BERTek emandako token errepresentazioak eta NER geruzako errepresentazioak konbinatzen dira, eta hainbat auto-atentzio (*self-attention* ingelesez) geruza ondoren RE ataza egiten dute hainbat sailkapen geruza erabiliz. Bestetik, UH-MAJA-KD (Rodríguez Pérez et al., 2020) taldeak informazio sintaktikoaz baliatzen da bere proposamenean (ikusi 16. irudia). Hain zuzen ere, karaktere mailako sare konboluzional baten bitartez lortutako hitz-bektoreak, BERT hitz-bektoreak eta informazio sintaktikoa kodetzen duten bektoreak erabiltzen dituzte CRFdun BiLSTM bat elikatzeko. Bertatik lortutako errepresentazioa

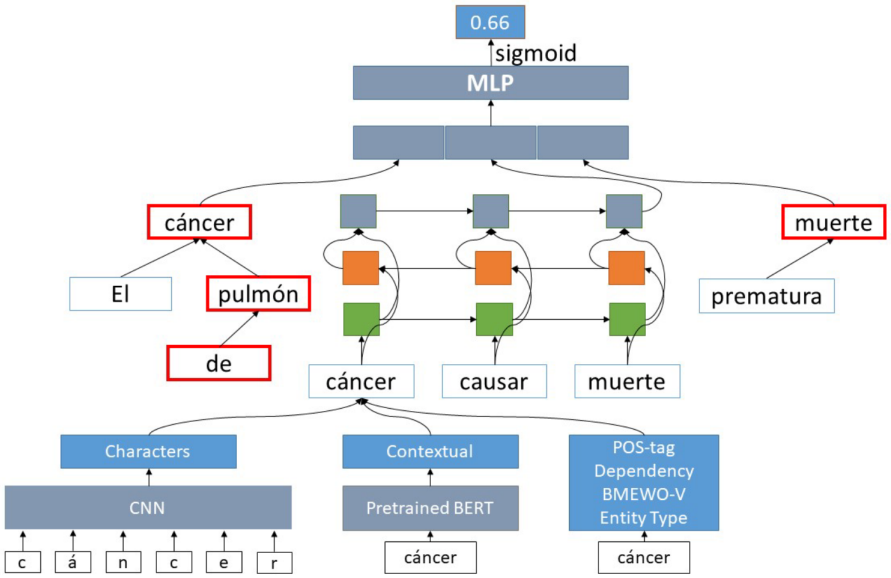


Irudia 15: Vicomtech taldeak proposatutako sistema. Irudia beraien artikulutik dago aterata.

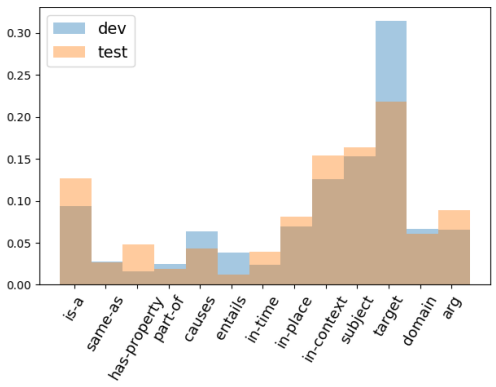
erlazioan parte hartzen duten entitateen dependentzi zuhaitza kodetzen duten errepresentazioekin konkatenatzen dute. Azkenik, sailkapen geruza bat gehitzen diote sailkapena egiteko. Ikus daitekeen bezala gure sistema, hau da hizkuntza-eredu bati sailkapen buruko bat gehitzea, baino arkitektura konplexuagoak planteatzen dituzte.

Bai gure hiru sistemen emaitzak, baita atazeko beste talde onenen emaitzak aurkezten ditugu 4. taulan. Taulak entrenamenduko, garapeneko eta test datu-multzotan lortutako micro doitasuna, estaldura eta F1 neurria aurkezten ditu. Hala ere, txapelketan testeko emaitzak besterik ez dituztenez partekatu, taulan beste taldeen testeko emaitzak agertzen dira soilik. Emaitzei erreparatuz, guk garatutako hiru sistemek beste partaideen sistemek baino emaitza hobeagoak lortzen dituzte, hain zuzen ere, F1 neurrian gutxienez 3,4ko tartea lortzen dute. Emaitza hauen ondorioz pentsa dezakegu oso konplexuak diren arkitekturak ez dutela zertan sinpleagoak diren arkitekturak baino hobeagoak izan beharrik.

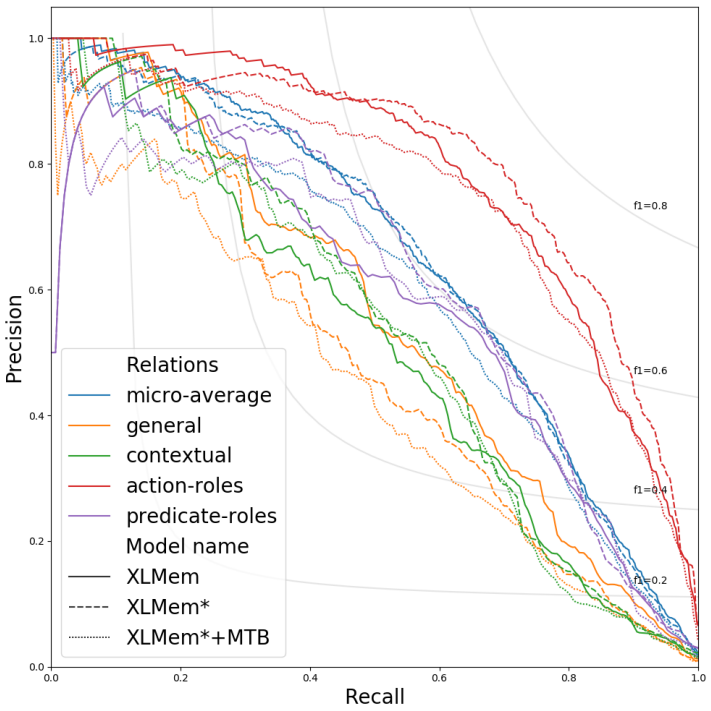
Gure sistemen emaitzetan soilik erreparatuz bi dira ikus daitezkeen gako argienak garapen datu-multzoan: entrenamendurako automatikoki anotatutako datu gehigarriak (XLM_{EM}*-ren eta XLM_{EM}*+MTB-ren kasua) erregularizazio gisa balio izan dute eta MTB aurrenentramendua sistemaren doitasuna igotzen du estaldura galduz. Test datu-multzoari erreparatuta portaera desberdina erakusten dute sistemek, MTB entrenamenduaren ondorioz lortutako doitasun igoera mantentzen den arren garapenean izan den sistema onena testean okerrena bihurtu da eta alderantziz. Portaera aldaketa hau ulertzeko eta sakonago zer gertatzen den ikusteko doitasun-estaldura kurbak erakusten ditugu 18. irudian. Ber-



Irudia 16: UH-MAJA-KD taldeak proposatutako sistema. Irudia beraien artikulutik dago aterata.



Irudia 17: Garapen eta testeko datumultzoen erlazio distribuzioa.



Irudia 18: Sistema desberdinen portaera aztertzen duten doitasun/estaldura kurbak.

	Entrenamendua			Garapena			Testa		
Eredua	Doitasuna	Estaldura	F1	Doitasuna	Estaldura	F1	Doitasuna	Estaldura	F1
XLM_{EM}	**0.861**	**0.849**	**0.855**	0.708	0.642	0.674	0.690	0.625	0.656
XLM_{EM}*	0.767	0.795	0.781	0.707	0.672	0.689	0.649	0.619	0.633
XLM_{EM}*+MTB	0.788	0.709	0.746	0.755	0.616	0.678	**0.707**	0.584	0.640
$XLMR_{large\text{-}EM}$*	0.809	0.743	0.775	0.741	**0.685**	**0.712**	0.699	**0.640**	**0.668**
$XLMR_{base\text{-}EM}$*	0.817	0.764	0.789	0.756	0.645	0.696	0.698	0.601	0.646
$XLMR_{base\text{-}EM}$*+MLM	0.851	0.739	0.791	**0.757**	0.583	0.658	0.693	0.538	0.606

Taula 5: Egin diren hobekuntzen eta esplorazioen emaitzen taula. Berriz ere * adierazten du entrenamendurako datu-multzo zaratatsu gehigarria erabili dela.

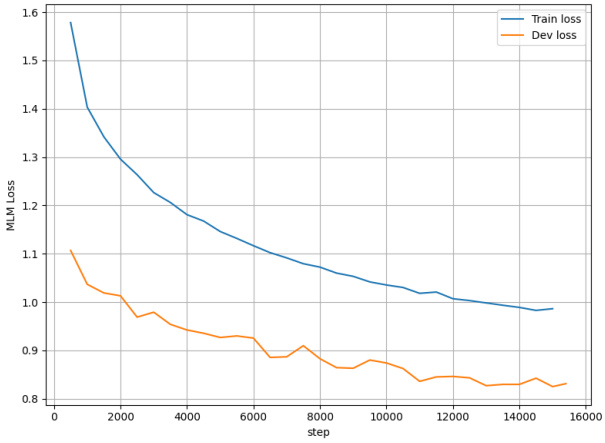
tan sistemen eta erlazio moten arabera multzokatuta erakusten ditugu doitasun-estaldura kurbak. Ikus daiteke nola erlazio mota desberdinen arabera sistema batek hobeto egiten duela besteak baino, pentsa dezakegu beraz automatikoki anotatutako datuak erlazio mota zehatz batzuei hobeto egiten diela besteei baino. Hala ere, horrek ez du guztiz azaltzen bi datu-multzoen arteko desberdintasuna, horretarako eta aurreko ideian oinarrituz, bi datu-multzoen erlazio mota distribuzioa erakusten dugu 17. irudian. Nabaria da bi datu-multzoen distribuzioen arteko aldea badagoela, eta ondorioz, datu-multzo batean espero ditugun emaitzak ezin direla guztiz bestean itxaron. Aipatu beharra dago ere, test datu-multzoak soilik 100 esaldiz osatuta dagoela, eta beraz, aldaketa txiki batek emaitzetan aldaketa nabarmena suposa dezakeela.

6.2 Sistemaren hobekuntza

Aurreko emaitzak hobetzeko helburuarekin fase honetan hizkuntza-eredu hobeago bat lortzen saiatu gara. Horretarako eta 5.3. atalean azaldu dugun bezala hasiera batean dagoeneko entrenatutako eredu gehiago esploratu ditugu. Esplorazio hori XLM eta XLMR familiako ereduetara mugatzea erabaki dugu, preseski aurreko fasean baino hiperparametro esplorazio sakonago bat ere egin nahi genuelako.

Lortutako sistema bakoitzeko emaitza onenak erakusten ditu 5. taula. Lehendabiziko konparaketa interesgarria $XLMR_{large}$ eta $XLMR_{base}$ artean dago, hain zuzen ere lehendabizikoak bigarrenari F1eko 1.6 puntu garapenean eta 2.2 puntu testean besterik ez dizkio ateratzen parametro kopurua bikoiztu arren. Hala ere, ez da konparaketa zuzen bat, izan ere $XLMR_{large}$ SGD optimizatzailea erabilita dago doituta eta $XLMR_{base}$ berriz AdamW erabiliz, Adam optimizatzailearen aldaera bat. Honen arrazoi nagusia Adam motako optimizatzaileek behar duten memoria extra izan da, eta beraz, $XLMR_{large}$-ekin erabili ezin izan dugun arren $XLMR_{base}$-ek garapeneko F1ean 7ko hobekuntza lortu du besterik gabe SGD optimizatzailea AdamW-rengatik aldatuz. Gure hipotesia da $XLMR_{large}$ AdamW-rekin doituz hobekuntza bat ere lortuko lukeela, txikiagoa izan arren, baina ezin izan dugu frogatu.

Gure esperimentazio plangintzan zegoen beste hobekuntza MLM birdoiketa da. Birdoiketa hau egiteko berriz $XLMR_{base}$ eredura egon gara mugatuta memoria arrazoiengatik.

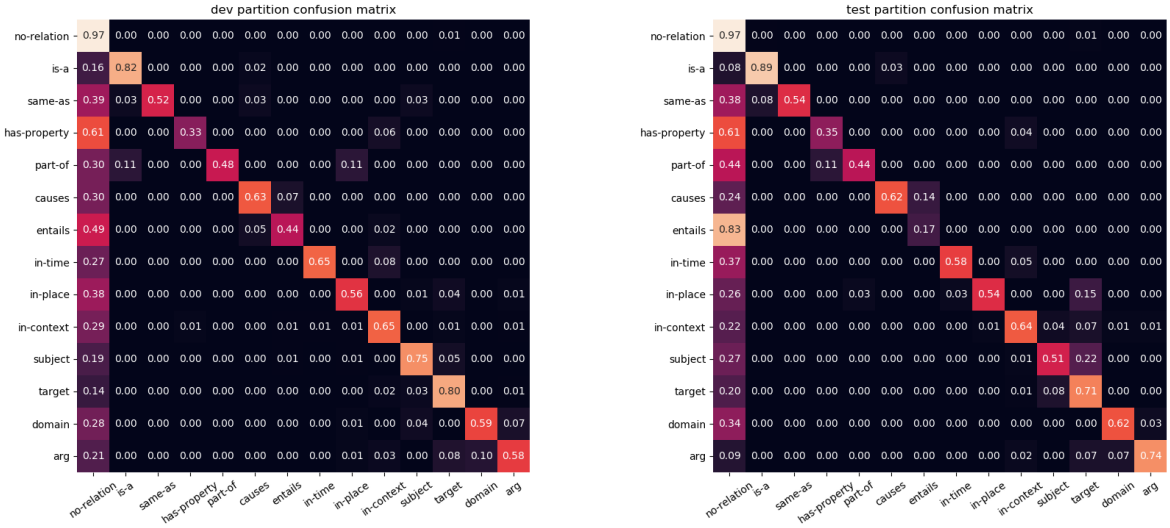


Irudia 19: $\text{XLMR}_{\text{base}}$ eredua MLM birdoiketan zehar izan dituen galera kurbak.

Hobekuntza honen emaitza jasotzen du $\text{XLMR}_{\text{base-EM}}$*+MLM eredua 5. taulan. Ikusi dezakegu orokorrean MLM birdoiketaren ondorioz errendimendua nahiko jaisten dela, zehazki orain arte lortutako sistemarik okerrena bihurtuta. Ez hori bakarrik, ikus dezakegu nola MTB birdoikuntza bezala honek ere doitasuna igotzeko joera erakusten duela. Portaera honen ondorioz sortu zaigun hipotesietako bat da eHealth-KD 2020 atazeko testuek ez dutela medikuntzako terminologia espezializatua bakarrik erabiltzen baizik eta domeinuko orokorreko hizkuntzatik gertuago dagoela. Hipotesia baieztatzeko 19. irudian erakusten dugu MLM birdoiketan ereduak lortu duen galera kurba, bai entrenamendu corpusean (MCB) bai garapeneko corpusean (eHealth-KD 2020 entrenamendurako testuak). Alde batetik irudian ikus daiteke nola hasieratik garapeneko testuetan egindako errorea entrenamendukoan baino txikiagoa mantentzen dela, honek pentsarazten digu gure hipotesia zuzena dela. Bestetik, entrenamenduko kurbarekin batera garapeneko kurba ere jaisten joan dela, eta beraz, domeinuko terminologia nolabait ikasten dabilela ematen du. Hala ere gero erlazio-erauzketa atazara eramaterakoan ezer ikasi baino gauzak ahaztu dituela dirudi.

6.3 Errore analisia

Azkenik, gure sistema onenak, $\text{XLMR}_{\text{large}}$, egindako erroreak aztertuko ditugu. Horretarako 20. irudian agertzen diren bi konfusio matrizeetan oinarrituko gara, hauek lerroka normalizatu ditugu diagonalean estaldura erakus dezaten. Ezkerrean garapenekoa eta eskuinean testekoa ikus dezakegu antza handia daukatela. Bi matrizeek erakusten duten konfusio nabarmenena `no-relation` eta klase positiboen artean dago. Hain zuzen ere, klase positiboekin batera klase negatiboa ere kontutan hartzen badugu sistemak **%93.5**-eko asmatze-tasa lortzen du. Horren arrazoi nagusia klase negatiboa eta positiboen arteko desoreka da, adibidez, garapeneko datu-multzoko %85.6 klase negatiboko adibideak dira.



Irudia 20: $\text{XLMR}_{\text{large}}$ sistemaren garapen eta testeko datu-multzoen gaineko konfusio matrizeak.

Klase positiboen arteko konfusioa aztertzen badugu ikus dezakegu garapeneko datu-multzoan sistemak nahiko ondo egiten duela. Aipatzeagatik, `in-place` eta `part-of` edo `is-a` eta `part-of` erlazioen artean dago konfusiorik handiena. Testeko datu-multzoan berriz nahasmena nabariagoa da, seguruenik adibide kopuru txikiagoa eta erlazio moten banaketa deberdina delako. Bertan ikus dezakegu adibidez, `causes` eta `entails`, `in-place` eta `target` edo `subject` eta `target` nahasmena handiagoa dela, beste batzuen artean. Bukatzeko, aipatzeko dago diagonalean lortutako balioak, hau da, klase bakoitzeko estaldura zuzenki proportzionala dela entrenamendu datu-multzoko erlazio-distribuzioarekiko.

7 Ondorioak eta etorkizuneko lana

Proiektu honetan planteatutako helburuak bete dira, alde batetik artearen egoerako erlazio-erauzketa orokorreko sistema bat berinplementatu egin da. Gero, garatutako sistema hori domeinu zehatz batera egokitzea lortu dugu eHealth-KD 2020 ataza partekatuan emaitza onak lortuz. Azkenik, datu-gehikuntza eta ezagutza-transferentziko teknikak aztertu dira sistemaren errendimendua hobetzeko asmoz. Hurrengo azpiataletan deskribatzen dugu betetako helburuen ondorioz lortutako ondorio nagusiak eta ekarpenak, baita etorkizunerako planteatutako jarraipen lerro batzuk.

7.1 Ondorio nagusiak

Lortutako emaitzei erreparatuta hauek dira lortu ditugun ondorio nagusiak:

1. Bai TACRED datu-multzoan baita eHealth-KD 2020 datu-multzoa ere entitate marketan oinarritutako sistemek emaitzak onenak lortzen dituzte, beraz erlazioen errepresentazioa lortzeko estrategia egokia dela ondorioztatzen dugu.

2. Geroz eta testu gehiago erabili aurrentrenamenduan (eta denbora luzez) bukaerako atazeko emaitzak hobeagoak dira. Hori dela eta XLMR-ek lortutako emaitzak XLM originala baino hobeagoak dira, baita $\text{XLMR}_{\text{base}}$, XLM originalaren parametro kopuru erdia edukita, garapeneko datu-multzoan ere.

3. MLM edo MTB birdoiketa helburuko testuak baino hizkuntza teknikoagoa duten testuekin egitea onura baino kalte egin dezake, hain zuzen ere doitasuna igo arren, estaldura dezente jaisten dutelako, batez ere MLM egitean. Tamaina txikiko corpora erabiltzea ere ez du lagundu.

4. Erlazio orokorren multzoko erlazioak, `is-a` izan ezik, testuinguruarekiko menpeagoak diren erlazioak baino zailagoak dira iragartzeko. Honen arrazoiak erlazioen maiztasuna edo erlazio horiek iragartzeko beharrezko ezagutza izan daitezke.

7.2 Ekarpenak

Lan honekin batera hiru izan dira egindako ekarpen nagusiak. Lehendabizikoa artearen egoerako erlazio-erauzketarekin du zerikusia, hain zuzen ere, BERT_{EM} sistemaren ber-inplementazio bat izan da. Sistema hau Baldini Soares et al. (2019)-ek proposatuta oraindik ez du inplementazio ofizialik, hori izan da lan honetan sistema ber-inplementatzeko arrazoi nagusia. Guk egindako sistemaren emaitzak eta beraiek erreportatutako emaitzak bat etorri ez arren nahiko gertu daudela ikusi dugu, eta seguruenik hardware egokia edukiz gero berdintsuagoak izatea lortuko genuke.

Bigarrena ezagutza-transferentziarekin eta datu-gehikuntzarekin du zerikusia. Zehatzago hizkuntzen arteko ezagutza-transferentzia eta domeinu zehatzetara doitzeko teknikaten egin dugun analisia izan da. Lan honetan momentuko artearen egoerako hizkuntza-

eredu eleanitzen azterketa bat egin dugu. MLM birdoiketaren bitartez domeinu orokorreko hizkuntza-eredu bat domeinu zehatz batera doitu dugu. Eredu horren emaitzak onak izan ez diren arren interesgarriak izan dira. Eta, azkenik MTB aurrentrenamendu erdigainbegiratua aplikatu dugu domeinu zehatz batean, sistemen doitasuna hobetuz.

Azkenik, egindako lan guztiaren ondorioz lortutako medikuntza domeinuko erlazio-erauzle sistema bera da egindako ekarpen nagusia. Ataza partekatuaren irabazle izan den sistema eta eduki dituen hobekuntza guztiak. Sistemaren kodea GitHub-en[15] egongo da publikoki eta entrenatutako ereduak eskaera pean.

7.3 Etorkizuneko lana

Proiektu honetan egindako lana emaitza onak eman dituela esan daiteke, hain zuzen ere eHealth-KD 2020 ataza-partekatuaren erlazio-erauzketako azpiatazaren irabazle izan den sistema garatua izan delako. Baina hori egia izan arren, oraindik badaude aipatutako puntu asko ikerketa sakonago bat behar dutenak. Horien artean hauek dira gure iritziz etorkizunean ikerketzeko interesagarri izan daitezkeen jarraipen lerro batzuk:

1. **Datu-multzo kopurua handitzea**

   Garatutako sistemak emaitza onak eman ditu eHealth-KD datu-multzoan, baina, aurretik aipatu dugun bezala datu-multzo hori nahiko txikia da. Normalean, sistema berri bat garatzean ohikoa da datu-multzo bat baino gehiagotan ebaluatzea, benetan beste sistemak baino hobeagoa den ziurtatzeko. Baita ere, sistemaren sendotasuna ebaluatu ahal izateko komeni da beste datu-multzo batzuetan probatzea. Horretarako hasiera batean beste urteetako eHealth-KD txapelketen datu-multzoak erabiltzea pentsatu dugu.

2. **Hizkuntza-eredu elebakarren azterketa**

   Lan honetan zuzenean eredu eleanitzekin egin dugu lan, elebakarrak diren eredu horiek kontutan hartu gabe. Egin daitekeen azterketa interesgarri bat da beste hizkuntzetatik transferitutako ezagutza benetan baliagarria den edo ez ikustea. Intuizioak esaten digu helburu hizkuntza horretan soilik entrenatua izan den hizkuntza-eredu batek hizkuntza-eredu eleanitz bat baino hobeto egin behar lukeela baldin eta datu kopuru handitan entrenatu bada. Adibidez, aurten aurkeztutako BERTEUS (Agerri et al., 2020) euskarazko MLM hizkuntza-eredua mBERT (Devlin et al., 2019) hizkuntza-eredu eleanitza baino emaitza hobeagoak lortzen ditu hainbat atazatan. Proiektu honetan gaztelerazko hizkuntza-eredu elebakarrik ez erabiltzearen arrazoi nagusia proiektuaren hasieran hauek oraindik publikatuta ez egotea izan da. Baina orain dela gutxi BETO (Cañete et al., 2020) aurkeztu dute eta interesagarria izan daiteke gure sisteman probak egitea.

3. **Sistemaren eleaniztasuna ebaluatzea**

[15] `www.github.com/osainz59/XLREMed`

Eredu eleanitzak erabiltzen hari garen arren gazteleran besterik ez gaude ebaluatzen. Beste hizkuntza bateko testu bat emanda ze errendimendu emango lukeen aztertzea oso ataza interesgarria izan daiteke. Ideia honen atzean badira proiektu garrantzitsuak, BETTER (Better Extraction from Text Towards Enhanced Retrieval, IARPA-BAA-18-05) adibidez, non ingelesezko datuetatik bakarrik ikasita beste hizkuntzetara gertaera-erauzketako ezagutza transferitzea du helburu. Baina ebaluaketa posible izateko pare bat hizkuntzatan anotatutako lagin txiki bat anotatzea beharrezkoa da, hori dela eta proiektu honen helburuetatik kanpo gelditu da.

4. **Datu gehikuntzarako tekniken azterketa eta hobekuntza**

   Lan honetan datu gehikuntzarako teknikak aplikatu diren arren ez dute emaitzetan eragin on handirik izan. MLM birdoiketaren kasuan batez ere onura baino kalte gehiago egin duelako. Hala eta guztiz ere jasotako emaitza hauek hausnarketa eta konklusio interesgarri batzuetara eraman gaituzte. Horietako bat erabilitako corpusekin du zerikusia, datu gehikuntzarako erabili ditugun corpusak nahiko txikiak izan baidira, batez ere MTB birdoiketa egiteko erabili dugun MA corpusa. Egin beharreko lan bat MCB corpusa handitzea eta modu automatiko batez entitateak anotatzea da, gero MTB birdoiketa corpus horren gainean egin ahal izateko. Bestetik MTBek beste kontraste bidezko ikasketa (*Contrastive Learning* ingelesez) metodo askok bezala bere arazoak ditu, horregatik Caron et al. (2020) autoreek aurkeztutako konparaketa bidezko multzokapena (*Contrastive Clustering* ingelesez) MTBrekin batera aplikatzea izan daiteke beste ikerketa lerro bat. Azkenik, beste metodo tradizionalagoak probatzea ere aukera dago, urruneko gainbegiraketa adibidez.

Erreferentziak

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, eta Eneko Agirre. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 2020.

Alan Akbik, Duncan Blythe, eta Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

Alan Akbik, Tanja Bergmann, eta Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1078. URL `https://www.aclweb.org/anthology/N19-1078`.

Alan Akbik, Tanja Bergmann, eta Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 724–728, 2019b.

Christoph Alt, Marc Hübner, eta Leonhard Hennig. Improving relation extraction by pre-trained language representations. In *Proceedings of AKBC 2019*, 2019. URL `https://openreview.net/forum?id=BJgrxbqp67`.

Edgar Andrés, Oscar Sainz, Aitziber Atutxa, eta Oier Lopez de Lacalle. IXA-NER-RE at eHealth-KD Challenge 2020: Cross-Lingual Transfer Learning for Medical Relation Extraction. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Mikel Artetxe, Gorka Labaka, eta Eneko Agirre. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*, 2020.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, eta Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL `https://www.aclweb.org/anthology/P19-1279`.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL `https://arxiv.org/pdf/2005.14165.pdf`.

HAP masterra

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, eta Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, eta Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*, 2020.

Nancy A. Chinchor. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998. URL https://www.aclweb.org/anthology/M98-1001.

Laura Chiticariu, Yunyao Li, eta Frederick R. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1079.

Kevin Clark, Urvashi Khandelwal, Omer Levy, eta Christopher D. Manning. What does bert look at? an analysis of bert's attention. *ArXiv*, abs/1906.04341, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, eta Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, eta Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, eta Jonathan Wright. Linguistic resources for 2012 knowledge base population evaluation. In *Proceedings of the Fifth Text Analysis Conference, TAC 2012, Gaithersburg, Maryland, USA, November 5-6, 2012*. NIST, 2012. URL https://tac.nist.gov/publications/2012/additional.papers/KBP2012_annotation_overview.proceedings.pdf.

Joe Ellis, Jeremy Getman, Justin Mott, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, eta Jonathan Wright. Linguistic resources for 2013 knowledge base population evaluations. In *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*. NIST, 2013. URL https://tac.nist.gov/publications/2013/additional.papers/KBP2013_annotation_overview.TAC2013.proceedings.pdf.

HAP masterra

Aitor García-Pablos, Naiara Perez, Montse Cuadros, eta Elena Zotova. Vicomtech at eHealth-KD Challenge 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, eta Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Zhijiang Guo, Yan Zhang, eta Wei Lu. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1024. URL `https://www.aclweb.org/anthology/P19-1024`.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, eta Wei Lu. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, 7:297–312, March 2019b. doi: 10.1162/tacl_a_00269. URL `https://www.aclweb.org/anthology/Q19-1019`.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, eta Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Jeremy Howard eta Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL `https://www.aclweb.org/anthology/P18-1031`.

Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez De Lacalle, eta Eneko Agirre. Removing noisy mentions for distant supervision. *Procesamiento del lenguaje natural*, (51):41–48, 2013.

H. Ji, Joel Nothman, eta Ben Hachey. Overview of tac-kbp 2014 entity discovery and linking tasks. 2015a.

Heng Ji eta Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P11-1115`.

Heng Ji eta Joel Nothman. Overview of TAC-KBP2016 tri-lingual EDL and its impact on end-to-end KBP. In *Proceedings of the 2016 Text Analysis Conference, TAC 2016, Gaithersburg, Maryland, USA, November 14-15, 2016*.

NIST, 2016. URL https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf.

Heng Ji, Joel Nothman, Ben Hachey, eta Radu Florian. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST, 2015b. URL https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_Trilingual_Entity_Discovery_and_Linking_overview.proceedings.pdf.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, eta Cash Costello. Overview of TAC-KBP2017 13 languages entity discovery and linking. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST, 2017. URL https://tac.nist.gov/publications/2017/additional.papers/TAC2017.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf.

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

Diederik P Kingma eta Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

ThomasÑ. Kipf eta Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/pdf/1609.02907.pdf.

John Lafferty, Andrew McCallum, eta Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. URL https://repository.upenn.edu/cis_papers/159/.

Guillaume Lample eta Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, eta Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Xuansong Li, Joe Ellis, Kira Griffitt, Stephanie M. Strassel, Robert Parker, eta Jonathan Wright. Linguistic resources for 2011 knowledge base population evaluation. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST, 2011. URL https://tac.nist.gov/publications/2011/additional.papers/KBP2011_annotation_overview.proceedings.pdf.

HAP masterra

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, eta Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018. URL https://arxiv.org/pdf/1801.10198.pdf.

Pilar López-Ubeda, José M. Perea-Ortega, Díaz-Galian Manuel C., M. Teresa Martín-Valdivia, eta L. Alfonso Ureña-López. SINAI at eHealth-KD Challenge 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Ilya Loshchilov eta Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

Diego Marcheggiani eta Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1159. URL https://www.aclweb.org/anthology/D17-1159.

Tomas Mikolov, Kai Chen, G. S. Corrado, eta J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Mike Mintz, Steven Bills, Rion Snow, eta Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P09-1113.

Jeffrey Pennington, Richard Socher, eta Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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

Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, eta Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, eta Andrés Montoyo. Overview of the ehealth knowledge discovery challenge at iberlef 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145 – 151, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00116-6. URL http://www.sciencedirect.com/science/article/pii/S0893608098001166.

HAP masterra

Alec Radford, Karthik Narasimhan, Tim Salimans, eta Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, eta Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alejandro Rodríguez Pérez, Ernesto Quevedo Caballero, Jorge Mederos Alvarado, Rocío Cruz-Linares, eta Juan Pablo Consuegra-Ayala. UH-MAJA-KD at eHealth-KD Challenge 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Oscar Sainz, Oier Lopez de Lacalle, Itziar Aldabe, eta Montse Maritxalar. Domain adapted distant supervision for pedagogically motivated relation extraction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2213–2222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.270.

Connor Shorten eta Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Ian Tenney, Dipanjan Das, eta Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Association for Computational Linguistics*, 2019. URL https://arxiv.org/abs/1905.05950.

Jörg Tiedemann eta Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ Gomez, Ł ukasz Kaiser, eta Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, eta R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Rebecka Weegar, Alicia Pérez, Arantza Casillas, eta Maite Oronoz. Recent advances in swedish and spanish medical entity recognition in clinical texts

using deep neural approaches. *BMC Medical Informatics and Decision Making*, 2020. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0981-y.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, eta Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Wenpeng Yin, Jamaal Hay, eta Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL https://www.aclweb.org/anthology/D19-1404.

Bowen Yu, Zhenyu Zhang, Tingwen Liu, Bin Wang, Sujian Li, eta Quangang Li. Beyond word attention: Using segment attention in neural relation extraction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5401–5407. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/750. URL https://doi.org/10.24963/ijcai.2019/750.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, eta Bowen Zhou. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1053. URL https://www.aclweb.org/anthology/P17-1053.

Daojian Zeng, Kang Liu, Yubo Chen, eta Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1203. URL https://www.aclweb.org/anthology/D15-1203.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, eta Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.