IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

IV. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2021eko ekainaren 9, 10 eta 11a
Gasteiz, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK ETA ARTEA

Erlazio-erauzketa testu klinikoetan hizkuntzaren prozesamenduaren bidez

Sergio Santana, Alicia Pérez, Arantza Casillas eta *Maite Oronoz*

59-66 or.
https://dx.doi.org/10.26876/ikergazte.01.07

ANTOLATZAILEA: LAGUNTZAILEAK:
UEU udako euskal unibertsitatea
eman ta zabal zazu UPV EHU
upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Universidad de Navarra
Deusto Universidad de Deusto Deustuko Unibertsitatea
Mondragon Unibertsitatea
Ayuntamiento de Vitoria-Gasteiz Vitoria-Gasteizko Udala
araba álava foru aldundia diputación foral
EUSKO JAURLARITZA GOBIERNO VASCO HEZKUNTZA SAILA DEPARTAMENTO DE EDUCACIÓN
UDALBILTZA

Erlazio-erauzketa testu klinikoetan hizkuntzaren prozesamenduaren bidez

Santana, S., Pérez, A., Casillas, A., Oronoz, M.

Ixa ikerketa taldea (UPV/EHU). HiTZ: Basque Center for Language Technology. M. Lardizabal 1. 20080 Donostia.

ssantana005@ikasle.ehu.eus

Laburpena

Testu klinikoetan informazio aberatsa dago, besteak beste, medikuntzako entitate izendunak ditugu (botika-izenak, gaixotasun-izenak, etab.) eta hauen arteko erlazioak. Informazio hori eraueteko, ikasketa sakoneko algoritmoak hartu ditugu oinarri eta antzeko patroiak, ez derrigor berdinak, topatzeko gai den sistema bat eraiki dugu. Sistemak bi entitate (adb. botika bat eta gaixotasun bat) erlazionatuta dauden ala ez detektatu eta erlazio mota (adb. kausa) esleitu behar du. Horretarako, Joint AB-LSTM deritzon algoritmoaren ahuleziak aztertu eta desoreka estatistikoari aurre egiteko bi modu proposatu ditugu. Hurbilpena nazioarte mailan ezaguna den ingelesez idatzitako BioNLP 2011 datu-sortan ebaluatu dugu.

Hitz gakoak: sare neuronal sakonak, joint AB-LSTM, hizkuntzaren prozesamendua

Abstract

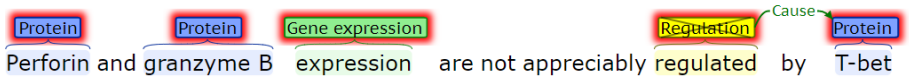
This work deals with relation extraction in clinical text mining. To this end we turn to algorithms based on deep learning: given examples from a set of texts, the aim is to build a system able to find similar though not necessarily equal patterns. In this work we analyse systems that automatically extract information from clinical texts. Given that the context is crucial in this task, we resorted to Joint AB-LSTM algorithm. The system has to detect if two entities (a medicine and a disease) are related, and if they are, assign a class (for example cause). In this work we explore the weak points of the algorithm and we propose two ways to tackle statistical inbalance. Our proposal was assessed in a well-known English medical data-set, i.e. BioNLP 2011.

Keywords: *deep neural networks, joint AB-LSTM, natural language processing*

1 Sarrera

Lan honek hizkuntza idatzia aztertu eta informazioa eraueteko sistema aurkezten du. Hain zuzen ere, medikuntza arloko testuetan entitateen arteko erlazioak eraueteko sistema. Erlazio kliniko baten adibide bat erakusten da 1. irudian Brat deritzon bistaratzea erabiliz (Stenetorp *et al.*, 2012). Adibidean, erlazioa geziaren bidez adierazi da eta erlazio-mota geziaren gainean idatzi da. Formalki, erlazioa $\mathcal{R}(e_i, e_j, text) = relType$ eran definitzen da, non e_i eta e_j erlazioan parte hartzen duten "entitateak" diren, hurrenez-hurren jatorria eta helmuga. Beste aldetik, "text" aldagaiak, entitateen testuingurua jasotzen du. Adibide baterako, 1. irudian bost entitate agertzen dira: *Perforin*, *granzyme B* eta *T-bet* PROTEIN motakoak; *expression* GENE EXPRESSION motakoa eta *regulated* REGULATION motako entitatea. Horretaz gain, bi entitateen arteko *Cause* motako erlazioa dago. $text$ aldagaian irudiko testu osoa legoke eta *Cause* motako erlazio bat agertzen da: $\mathcal{R}(regulated, T-bet, text) = Cause$. Iruditik, inplizituki, gainontzeko entitateak erlazionatuta ez daudela ulertzen da, formalki: $\mathcal{R}(regulated, expression, text) = NO\_RELATION$.

1. Irudia: Erlazio kliniko baten adibidea BioNLP 2011 GE azpi-atazan. Entitateak kolorez azpimarratuta agertzen dira testuan eta entitateei dagokien mota gainean idatzi da (adb. *T-bet* delakoa "Protein" motako entitatea da). Entitateen arteko erlazioak gezien bidez adierazten dira eta gezian erlazio mota adierazten da).



Erlazio-erauzketa prozesuan, lehenik eta behin, entitateak ezagutu behar dira. Adibidez, 1 irudian *Perforin*, *granzyme B* eta *beta-catenin* PROTEIN motako entitate gisa ezagutu dira. Entitate klinikoak erauzteko metodo automatikoek kalitate altua lortu ohi dute, 0.93ko F1 neurria lortu izan da azken urteotan (Goyal *et al.*, 2018; Yadav eta Bethard, 2019; Li *et al.*, 2020). Ez dago, ordea, lan honetako helburuen artean entitateak ezagutzea, izan ere, entitateak anotatuta dituzten datu-sortak erabiliko ditugu. Aitzitik, **helburua**, entitateen arteko erlazioak automatikoki erauztea da. Osasunaren arloan oso garrantzitsua da erlazioak ongi detektatzea. Adibidez, botika bat hartzearen ondorioz sortutako sintomak edo gaixotasunak detektatzean, drogen aurkako erreakzioak detektatzean, botiken eta gaixotasunen arteko erlazioak finkatu behar dira. Ezinbestekoa da, lotura ongi egitea, ezin baitugu ez botika gaizki identifikatu, ez botika hori hartzearekin zerikusirik ez duten sintomak lotu. Erlazio erauzketa automatikoa egiteko, adimen artifizialaren ildotik joko dugu. Hain zuzen ere, hizkuntzaren prozesamenduaren eta ulermen alorrean kokatzen da gure lana.

Azken hamarkadan, erlazio-erauzketan erabili izan diren hurbilpenak aztertuz, bilakaera nabarmena sumatzen da. Hasierako sistemek testu-ingururik gabeko errepresentazio dentsoak (*embedding* deritzonak) erabiltzen zituzten. Izan ere, joera entitate-bikoteak bektore-espazio horretan proiektatu eta entitate-bikoteen arteko distantzien arabera aurreikusten zen erlazionatuta zeuden ala ez (Bordes *et al.*, 2013; Kang *et al.*, 2014). Distantziak ez ezik, entitate-bikoteek osatzen zuten bektorearen norantza ere esanguratsua zela ohartarazi zen (Pérez *et al.*, 2016). Testuinguru honetan, sintaxiak paper garrantzitsua jokatzen duela frogatu zen (Li *et al.*, 2016). Testuen errepresentazioa lortzeko, eskuz aukeratutako ezaugarriak erabiltzetik, automatikoki erauzitako ezaugarriak igaro da (Nguyen eta Grishman, 2015). Aurrerago, testuingurua kontuan hartzen dituzten sare neuronalak gailendu dira (He *et al.*, 2018). Sare neuronal konboluzionalez gain *Long short-term memory (LSTM)* sare neuronaletan oinarritutako hurbilpenak erabili dira, adibidez Sahu eta Anand, 2018 aipamena, gure lanaren abiapuntua. Testuingurua erabakiorra izaten denez erlazio erauzketan, gure kasuan Joint AB-LSTM (Sahu eta Anand, 2018) hurbilpena aztertu dugu, izan ere, bi LSTM konbinatzen baititu. Sistemak testuingurua kudeatzeko gai bada ere, ahuleziak ditu, hurrengo atalean azalduko dugun bezala, eta horiei aurre egiteko metodo bat planteatu dugu.

2 Metodologia

Erlazioak erauzteko eredua lortzeko algoritmoak datu-sorta gainbegiratuak darabiltza, alegia, erlazioak markatuta dituen datu-sorta edo urre-patroia. Aldez aurretik erlazioak etiketatuta dituen testu-bildumatik, ikasketa automatikoan oinarritutako algoritmoen laguntzaz, testuak etiketatzeko gai den eredua sortzea da gure ataza. Algoritmoak, eredua entrenatu ahal izateko, datu-sortan kasu positiboak (erlazionatutako entitate-bikoteak) eta negatiboak (erlaziorik gabeko entitate-bikoteak) behar ditu. Orokorrean, erlazioak erauzteko eredu iragarlearen inferentziak dakarren **erronka** nabarmenena, klaseak har ditzakeen balioen arteko desoreka da. Potentzialki erlazionatuta egon litezkeen entitate-bikote kopurua nabarmenki handia da errealitatean erlazionatuta daudenekin alderatuta. Desorekaren ondorioz, ikasketa automatikoan oinarritutako algoritmoen joera, erlazio eza iragartzea da. Beste hitzetan, sistema maizen agertzen den klasera alboratuta egoten da. Proposatu dugun metodologiak, erronka honi aurre egiteko ekarpenak dakartza.

2.1 Kasu negatiboen sorkuntza

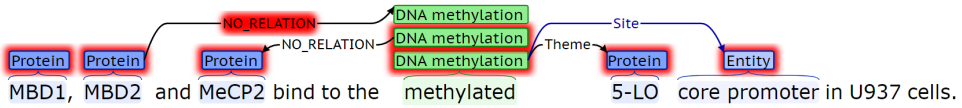
Berez, urre-patroian kasu positiboak baino ez daude etiketatuta. Entrenamendurako negatiboak diren erlazioak automatikoki sortu dira, baina ez denak. Erlazionatuta ez dauden guztiak sortu izan balira, positibo eta negatiboen arteko desoreka nabarmena eragingo genuke. Kasu negatiboak sortu behar dira neurriz, sistemak ez dezan soilik distribuzioa ikas. Garrantzitsuena, positibo eta negatibo posibleen artean bereizten ikastea da. Horretarako, gakoa da algoritmoari antzekoak edo nahasgarriak izan dakizkiokeen adibideak (positiboak eta negatiboak) aurkeztea da.

Kasu negatiboak sortzeko bi murriztapen ipini ditugu: batetik, esaldi bereko entitateak lotu dira eta ez esaldien artekoak; bestetik, nahasgarriak izan daitezkeen erlazio negatiboak baino ez dira kontsideratu. Erlazio bat sortzea ahalbidetu da baldin eta entrenamendu multzoan entitate-bikote mota bera erlazionatuta agertu bada. Honela, erlazionatuta egon litezkeen entitate-bikoteak aurkeztuko zaizkio sistemari, batzuk positiboak eta besteak negatiboak. Ezberdintasuna ez datza entitate-bikotean, testuinguruan baizik. Honek ulermen gaitasuna areagotzea du helburu.

Murriztapen hauek argitzearren adibide bat ipini dugu 2. irudian. Irudiko erlazio guztiak esaldi berean daude anotatuta (1. murriztapena) eta NO_RELATION agertzen da, gorriz markatua, erlazionatuta egon litezkeen bikoteen artean alegia, *Protein* eta *DNA methylation* motako entitateen artean (2. murriztapena). Datu-sortan $\mathcal{R}(methylated, 5 - LO, text) = Theme$ erlazioa existitzen denez, $\mathcal{R}(methylated, MeCP2, text) = NO\_RELATION$ erlazioa sor daiteke (2. murriztapena). Aitzitik, $\mathcal{R}(MDB2, methylated, text) =$

NO_RELATION erlazioa, gorriz dagoena, ez da sortuko, datu-sorta osoan ez baitago *Protein* eta DNA *methylation* lotzen duen adibiderik.

2. Irudia: BioNLP 2011 corpuseko dokumentu batean sor daitekeen NO_RELATION erlazio hautagaietako bat (fondo zuriz markatua), eta NO_RELATION hautagai bat ezin daitekeena sor (gorriz markatua).



2.2 Ikasketa sakoneko hurbilpena

Aitzindariak aztertuta, Joint AB-LSTM (Sahu eta Anand, 2018) aipamena erabiltzea erabaki dugu, alde batetik, testuingurua kontuan hartzeko duen gaitasunagatik eta beste aldetik oso alboratutako sailkapen bitarrean (bai / ez klaseekin) sendoa izan zelako (Santiso *et al.*, 2018). Gure lanaren ekarpenetako bat sistema hori klase anitzeko datuetara hedatzea izan da.

Joint AB-LSTM sareak hizkuntzaren prozesamenduan aurrerapen handia ekarri dituen Bi-LSTM motako (Hochreiter eta Schmidhuber, 1997) bi sare uztartzen ditu era ezberdinetan parametrizatuta. Bi-LSTM sare neuronalen ezaugarririk aipagarriena testuingurua bereizteko duten gaitasuna da, hitzari, testuinguruaren araberako ezaugarri numerikoak esleitzen baitizkio.

Izanak izan, ez dugu *Joint AB-LSTM* (Sahu eta Anand, 2018) bere horretan erabili, jatorrizko inplementazioak ahulezia batzuk baititu. Alde batetik, ez darabil entitateen hedadura (alegia, nondik-nora doan entitatea) eta, ondorioz, erauzitako erlazioak ez dira itxaron litezkeen erlazioen baliokideak hedadurari dagokionean. Beste aldetik, jatorrizko inplementazioak *F1 mikro-batezbestekoa* (ingelesez *micro-averaged F1)* neurria darabil sailkatzailea entrenatzeko optimizazio prozesuetan. Erlazio guztiekiko duen F1 neurriaren batezbestekoa lortzeko modu honetan maizen agertzen den klasearekiko (alegia erlaziorik ezarekiko) F1 neurria gailentzen da, alegia, alborapena sustatzen du. Adibidez, sailkatzaileak edozein entitate-bikote bat emanda erlazionatuta ez daudela esateko joera sustatuko luke, maizen agertzen den erlazio mota, erlazio eza baita. Entrenamendu prozesuan, ordea, gutxiengoan dauden erlazioak doitasunez eta estaldura handiz iragartzea da helburua.

Honenbestez, gure **ekarpenak** hauexek dira: batetik, jatorrizko inplementazioa klase bitarrekin lan egiteko prestatuta zegoen eta guk klase anitzekin lan egiteko egokitu dugu; bigarren, optimizazio prozesuan, *F1 makro-batezbestekoa* (ingelesez, *macro-averaged F1)* darabilgu neurri honi klaseen alborapenak ez baitio eragiten; hirugarren, entitateen hedadura kontuan hartzen dugu esplizituki eta, honenbestez, trazabilitatea errazten da.

2.3 Desorekari aurre egiteko arkitekturak

Datu-sortan zeuden kasu positiboei kasu negatibo sintetikoak erantsi zaizkie 2.1. atalean adierazi den moduan. Murriztapenekin erantsi diren arren, oraindik desoreka nabaria da. Alegia, erlazio positiboak erlazio posibleekiko arbuiagarriak dira. Desorekaren eraginari aurre egiteko, erlazioen sailkapena ondoz-ondoko bi urratsetan egitea proposatu dugu, ohikoena urrats bakarrean egitea bada ere. Honetan datza arkitektura bakoitza:

1. Urrats batean: zuzenean, erlazio hautagai guztiak planteatu eta dagokien klasea iragarriko duen sailkatzailea sortu. Hauxe da oinarri-lerroa, alegia, klase anitzetan oinarritzen dena.
2. Bi urratsetan: lehendabizi bi entitate erlazionatuta dauden ala ez erabakitzen duen sailkatzaile bitarra dago. Honek erlazionatuta ez dauden bikoteak baztertzen laguntzen du eta horrekin klaseen alborapena txikitu egiten da. Ondoren, erlazio positiboei dagokien klasea iragarriko duen sailkatzailea dago. Hauxe da sustatu dugun hurbilpena, alegia, hurbilpen bitarra deitu duguna.

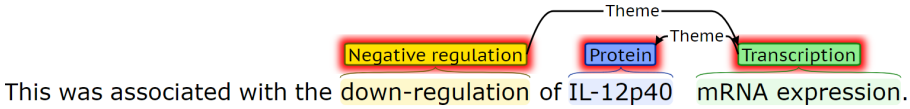
3 Esperimentazioa

3.1 Datu-sorta

Gure sistema nazioarteko erronketan erabili diren hiru atazetan ebaluatu da, hain zuzen ere BioNLP 2011 (Kim *et al.*, 2011a) datu-sortaren hiru azpi-atazekin: EPI ataza, non aldaketa epigenetikoak deskribatzen dituzten dokumentuak ematen diren (Ohta *et al.*, 2011); GE ataza, non geneetako gertaerak deskribatzen dituzten dokumentuak

ematen diren (Kim *et al.*, 2011b) eta REL ataza, non gene eta proteinen arteko interakzioak deskribatzen diren (Pyysalo *et al.*, 2011). BioNLP 2011 GE azpi-atazaren adibide bat jaso dugu 3. irudian.

3. Irudia: BioNLP 2011 GE azpi-atazako erlazioen adibidea.



Datu-sorta honen deskribapen kuantitatiboa 1. taulan ematen da eta erlazio klasearen banaketa 2. taulan. Gogoan izan behar da 1. taulan ematen diren erlazio kopuruan ez direla kontatzen NO_RELATION motakoak, hauek ez direlako corpusaren parte, sintetikoki sortutako erlazioak baizik. 2 taulan sortutako NO_RELATION erlazioak ere ematen dira, beste klaseetan dauden erlazio kopuruarekin konparatu ahal izateko.

1. Taula: BioNLP 2011 datu-sortaren deskribapen kuantitatiboa: Dokumentu kopuru osoa; Anotatutako entitateak; Erlazio kopuru osoa; Hitz kopurua multzo osoan; Batezbesteko hitz kopurua dokumentuka; Hiztegiaren tamaina.

Ataza	Partiketa	Dokumentuak	Anotazioak	Erlazioak	Hitz kopurua	Hitz kopurua dokumentuz	Hiztegiaren tamaina
EPI	Entrenamendua	600	7595	1852	127312	212	19454
	Garapena	200	2499	601	43497	217	9373
	Ebaluazioa	400	5096	1261	91929	230	15439
GE	Entrenamendua	805	11625	10310	205729	255	19482
	Garapena	155	4690	3250	64132	414	9768
	Ebaluazioa	264	4301	4487	79047	300	12035
REL	Entrenamendua	800	9297	1857	176146	220	16648
	Garapena	150	2080	480	33827	226	5934
	Ebaluazioa	260	3589	497	57256	220	9314

2. Taula: BioNLP 2011 datu-sortaren erlazio klaseekiko instantzia-banaketa erlazio negatiboak sortu eta gero

Klasea	EPI ataza		GE ataza		REL ataza	
	Entrenamendua	Garapena	Entrenamendua	Garapena	Entrenamendua	Garapena
atloc	-	-	62	8	-	-
cause	166	60	1394	605	-	-
contextgene	49	13	-	-	-	-
csite	-	-	42	13	-	-
equiv	391	153	509	113	433	85
no_relation	14800	4640	36718	15877	17588	4724
protein-component	-	-	-	-	1210	290
sidechain	60	24	-	-	-	-
site	642	215	446	203	-	-
subunit-complex	-	-	-	-	479	142
theme	1399	462	9427	3099	-	-
toloc	-	-	66	18	-	-

3.2 Aurre-prozesamendua

Erabili diren sarrerako datu-sortak Stav deritzon anotatzailearen (Ananiadou eta Tsujii, 2012) formatuan datoz. Formatu hau Joint AB-LSTM sare neuronalak eskatzen duen formatura bihurtu behar da. Jarraian, hizkuntzaren prozesamenduan ohikoa den bigarren aurre-prozesamendu bat aplikatzen zaie: sarrerako testua letra xehera bihurtzen da eta gero tokenizatu egiten da. Formatu-bihurketa bereziki eta, beste maila batean, aurre-prozesamendua, esker onekoak ez diren lanak direnez, irakurleak lan hau errepikatu ahal izateko eskuragarri[1] utzi ditugu programak bigarren mailako **ekarpen** gisa.

[1] Datuen egokitzapenerako inplementatutako softwarea eskuragarri utzi dugu hemen: `https://github.com/Porobu/HAPLAP-MAL`

3.3 Emaitza esperimentalak

Sarrerako datu hauek aurreprozesatuta daudenean Joint AB-LSTM sare neuronala entrenatu eta ebaluatzen da. Sare neuronalak erlazioak sailkatzeko duen gaitasuna ebaluatzeko erabili diren **metrikak** *doitasuna*, *estaldura* eta F1-neurria dira. Neurri hauek erlazio mota bakoitzerako lortzen dira eta ikuspegi orokorra izatearren, neurrien batezbestekoak ateratzen dira: *mikro-batezbestekoa*, *makro-batezbestekoa* eta *batezbesteko haztatua* (ingelesez, *weighted average*) Sokolova eta Lapalme (2009)-ren lanean azaltzen den moduan. Azpi-ataza bateko (REL azpiatazako) emaitzak klaseka aurkezten dira 3. taulan.

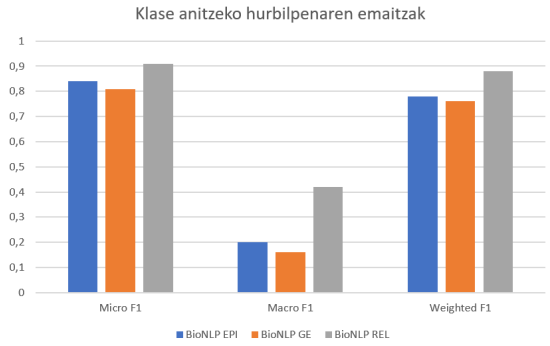
3. Taula: BioNLP 2011 REL datu-sortarekin lortutako emaitzak klaseka zehaztuta eta klaseekiko batezbestekoak (mikro-, makro- eta bataz-besteko haztatua)

Klasea	Urrats batean			Bi urratsetako arkitektura					
	Klase anitzeko hurbilpena			Hurbilpen bitarra 1. fasea			Hurbilpen bitarra 2. fasea		
	Doit.	Estal.	F1	Doit.	Estald.	F1	Doit.	Estald.	F1
equiv	0,66	0,45	0,53	-	-	-	0,87	0,80	0,83
erlazioa	-	-	-	0,50	0,59	0,54	-	-	-
no_relation	0,91	0,99	0,95	0,95	0,93	0,94	-	-	-
protein-component	0,50	0,10	0,17	-	-	-	0,81	0,83	0,82
subunit-complex	0,40	0,01	0,03	-	-	-	0,64	0,64	0,64
Mikro	0,91	0,91	0,91	0,90	0,90	0,90	0,79	0,79	0,79
Makro	0,62	0,39	0,42	0,73	0,76	0,74	0,77	0,76	0,77
Batez. haztatua	0,87	0,91	0,88	0,91	0,90	0,90	0,79	0,79	0,79

Laburtzearren, gainontzeko azpi-atazetan ez ditugu emaitzak klaseka aurkeztuko. Aitzitik, ikuspegi orokorraren adierazle gisa, *makro* F1 neurriari erreparatuko diogu eta azpi-ataza guztiak batera aurkeztuko ditugu era grafikoan. Urrats batean gauzatzen den arkitekturaren bidez lortutako batezbesteko emaitzak 4. irudian biltzen dira. Arkitektura honetan erlazio guztiak (positiboak eta negatiboak) batera sailkatzen dira. Bi urratsetan gauzatzen den arkitekturaren bidez lortutako batezbesteko emaitzak 5a. eta 5b. irudietan jaso dira. Lehenengo urratsean erlaziorik badagoen iragartzen da, eta bigarren urratsean erlazioei mota esleitzen zaie. Honegatik, 3. taulan, bi urratsetako arkitekturan, 1. fasean erlazioa ala erlazio-eza baino ez da iragartzen eta 2. fasean, ordea, erlazio-mota zehazten da.

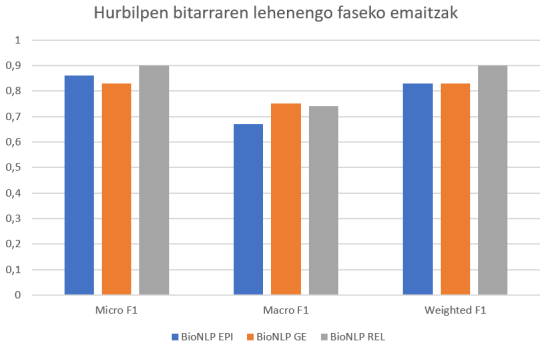
Emaitza esperimentalen **analisia** burutuko dugu jarraian. 3 taulan ikusten den bezala, urrats bateko sistemak *batezbesteko haztatu* hobea du, baina klaseka ematen diren emaitzetan sistemak asko ikasi ez duela nabaria da, gehienetan erlaziorik ez dagoela iragarri du eta. Urrats bateko eta bi urratsetako sistemaren bigarren fasea konparatuz, desorekari aurre egiteko proposatutako hurbilpena egokia suertatu da, nahiz eta *F1 haztatua* jaitsi, klase positiboen F1 asko igo baita, *protein-component* klasearen kasuan $0,03$tik $0,64$ra igo baita. Portaera hau aztertutako hiru azpi-atazetan ematen da. 4 irudian ikus daitekeen bezala *F1 haztatua* oso altua da, aurreko kasuan bezala, hiru sistemek erlaziorik ez dagoela iragartzeko joera dutelako, hain zuzen ere. Klase anitzeko eta hurbilpen bitarraren emaitzen arteko *makro F1* neurriak konparatzen baditugu, $0,42$ eta $0,77$, agerian geratzen da hurbilpen bitarrak sistemaren sendotasuna hobetu duela alborapenaren aurka.

4. Irudia: Oinarri-lerroa: urrats bakarreko arkitektura, alegia, klase anitzeko hurbilpenarekin lortutako emaitzak.

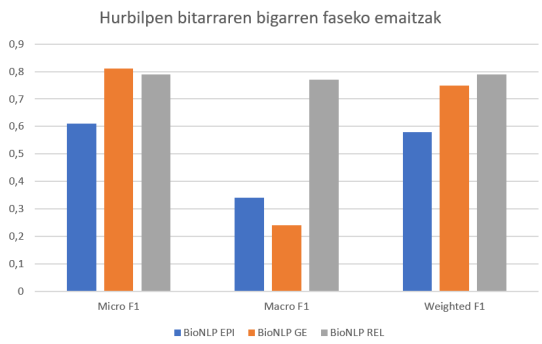


5. Irudia: Bi urratsetako arkitekturarekin lortutako emaitzak: 1. erlaziorik badagoen; 2. erlazio-mota esleitu.

(a) 1. urratsean: erlaziorik dagoen aurreikusi



(b) 2. urratsean: erlazio mota iragarri



Jarraian, gure proposamena **aurrekariek** burututako lanarekin alderatuko dugu, nabarmenenak 4. taulan laburbildu ditugu. Hasteko, BioNLP 2011 REL erronkan lortutako hurbilpenak eta emaitzak bildu ditugu 4. taulako lehenengo lau errenkadetan. Nabarmendu behar da, BioNLP 2011 erronkan, parte-hartzaileek erlazio-erauzketa ez ezik, entitateak ere erauzten zituztela. BioNLP 2011 lehiaketako ebaluazioan *protein-component* eta *subunit-complex* erlazioak bakarrik ebaluatu ziren, baina gure sistema ebaluatzeko, corpusean anotatutako erlazio guztiak erabili dira, *equiv* erlazioa barne. Hau dela eta, emaitzak ezin dira zuzenean alderatu. 2011ko aurrekarien artean, kalitate onena *Support Vector Machine* (SVM) (Cortes eta Vapnik, 1995) motako sailkatzailearen bidez lortu zen (Björne eta Salakoski, 2011; Van Landeghem *et al.*, 2011). 3. eta 4. sistemek (Kilicoglu eta Bergler, 2011; Le Minh *et al.*, 2011) hiztegian oinarritutako entitate erauzketa erabili zuten erregelekin konbinatuta. 2011-ko sistemek eskuz planteatutako ezaugarriak erabiltzen zituzten eta nekez ustiatzen zuten hizkuntzaren semantika, testuingurua etab.

4. Taula: BioNLP 2011 REL azpi-atazan lortutako emaitzak, ofizialak eta ikasketa sakoneko inplementazio berri birekin lortutakoak, MAT-LSTM eta LBERT, eta amaieran lan honetan aurkeztutakoak.

	Urtea	Taldea	Sistema	Doitasuna	Estaldura	F1
1	2011	University of Turku	SVM	0,68	0,50	0,58
2	2011	VIB - Ghent Univeristy	SVM	0,37	0,48	0,42
3	2011	Concordia University	Dict + Rules	0,47	0,25	0,32
4	2011	University of Science, VNU, HCMC	Dict + Rules	0,23	0,16	0,19
5	2018	MAT-LSTM	Attentive LSTM	0,48	0,77	0,59
6	2020	LBERT	BERT	0,62	0,59	0,59
7	2021	Ixa@IkerGazte: Urrats batean	Joint AB-LSTM	0,87	0,91	0,88
8	2021	Ixa@IkerGazte: Bi urratsetan	Joint AB-LSTM	0,79	0,79	0,79

Azken bost urteotako sistemek, automatikoki erauzitako ezaugarriak erabiltzen dituzte. Alegia, testuak karakterizatzeko latente dauden ezaugarriak erauzten dira automatikoki sare neuronalen bidez. Aldaketa nabaria izan da testuak deskribatzeko eskuz planteatutako ezaugarriak erabiltzetik ezaugarri deskriptibo latenteak erabiltzera pasatzea. Sare neuronalak berak testuak deskribatzeko gaitasuna du. Gaitasun hori urteekin aberastu da eta testuinguruak dakarren ñabardurak modelatzerako bidean garatu da. BioNLP 2011 lehiaketako REL azpi-ataza ikasketa sakonean oinarritutako sistemekin ere ebatzi da, 4. taulako 5. eta 6. hurbilpenetan adibidez. MAT-LSTM ikasketa sakoneko inplementazioak atentzio anizkun LSTM bat darabil (Bai *et al.*, 2020). Sarea entrenatzeko corpus erraldoiak erabili dituzte edozein motako erlazio klinikoak sailkatzen irakasteko asmoz. Gure sistemak MAT-LSTM hurbilpenak baino doitasun hobea lortu du antzeko estaldurarekin. LBERT (Warikoo *et al.*, 2020), domeinu klinikorako adaptatutako BERT (Devlin *et al.*, 2018) sare neuronalean oinarritzen da (4. taulako 6. lerroan). Gure sistemak eta MAT-LSTM sistemak emaitza hobeak lortu dituzte estalduran. Gure kasuan, Bi-LSTM bi uztartzen duen eredua aukeratu dugu, alegia, ikasketa sakoneko hurbilpena da. Ez dauzka BERT-ek dituen gaitasun aurreratuak testuak karakterizatzeko, aitzitik, erlazio erauzketa atalerako ikasketa algoritmoa sendoa dela dirudi emaitzen arabera.

4 Ondorioak

Lan honetan erlazioak erauzteko sistema bat aurkeztu da. Sistema honen erronka nagusia alboratuta dauden datuekin lan egiten duten atazei aurre egitea izan da. Honetarako, Joint AB-LSTM hurbilpena hartu dugu oinarri gisa, sendotu dugu optimizazio prozesua desorekarekiko sentikortasuna txikitzearren, *makro F1* neurriak erabiliz optimizazio prozesuan eta hedatu egin dugu klase anitzeko atazei aurre egin ahal izateko. Bi arkitektura konparatu ditugu: urrats bakarrekoa eta bi urratsetan gauzatzen dena. Emaitzak ikusita, batez ere *F1 makro* neurriari erreparatuta, agerian dago sistema nahiko sendoa dela eta alboratutako datuekin lan egiteko gai dela. Lanean atazarako egokia den sare neuronala ez ezik, aurreprozesamenduan estrategia bat inplementatu da alborapena gutxitzeko. Alborapena gutxitzeko teknika hau ere findu egin da, kasu negatiboak neurriz sortzeko, esponentziaki haz ez daitezen. Emaitzetan oinarrituta, bi urratsetako estrategiaren bidez kalitatea hobetzen da era nabarian (+0,35 F1 REL datu-sortan).

5 Etorkizunerako lana

Etorkizunerako erronkak gelditzen dira oraindik eta horietako bat sistemari entitate ezagutzailea integratzea da. Era horretara, sistema entitateak eta erlazioak era bateratuan erauzteko gai izango litzateke. Lan honetan bakarrik ingelesez dauden dokumentuak erabili dira, baina beste hizkuntzetara aplikatu daiteke, adibidez euskarara. Oraindik ez dago euskaraz medikuntza-arloko corpus handirik, baina hau biltzeko lanak martxan daude. Hurrengo urratsak medikuntza alorreko testuingurudun *embedding*ak erabiltzea litzateke (BioBERT edo BioXLNet adibidez). BERT-en oinarritutako karakterizazioak sintaxia eta semantika konplexua biltzeko gai baitira eta honek ereduaren ulermen artifiziala hobetzeko gaitasuna erantsiko duelakoan gaude. Amaitzeko, Joint AB-LSTM sare neuronalak darabilen sarrerako datu formatua ez da egokia hizkuntzaren ezaugarri konplexu guztiak modelatzeko, formatu malguago bat beharko luke, adibidez BILOU (Ratinov eta Roth, 2009). Datu-sortetan klase desoreka oso handia egon da klase positibo eta negatiboen artean. Hau ekiditeko datu-sorta bitarra eraiki da; hain zuzen ere, klase negatiboak iragazteko. Teknika hauek sailkapenean lagundu dute, baina oraindik sailkatzaileak klase negatiboa iragartzeko joera du. Hau ekiditeko teknika gehiago probatu nahi dira, *smoothing* eta *re-inforce learning* adibidez.

Erreferentziak

Ananiadou, Sophia, eta Jun'ichi Tsujii. 2012. stav: text annotation visualiser. In *Proceedings of the Eighteenth Annual Meeting of the Association for Natural Language Processing*, 341–344.

Bai, Tian, Chunyu Wang, Ye Wang, Lan Huang, eta Fuyong Xing. 2020. A novel deep learning method for extracting unspecific biomedical relation. *Concurrency and Computation: Practice and Experience* 32.e5005.

Björne, Jari, eta Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, 183–191.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, eta Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, 2787–2795.

Cortes, Corinna, eta Vladimir Vapnik. 1995. Support vector machine. *Machine learning* 20.273–297.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, eta Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Goyal, Archana, Vishal Gupta, eta Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review* 29.21–43.

He, Bin, Yi Guan, eta Rui Dai. 2018. Convolutional gated recurrent units for medical relation classification. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 646–650. IEEE.

Hochreiter, Sepp, eta Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9.1735–1780.

Kang, Ning, Bharat Singh, Chinh Bui, Zubair Afzal, Erik M van Mulligen, eta Jan A Kors. 2014. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics* 15.64.

Kilicoglu, Halil, eta Sabine Bergler. 2011. Adapting a general semantic interpretation approach to biological event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, 173–182.

Kim, Jin-Dong, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, eta Jun'ichi Tsujii. 2011a. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP shared task 2011 workshop*, 1–6. Association for Computational Linguistics.

——, Yue Wang, Toshihisa Takagi, eta Akinori Yonezawa. 2011b. Overview of genia event task in bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, 7–15. Association for Computational Linguistics.

Le Minh, Quang, Son Nguyen Truong, eta Quoc Ho Bao. 2011. A pattern approach for biomedical event annotation. In *Proceedings of BioNLP Shared Task 2011 Workshop*, 149–150.

Li, Jing, Aixin Sun, Jianglei Han, eta Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* .

Li, Lishuang, Jieqiong Zheng, Jia Wan, Degen Huang, eta Xiaohui Lin. 2016. Biomedical event extraction via long short term memory networks along dynamic extended tree. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, 739–742. IEEE.

Nguyen, Thien Huu, eta Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 39–48.

Ohta, Tomoko, Sampo Pyysalo, eta Jun'ichi Tsujii. 2011. Overview of the epigenetics and post-translational modifications (epi) task of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, 16–25. Association for Computational Linguistics.

Pérez, Alicia, Arantza Casillas, eta Koldo Gojenola. 2016. Fully unsupervised low-dimensional representation of adverse drug reaction events through distributional semantics. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, 50–59.

Pyysalo, Sampo, Tomoko Ohta, eta Jun'ichi Tsujii. 2011. Overview of the entity relations (rel) supporting task of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, 83–88. Association for Computational Linguistics.

Ratinov, Lev, eta Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 147–155, Boulder, Colorado. Association for Computational Linguistics.

Sahu, Sunil Kumar, eta Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of biomedical informatics* 86.15–24.

Santiso, Sara, Alicia Pérez, eta Arantza Casillas. 2018. Exploring joint ab-lstm with embedded lemmas for adverse drug reaction discovery. *IEEE journal of biomedical and health informatics* 23.2148–2155.

Sokolova, Marina, eta Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45.427–437.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, eta Junichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107.

Van Landeghem, Sofie, Thomas Abeel, Bernard De Baets, eta Yves Van de Peer. 2011. Detecting entity relations as a supporting task for bio-molecular event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, 147–148.

Warikoo, Neha, Yung-Chun Chang, eta Wen-Lian Hsu. 2020. Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations. *Bioinformatics* .

Yadav, Vikas, eta Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470* .

6 Eskerrak

Lan honetarako finantziazioa honako iturrietatik lortu da: Espainiako Gobernuko Zientzia eta Berrikuntzako Ministerioa (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), Europako Batzordea (FEDER) eta Eusko Jaurlaritzako diru-laguntzetatik (IXA IT-1343-19 eta Sergio Santana ikasleari emandako diru-laguntza, 12/03/2020 EHAAn argitaratua).