V. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2023ko maiatzaren 17, 18 eta 19a
Donostia, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Osasun-arloko entitate izendunen etiketatzea

Paula Ontalvilla Gutierrez,
Aitziber Atutxa Salazar
eta *Maite Oronoz Antxordokiz*

91-98 or.
https://dx.doi.org/10.26876/ikergazte.v.03.12

ANTOLATZAILEA: udako euskal unibertsitatea
BABESLEAK: Eusko Jaurlaritza Gobierno Vasco Hezkuntza Saila Departamento de Educación; Donostia San Sebastián Euskara; Gipuzkoa Etorkizuna Orain
LAGUNTZAILEAK: eman ta zabal zazu; Deusto Universidad de Deusto Deustuko Unibertsitatea; Université de Pau et des Pays de l'Adour; UNED Bergara; Mondragon Unibertsitatea; upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa; Universidad de Navarra; Udalbiltza

Osasun-arloko entitate izendunen etiketatzea

Paula Ontalvilla, Aitziber Atutxa, Maite Oronoz

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

pontalvilla001@ikasle.ehu.eus

Laburpena

Lan honek helburu bikoitza du: alde batetik, transformerretan oinarritutako hizkuntza-ereduak erabiliz medikuntzaren alorreko entitate izendunen identifikazioa egiten du, eta bestetik, identifikatutako entitate klinikoak Wikidata ezagutza-baseko gaixotasunekin eta sintomekin lotzen ditu. Entitateak ezagutzeko, biomedikuntzako *MedMentions* corpusaren gainean, aldez aurretik entrenatutako BERT hizkuntza-eredu orokor batekin (BERT small) eta bi BERT espezializaturekin (BiomedNLP-PubMedBERT eta BioBERT) egin dira esperimentuak. Token segida batek medikuntzako entitate bat osatzen ote duen ebaluatu denean, 0,819ko F1 balioa lortu da, eta entitatea zein klase zehatzetakoa den ebaluatu denean, 0,62ko F1 balioa. Gainera, Levenhstein distantzia erabiliz ezagututako entitateak Wikidatarekin lotzeko lehenengo saiakeran %50 inguruko estaldura lortu da.

Hitz gakoak: Entitate izendunen ezagutza, hizkuntza-ereduak, Wikidata, medikuntza

Abstract

This work has a double objective: on the one hand, it identifies named entities using language models based on transformers and, on the other hand, it links the identified clinical entities with the diseases and symptoms of the Wikidata knowledge base. To identify the entities, experiments have been performed on the MedMentions biomedical corpus with a generalpre-trained language modeñ BERT (BERT small) and two specialised BERTs (BiomedNLP-PubMedBERT and BioBERT). When assessing whether a succession of tokens constitutes a medical entity, an F1 value of 0.819 was obtained, while assessing the specific class to which the entity belongs, an F1 value of 0.62 was obtained. In addition, a recall close to 50% has been achieved in the first attempt to associate Wikidata to known entities using the Levenhstein distance.

Keywords: *Named Entity Recognition, language models, Wikidata, medicine*

1 Sarrera eta motibazioa

Entitate izendunen ezagutza, *Named Entity Recognition (NER)* ingelesez, informazio erauzketaren azpiataza bat da, zeinetan testuetan agertzen diren entitate izendunak identifikatu eta kategoria zehatz batzuetan sailkatzen diren (azpiataza hau NERC ere deitu izan da, *Named Entity Recognition and Classification* ingelesez[1]). Ohiko NER atazetan entitateak pertsona-izenetan, leku-izenetan eta erakunde-izenetan sailkatu ohi dira. Adibidez, Barak Obama-pertsona, EHU-erakundea, Euskadi-lekua. NER informazioaren kudeaketarako funtsezkoak diren beste arlo askotarako oinarria da, hala nola, anotazio semantikoa, galdera-erantzun sistemak, ontologien populatzea eta iritzien meatzaritza (Marrero et al., 2013). "*Entitate izendun*" terminoa lehenengo aldiz 6th MUC konferentzian (Grishman eta Sundheim, 1996) erabili zen. Hemendik aurrera NER inguruko interesa piztu da eta ekarpen desberdinak egin dira konferentzia ezberdinetan, adibidez: CoNLL03 (Sang eta De Meulder, 2003), ACE (Doddington et al., 2004), IREX (Demartini et al., 2010) eta TREC Entity Track (Balog et al., 2010). NER orokorraz haratago, domeinu espezializatuagoetan aritzen diren NER sistemak garatu dira, batez ere biomedikuntzaren arloan. Medikuntzaren domeinuan NER erabiltzearen helburua, garrantzitsuak diren entitate klinikoak, hala nola, gaixotasunak, sintomak, botikak . . . detektatzea da, hori guztia, testu klinikoetan. Ataza honi Clinical NER (Kundeti et al., 2016) edo NER klinikoa egitea deritzo.

NER klinikoa oso erabilgarria izan daiteke ataza desberdinak aurrera eramateko; hala nola, sintomak testuan

[1] Aurreratzean eta sinplifikatzearren, euskaraz ere, NER laburtzapena erabiliko dugu.

identifikatuta gaixotasunak iragar ditzakegu, edo adibidez, testuko sintomak dagozkion testuko gaixotasunekin lot ditzakegu. Era berean, gaixotasunak eta botikak deskribatzen dituzten entitateak ezagutuz gero, gaixotasunen eta horiek sendatzeko emandako botikak egokiak diren jakiten saia gaitezke, etab. Hori guztia egiteko entitate izendunak testu klinikoetan ezagutuko dituen sistema bat behar dugu. Sistema hori eraikitzeko, hizkuntza-eredu bat erabiliko dugu hizkuntza errepresentatzeko. Hizkuntza-eredua domeinukoa den testu bilduma batekin findu ohi da ikasketan espezializatzeko, eta kasu honetan biomedikuntzako entitatedun testu-bilduma bat erabiliko dugu, arlo klinikoari dagozkion entitate-klasetan arreta berezia ipiniaz. Testu klinikoak normalean era librean idatzita daude, inolako uniformetasunik eta egiturarik gabe. Hortaz, egituratu gabeko testu horietatik informazioa egituratzeko lehenengo urratsa NER egitea da (Kundeti et al., 2016).

Lan honetan, medikuntzako NER sistema gainbegiratu bat egin nahi da, hau da, testuan osasun-arloko entitateak identifikatu nahi dira, gero ezagutza-base batekin lotzeko. Ezagutza-baseetan orokorrean informazio oso aberatsa gordetzen da, hala nola, kontzeptuen arteko erlazio motak, sinonimoak, eta hainbat erlazio hierarkiko. Era berean, kontzeptuak kodeen bitartez identifikatzen dira, adibidez, UMLS (*Unified Medical Language System*) (Bodenreider, 2004) ezagutza-basean UMLS kodeak erabiltzen dira kontzeptuen lexikalizazioak diren entitateak identifikatzeko. Izan ere, erabiliko den ezagutza-baseak Wikidatako gaixotasunak ditu bakoitzeko sinonimoak, sintomak, kausak, UMLS kodeak, etab. gordetzen dituelarik. Modu horretan, testuak ezagutza-baseko medikuntzari buruzko ezagutza, osagarri izango du. Esan bezala, medikuntzako entitate ezagutzaile on bat entrenatu nahi dugu, aukera ezberdinak aztertuz (3. atalean). Ezagutza-basea aldez aurretik landu eta sortu denez[2], artikulu honetan testuaren eta ezagutza-basearen arteko lotura soilik landuko da.

2 Arloko egoera, eta helburuak

NER klinikoan eta biomedikoan azken urteetan aurrerapen nabarmenak egin dira. Alde batetik, arkitektura berriek, adibidez, neurona-sare errekurrenteak, haien artean Long Short-Term Memory (LSTM) (Hochreiter eta Schmidhuber, 1997) deritzanak eta Conditional Random Field (CRF) (Lafferty et al., 2001) algoritmoak, 2017an lortu zituzten artearen egoerako emaitzak hobetzea (Habibi et al., 2017). Bestetik, aurrerapen horiek azken urteotan garatutako ELMo (Peters et al., 2018) eta BERT *(Bidirectional Encoder Representations from Transformers)* (Devlin et al., 2018) moduko hizkuntza-ereduak erabilita, sortutako sistemen eskutik ere etorri dira. Adibidez, ELMo-k emaitza oso onak lortu zituen kontzeptu klinikoak erauzten (Zhu et al., 2018); (Si et al., 2019) lanean aldiz, BERT erabili zuten kontzeptu klinikoen erauzketa hobetzeko.

BERT eta ELMo domeinu orokorreko testuen gainean entrenatuta daude, Wikipedian adibidez. Hortaz, zaila da biomedikuntzako testuetan izango duten portaera aurreikustea. Gainera, corpus orokorretan eta biomedikuntzako corpusetan hitzen distribuzioa nahiko desberdina denez domeinu orokorrean entrenatutako hizkuntza-ereduen portaera biomedikuntzatako atazeetan zalantzagarria da. Biomedikuntza bezalako domeinu espezializatuetan, aurretik egindako lanek erakutsi dute domeinu-testua erabiltzeak emaitzak hobetzen dituela. BioBERT izeneko ereduan adibidez, (Lee et al., 2020) BERT bat entrenatzen da PubMed-eko laburpenen eta *PubMed Central*-eko artikulu osoen (PMC) gainean. PubMedBERT (Gu et al., 2021), aldiz, PubMed artxibategiko artikuluen laburpenen eta PubMedCentral artikuluen gainean aurre-entrenatuta dago. Bi BERT horien kasuan, biomedikuntzan espezializatua egoteak biomedikuntzako atazetan, NERen adibidez, emaitzak hobetzen dituela frogatu da. Domeinu klinikoan, aldiz, ClinicalBERT (Alsentzer et al., 2019) ereduak emaitza onak lortu ditu txosten klinikoekin.

Artikulu honetan MedMentions (Mohan eta Li, 2019) biomedikuntzako corpus etiketatua erabiliko dugu, transformerretan oinarritutako BERT hizkuntza-eredu ezberdinak baliatuta NER egiteko, eta BERT+NER sistema horiek konparatzeko. Corpusa PubMed-en[3] argitaratutako 4.392 artikuluen izenburua eta laburpenaz osatutako dago eta UMLS-ko 128 entitate-klase etiketatuta ditu (gaixotasunak, birusak, botikak, etab). BERT orokor bat eta biomedikuntzan espezializatuak dauden bi BERT espezializatu desberdin probatu dira: PubMedBERT eta BioBERT, azkenengo horiek hobeto funtzionatzen dutela baieztatzeko asmoz. Izan ere, hori horrela dela frogatzen da (Fraser et al., 2019) lanean, non corpus berdinaren gainean antzeko hurbilpen bat egin duten 0.56ko F1-neurria lortuz. Bestalde, testuan ezagututako entitateak, Wikidatarekin (Waagmeester et al., 2020) lotu dira arloan ekarpen berria eginez.

Wikidata ezagutza-base libre eta irekia da. Arlo askotako informazioa biltegiratzen du, hala nola, medikuntza, geografia, historia, etab-eko artikuluetako informazio laburtua gordetzen du RDF tripletetan (entitate1, erlazio, entitate2). Wikidata informazio-iturri oso aberats eta erabilgarria da. Adibidez, gaixotasunen inguruan hizkuntza

[2] https://github.com/paulaonta/InfoExtractor
[3] https://pubmed.ncbi.nlm.nih.gov/

desberdinetan informazio asko gordetzen du; hala nola, sintomak, jatorria, diagnostikoak, etab. Zehazki, 16 mila gaixotasun inguru ditu biltegiratuta Wikidatak (Waagmeester et al., 2020).

Lan honek bi **helburu** nagusi ditu. Alde batetik, transformerretan oinarritutako hizkuntza-ereduak erabiltzen dituen eta estaldura eta doitasun altuak izango dituen medikuntzaren alorreko entitate ezagutzaile batzuk garatzea (ikus 1 eta 2 atazak). Beste alde batetik, testuaren eta Wikidata ezagutza-basearen arteko lotura nola egin aztertzea (ikus 3. ataza).

Aipatutako bi helburuak betetzeko, ondorengo hiru atazak garatu dira:

- **1. ataza**. MedMentions corpusa erabiliz NER sistema orokor bat entrenatu, termino bat UMLS-koa den, hau da, osasunarekin zerikusia ote duen, jakiteko. Lehen ataza honetan terminoak identifikatuko dira, klase bakarra irteeran dagoelarik (MED klasea, termino bat medikuntzakoa dela adierazten duena).
- **2. ataza**. Klase bakarra identifikatu beharrean, bigarren hurbilpen honetan, MedMentions corpuseko klase semantiko ezberdinak bereiziko dira. Ez dira MedMentions-eko 128 klaseak iragarriko, antzemate finagoa egingo da. Zehazki, 1. taulako 23 klaseak detektatuko dira gaixotasunekin erlazio zuzena duten klaseak direlako.

1. Taula: Detektatu diren klaseak eta hauen kodeak

Klase mota	IDa	Klase mota	IDa
Health Care Activity	T058	Embryonic Structure	T018
Laboratory Procedure	T059	Fully Formed Anatomical Structure	T021
Diagnostic Procedure	T060	Body Part, Organ, or Organ Component	T023
Therapeutic or Preventive Procedure	T061	Anatomical Abnormality	T190
Disease or Syndrome	T047	Congenital Abnormality	T019
Mental or Behavioral Dysfunction	T048	Acquired Abnormality	T020
Neoplastic Process	T191	Clinical Drug	T200
Experimental Model of Disease	T050	Pharmacologic Substance	T121
Organism	T001	Antibiotic	T195
Virus	T005	Finding	T033
Bacterium	T007	Sign or Symptom	T184
Anatomical Structure	T017		

- **3. ataza**: Behin gaixotasun bat detektatu denean, Wikidatarekin lotura egingo da Levenhstein distantzia (Navarro, 2001) neurtuz. Horrela automatikoki testuan gaixotasuna detektatu eta Wikidatara lotuko da. Levenshtein distantzia berera, adibidez 2 distantziara, hainbat medikuntzako sarrera baleude, sarrera guztiekin egingo da lotura. Berdina egingo da sintomekin.

Lehenengo bi atazak aurrera eramateko, hizkuntza-eredu desberdinak doituko dira, hau da, *fine-tuning*-a egingo da. BERT orokor batekin eta medikuntzaren domeinuan espezializatuak dauden bi BERT desberdinekin egingo dira probak asmatze-tasa handitzeko asmoarekin.

3 Ikerketaren muina

Atal honetan, hasteko, erabilitako datuak deskribatuko dira. Ondoren, problema ebazteko sistema azalduko da hobetzeko asmoarekin egin diren aldaketak azalduz. Jarraitzeko, diseinatutako sistemen errendimendua neurtzeko eta domeinu orokorretan eta espezializatuetan aurre-entrenatutako BERT ereduak konparatzeko diseinatutako ebaluazio-sistema eta honen emaitzak azalduko dira.

3.1 Datuak

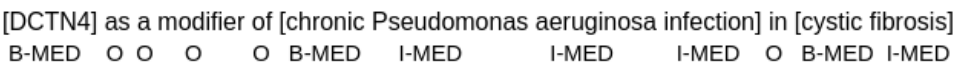
Aipatu bezala, MedMentions-eko corpusa (Mohan eta Li, 2019) erabili da lan honetan. Esaldiei dagokionez, 42602 esaldi ditu esaldiaren luzera batezbeste 27,6 tokenekoa delarik; guztira, corpusak 1176058 token eta 579839 token anotatu ditu. Bestalde, corpusak *train*, *dev* eta *test* partizioak eginda ditu dagoeneko bakoitzean corpuseko %60, %20, %20-ko ausazko kasuak dituelarik. Partizio bakoitzean dagoen testu (izenburu eta laburpen) bakoitzeko PMID kodea, corpus honetan kasu bakoitza identifikatzeko erabiltzen dena, dago bakarrik. Beraz, hori harturik

kasu bakoitzeko, hasteko, dagokion izenburua eta laburpena lortu dira. Horrela partizio bakoitzean testua egongo da token eta token anotatuekin batera.

Hori egin ostean, datuak definitutako sistemen ikasketarako egokia den etiketatze-formatura pasatu dira. Erabiliko diren sistemek, aurrerago azalduko direnak, *Beginning-Inside-Outside* edo BIO etiketatzea eskatzen dute. BIO etiketatzeak, tokenak etiketatzeko oso ohikoa denak, *B* etiketa erabiltzen du entitate baten hasiera adierazteko; *I* etiketa entitate baten barruan dagoela adierazteko eta *O* etiketa entitate kanpokoa dela adierazteko (Ramshaw eta Marcus, 1999). Hortaz, datuak BIO etiketatzera pasatu dira kontuan izanik ataza bakoitzean etiketa desberdinak egongo direla.

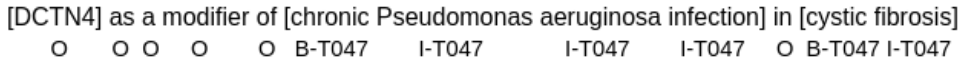
Klase bakarra (MED) detektatu behar da 1. atazan, MED izeneko klaseak hitzak edo hitz-segidak medikuntzarekin zerikusia ote duen adierazten du. Beraz, tokena entitate baten hasiera baldin bada `B-MED` etiketa jasoko du. `I-MED` izango du, aldiz, entitatearen barruan baldin badago eta `O` ez bada entitate baten parte. 1. irudian adibide bat aurkezten da.

1. Irudia: Etiketa bakarra (MED) daukan BIO etiketatzearen adibidea.



Klase anitz (ikus 1. taula) detektatu behar dira 2. atazan. Beraz, tokena entitate baten hasiera baldin bada `B-klasearenID` etiketa jasoko du. `I-klasearenID` izango du, aldiz, entitatearen barruan baldin badago eta `O` ez bada entitate baten parte edota klase horren identifikadorea 1. taulan ez badago. Halaber, token batzuek bi klase dituzte, kasu horietan etiketatzeko klaseen identifikadoreak erabili dira komaz bereziturik; adibidez, `B-T019,T047`. Hurrengo irudian (ikus 2. irudia) aurkezten den adibidean, *DCTN4* entitate bat da baina honen klaseen identifikadoreak T116, T123 dira eta hauek ez daudenez aukeratu ditugun etiketen artean, `O` etiketa hartzen du. T047 klasea ordea, 1. taulan zehaztutako klaseen artean dagoenez etiketatzearen parte da:

2. Irudia: Hainbat etiketa dituen BIO etiketatzearen adibidea.



Jakinda 1. taulan 23 klase aurkezten direla eta corpusa BIO moduan etiketatuta dagoela; entitate bat dagoela adierazteko, klase bakoitzeko bi etiketa egongo dira (`B-klasearenID` eta `I-klasearenID`), eta bestalde, `O` etiketa egongo da tokena entitatearen parte ez dela adierazteko. Corpuseko entitate batzuek bi klase esleituak izan ditzakete, eta hori horrela, 47 etiketa (klase bakoitzeko bi + O = 23*2+1= 47) egon beharko lirateke asmatzeko. Baina `I-T021` etiketa ez dagoenez eta lau kode ([`B-T019,T047`], [`I-T019,T047`], [`B-T047,T190`] eta [`I-T047,T19`]) anbiguo daudenez, hots, bi UMLS kodeekin, 50 etiketa daude asmatzeko. Honek arazo bat suposa dezake biko konbinazio asko daudelako eta gerta daitekeelako etiketa batzuk entrenamenduko partizioan ez egotea. Gainera, egoera honetan dauden etiketa asko baldin badaude ereduaren asmatze-tasa txikia izan daiteke. Hala ere, 2. taulan ikus daiteke garapeneko 2 etiketa baino ez daudela entrenamenduko partizioan. Horrek esan nahi du 46 etiketa badaudela, hau da, %95a badagoela *train*-en. Beraz, ez dira asmatze-tasa baxuak espero.

Nahiz eta emaitza oso txarrak ez espero izan, probatu nahi izan da, bi klase-etiketa dituzten entitateentzako bakarra erabiltzean zer gertatzen ote den emaitzekin. Beraz, bi klase dituen entitate bat topatzen denean hainbat aukera daude batekin etiketatzeko (1) lehenengo klasea hartu edo (2) klase orokorrena hartu. Kasu honetan lehenengo aukera egin da.

2. Taula: Partizio bakoitzean dauden etiketa kopurua eta bakoitzeko *train*-en ez dauden etiketak, hurrenez hurren.

	Etiketa kopurua	Ez dauden etiketak
Entrenamendu partizioa *(train)*	50	—
Garapen partizioa *(dev)*	48	B-T050,T191, I-T050,T191
Ebaluazio partizioa *(test)*	49	∅

3.2 Sistema

Problema hau ebazteko sistema nagusi bat eraiki da zeini aldaketak egin zaizkion hobetzeko asmoarekin. Lehenik eta behin azalduko da egindako sistema nagusia, eta gero, aplikatutako hobekuntzak.

Lehenengo bi atazak burutzeko, entitate izendunen ezagutzari dagozkionak, hain zuzen ere, tokenizatzaile eta transformer bat definitu dira. Transformerrari dagokionez, domeinu orokorreko BERT erabili da hasieran eta honen aldaerak erabili dira geroago, emaitzetan aldaketarik baden aztertzeko. Tamaina desberdinetako BERT-ak daude, geruza eta dimentsio kopuru desberdinak dituztelarik. Zenbat eta txikiago izan eredua, azkarrago entrenatzen da baina lortzen diren emaitzak okerragoak izan ohi dira. Beraz, erdiko tamaina daukan *bert-small* erabili da (Turc et al., 2019) lehen esperimentuan. Tokenizatzailearen kasuan, BERT ereduek NER egiteko beharrezko moldaketak egin dira kodean. BERT tokenizatzaileak corpuseko hitzak azpi-hitzetan banatzen ditu, hau dela-eta etiketak egokitu behar dira BIO etiketatzea sekuentzia tokenizatuaren luzera berriarekin bat etortzeko. Hurrengo irudian, 3. irudian, ikus daiteke nola jatorrizko hitzaren BIO etiketa, dagozkion hitz guztiei esleitu zaion.

3. Irudia: BERT tokenizatzailearen adibide bat etiketak egokituta.

```
input in sentences (3 tokens):

       words: Hello San Sebastian!
       tags:  O     B   I

input as organized for batch_x and batch_adjusted_tags:

       subwords:    [CLS] Hel #lo San Sebast #ian ! [SEP] [PAD] ... [PAD]
       .word_ids(): None  0   0   1   2      2    2 None  None      None
       tags:        -1    O   O   B   I      I    I -1    -1        -1
```

Emaitza hobeak lortzeko asmoarekin aldaketak egin dira. Zehazki, bi BERT espezializatuekin, hau da, biomedikuntzako corpusen gainean aurre-entrenatuta dauden eredu handi hauekin probak egin dira: BiomedNLP-PubMedBERT eta BioBERT. Ereduak PubMed artikuluen gainean aurre-entrenatuta daudela eta MedMentions corpusak PubMed laburpenak dituela jakinda, asmatze-tasa igo dezateken aztertu nahi dugu.

Gauzak horrela, BERT desberdinak probatu dira NER etiketatzea egiteko. Ereduak entrenamenduko *(train)* datuekin doitu egin dira deskribatutako ataza zehatzean hobeto egiten ikasteko, hau da, fine-tunning egin da. Gero garapenerako corpusarekin (*dev*) hainbat aldiz egiaztatu da ereduek ondo egiten dutela, eta azkenik, ebaluazioko corpusean (*test*) ebaluatu dira.

Hirugarren ataza burutzeko, hau da, gaixotasunak eta sintomak Wikidata ezagutza-basearekin lotzeko, Levenshtein distantzia (Navarro, 2001) erabili da. Wikidatako ezagutza-basetik gaixotasunen eta sintomen informazioa erauzi ondoren, informazio hori egitura originaletik testura bihurtu zen eta corpus hori da lan honetan erabili dena. Wikidatako corpus honek gaixotasun bakoitzeko gaixotasunaren izena, sintomak, Wikipediarako esteka, tratamendurako botikak, etab. gordetzen ditu. Levenshtein distantzia, zenbaki bat da[4] hitz batetik abiatuta beste hitz bat lortzeko egin behar diren gutxieneko eragiketen kopurua adierazten duena. Entitate izendunak adierazita dituen corpusaren eta Wikidata ezagutza-baseko gaixotasunak eta sintomak gordetzen dituen corpusaren arteko lotura egiteko, probatutako BERT onenak detektatutako gaixotasunen eta sintomen eta Wikidatako gaixotasun guztien arteko distantzia neurtu da; distantzia 2 baino txikiago edo berdina baldin bada gaixotasun horren Wikipediako esteka lortu da. Gainera, sintoma bat Wikidatan sintoma moduan agertzen baldin bada markatu egin da [WikidataSympton] jarriz.

3.3 Emaitzak

Lehenengo bi atazetan diseinatutako hiru sistemak ebaluatzeko, zein BERT orokorra, zein espezializatuak, asmatze-tasa kalkulatzea ez zaigu egokia iruditu. Neurri honek ematen dituen emaitzak txarrak izateaz gain, oso altuak dira datuetan O etiketa asko baitaude. Horren ordez, doitasuna, estaldura eta F1-neurria [5] erabili dira. Doitasuna hautemandako elementu guztietan zuzen hautemandako elementuen proportzioa da eta estaldura hauteman beharreko elementu guztietatik hautemandako elementuen proportzioa da. F1-neurriak aldiz, beste bi metriken,

[4] Adibidez, *tonsillitis* Wikidatan agertzen den hitz zuzena balitz eta testukoa *tonsilitis*, bien arteko distantzia 1 da *l* bat txertatuz gero Wikidatakoa lortuko genukeelako.

[5] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

doitasuna eta estaldura metriken, batezbesteko harmonikoa kalkulatzen du. Bi metrika horiek alderantzizko proportzionalak direnez F1 metrikaren bidez kalkula dezakegu doitasunaren eta estalduraren balioak zein onak diren. Horrela, asmatze-tasa tokenen gainean egin beharrean entitate izendunen gainean egiten da. Ez dira emaitza hain altuak lortzen baina emaitzak errealistagoak dira O etiketak ez direlako zuzenean kontutan hartzen.

BERT desberdinekin eta zeregin bakoitzeko lortutako emaitzak, 3. taulan ikus daitezke. Lehenengo atazaren kasuan, PubMedBERT oso eredu handia denez bi sorta-tamainekin (ingelesez *batch size*) probatu da. Modu honetan, ikusi da sortaren tamainak nolako eragina duen emaitzetan. Baina 2. atazako bi bertsioetan sortaren tamaina 32ra ezarrita probatu da 1. atazan honekin lortu direlako emaitzarik onenak. Aipatu behar da, hobeto ulertze aldera, 2. atazako 1. bertsioan entitateek bi klase desberdin izan ditzaketela etiketan (taulan EntBatNKlase[6]) eta 2. atazako 2. bertsioan, aldiz, entitate bakoitzak klase bakarra dauka esleituta (taulan EntBatKlaseBat[7]).

3. Taula: Zeregin bakoitzeko, BERT desberdinekin lortutako doitasuna, estaldura eta F1 metriken emaitzak, hurrenez hurren. [1] sortaren tamaina = 4; [2] sortaren tamaina = 32.

Eredua	**1. Ataza**			**2. Ataza (EntBatNKlase)**			**2. Ataza (EntBatKlaseBat)**		
	Doi	Est	F1	Doi	Est	F1	Doi	Est	F1
BERT small	0,848	0,766	0,805	0,578	0,518	0,546	0,591	0,520	0,553
PubMedBERT-	0,823	0,796	0,809 [1]	0,649	0,559	**0,623**	0,656	0,590	0,621
BiomedNLP	0,811	0,826	**0,819** [2]						
BioBERT	0,820	0,817	**0,819**	0,627	0,608	0,618	0,661	0,591	**0,624**

Alde batetik, NER ebaluatzeko erabili diren metrika berdinak, hau da, F1, doitasuna eta estaldura erabili dira 3. ataza ebaluatzeko. Beste alde batetik, zein den gaixotasunen eta sintomen estaldura Wikidatan kalkulatu da. Horretarako, kalkulatu da zenbat gaixotasun detektatu dituen BERT ereduak eta horietatik zenbat dauden Wikidatan, 1. formulan ikus daitekeen bezala. Sintomen kasuan, 2 formulan ikus daitekeenez, berdina egin da.

$$\frac{Wikidatako\ gaixotasun\ kopurua}{BERTak\ detektatuko\ gaixotasun\ kopurua} \tag{1}$$

$$\frac{Wikidatako\ sintoma\ kopurua}{BERTak\ detektatuko\ sintoma\ kopurua} \tag{2}$$

Eredu onena erabili denez 3. ataza garatzeko, EntBatKlaseBat atazako BioBERT eredua erabili da. Hortaz, doitasuna, estaldura eta F1- neurria3. taulan ikus daitezke. Bestalde estaldurei dagokienez, gaixotasunen kasuan testuko entitateen **%46** eta sintomen kasuan **%13** lotzea lortu da.

3.4 Analisia

Kasu guztietan emaitza hobeak BERT espezializatuekin lortzen dira, 3. taulan ikus daitekeenez. Hori horrela izanda, esan daiteke ziurtasun osoz biomedikuntza domeinuan espezializatuak dauden BERTek emaitza hobeak ematen dituztela domeinu orokorretan entrenatuta daudenak baino. Bestalde, zereginen zailtasuna kontutan hartzen bada, errazagoak diren atazetan emaitza hobeak lortu dira. Izan ere, 1. atazarekin eta EntBatKlaseBat atazarekin emaitza hobeak lortu dira.

MED klase bakarra iragarri behar izan denean sortaren tamaina (*batch size*) desberdinekin lortu diren emaitzak aztertzen baldin badira, ikus daiteke emaitzen artean ez dagoela desberdintasun nabarmenik. Baina sortaren tamaina handiago denean lortu dira emaitza hobeak (F1 kontutan harturik). Hala ere, esan beharra dago, bi sorten tamaina desberdinekin probatu denez ezin dela ondorioztatu sorta-tamaina handiek emaitza hobeak lortzen dituztenik.

Hainbat klase iragartzearen atazan lortzen diren emaitzak MED klase bakarra iragarri behar denekoarekin konparatuz, baxuagoak dira. Hau gertatzen da, klase desberdinak iragartzea ataza zailagoa delako. Gainera 2. atazan garatutako bi bertsioak konparatzen baldin badira, EntBatKlaseBat atazarekin EntBatNKlase atazarekin baino emaitza hobeak lortu dira. Hala ere, nahiz eta EntBatNKlase ataza zailagoa izan, entitate batzuk bi klase dituztelako, EntBatNKlase ataza sinpleagoaren oso antzeko emaitzak lortzen dira, 0,01ko diferentzia baino ez dago.

[6] Aurreratzean EntBatNKlase etiketa erabiliko da erreferentzia egiteko.
[7] Aurreratzean EntBatKlaseBat etiketa erabiliko da erreferentzia egiteko.

Azkenengo atazako emaitzak ikusiz, gaixotasunen estaldura sintomena baina nabarmen altuagoa da. Honek zentzua du sintomak Wikidatako gaixotasunekin parekatzen direlako. Bestalde, gaixotasunen estaldura ez da oso altua ez baita ailegatzen %50ra. Honen arrazoia izan daiteke Levenshtein distantzia txikia ezarri izan dela goi-muga bezala.

4 Ondorioak

NER ataza egiteko teknika desberdinak daude. Artikulu honetan transformerrak erabili dira problema honi aurre egiteko. Zehazki, aurre-entrenatutako transformerretan oinarritutako hizkuntza-eredu edo BERT desberdinak erabili dira. Domeinu orokorrean aurre-entrenatuta dagoen *bert-small* eta gure corpuseko domeinu espezializatu berdinean, biomedikuntzan, aurre-entrenatuta BioBERT eta PubMedBERT erabili dira.

Egindako esperimentuen bidez, ikusi da BERT espezializatuek hobeto funtzionatzen dutela beti. Bestalde, ataza zenbat eta errazagoa izan, orduan eta eta emaitza hobeak lortzen dira. Hala ere, hainbat klase detektatu behar diren atazaren kasuan ataza errazenarekin (entitatekoa den token bakoitzari klase bakarra egokitzen zaio (edo ez)) ez da ataza zailarekin (entitatekoa den token bakoitzari hainbat klase esleitu ahal zaizkio) lortutako emaitzak asko hobetzea lortu.

Hori guztia horrela izanda, domeinu espezializatuen entitate izendun klinikoen ezagutza egin nahi bada hobe da domeinu horretan espezializatuak dauden BERT ereduak erabiltzea. Hainbat klase detektatu behar diren atazan egindako bi bertsioekin emaitza oso antzekoak lortzen dira. Hortaz, beti ataza sinplifikatzeak ez du merezi. Gainera lan honetan ez sinplifikatuak, hau da, EntBatNKlase atazak informazio gehiago eskuratzen du.

5 Etorkizunerako planteatzen den norabidea

Lan honi jarraipena emateko, hasteko, BERT espezializatu desberdinekin probatu daiteke asmatze-tasa igotzen ote den aztertzeko; adibidez, ClinicalBERT (Alsentzer et al., 2019) ereduarekin egin dezakegu proba.

Bestalde, entitateen eta Wikidata ezagutza-baseko medikuntza alorreko ezagutzaren arteko lotura hobetu daiteke. Horretarako, alde batetik, estaldura handitu daiteke Levenshtein distantzia minimoa igoz, 2ko distantzia 4ko distantziarekin aldatuz, adibidez. Gainera sintomak Wikidatan sintoma moduan bila daitezke eta modu honetan lortutako emaitzaren estaldura neurtu. Izan ere, sintoma-sintoma lotura eginda lortutako estaldurak lan honetan erdietsitako %13a baino altuago izan beharko luke. Beste alde batetik, entitateen desanbiguazioa, *entity linking* ingelesez, egin daiteke. Hau, Levenshtein ez bezala, antzekotasun semantikoan oinarritzen ohi da. Levenshtein neurriak hitzen azaleko formari erreparatzen dio, aldiz distantzia edo antzekotasun semantikoak hitzen esanahia hartzen du oinarri gisa. Horrela, adibidez, testuan *influenza* agertuz gero eta Wikidatan hau ez baldin badago entitateen desanbiguazioaren bidez Wikidatako *flu* gaixotasunarekin lotura ezarriko litzateke semantika erabiltzen duelako lotura ezartzeko. Levenshtein distantziarekin, aldiz, ezinezkoa izango litzateke lotura ezartzea. Gainera, ezagutza-basearekin lotura ezartzerakoan hainbat gaixotasun lortzen diren kasuetan, loturarekiko anbiguotasuna dagoenean, ezagutza-baseko gaixotasun zuzena identifikatu daiteke testuingurua kontutan izanik. Lan honetan Wikipediako orriekin lotura egiten da, hori dela eta zehazki, Wikification ataza egin beharko litzateke. Wikification testuko entitateak Wikipediako orri egokiarekin lotura ezartzearen ataza da. Hortaz, ataza hau egin daiteke Lin eta Zeldes autoreek (2021) lanean egin duten antzera.

Erreferentziak

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Balog, K., Serdyukov, P., & Vries, A. P. d. (2010). Overview of the trec 2010 entity track. Technical report, Norwegian univ of Science and Technology Trondheim.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Demartini, G., Iofciu, T., & De Vries, A. P. (2010). Overview of the inex 2009 entity ranking track. In *Focused Retrieval and Evaluation: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers 8*, 254–264. Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, 837–840. Lisbon.

Fraser, K. C., Nejadgholi, I., De Bruijn, B., Li, M., LaPlante, A., & Abidine, K. Z. E. (2019). Extracting umls concepts from medical text using general and domain-specific deep learning models. *arXiv preprint arXiv:1910.01274*.

Grishman, R. & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kundeti, S. R., Vijayananda, J., Mujjiga, S., & Kalyan, M. (2016). Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, 1937–1945. IEEE.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lin, J. & Zeldes, A. (2021). WikiGUM: Exhaustive entity linking for Wikification in 12 genres. *arXiv preprint arXiv:2109.07449*.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Mohan, S. & Li, D. (2019). Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ramshaw, L. A. & Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural language processing using very large corpora*, 157–176.

Sang, E. F. & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., et al. (2020). Wikidata as a knowledge graph for the life sciences. *Elife*, 9:e52614.

Zhu, H., Paschalidis, I. C., & Tahmasebi, A. (2018). Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*.

6 Eskerrak eta oharrak

Lan honetarako finantziazioa honako iturrietatik lortu da: Espainiako Gobernuko Zientzia eta Berrikuntzako Ministeritza (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), Europar Batasunaren NextGenerationEU/PRTR funtsetatik (Antidote, PCI2020-120717-2) MCIN/AEI /10.13039) eta Eusko Jaurlaritzako diru-laguntzetatik (IXA IT1570-22 eta Paula Ontalvilla ikasleari emandako diru-laguntza, 2022/07/11 EHAAko deialdiko erresoluzioan argitaratua).