eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Hitz-adiera desanbiguazio neuronala

Egilea

Josu Murua Larrarte

2020

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Hitz-adiera desanbiguazio neuronala

Egilea

Josu Murua Larrarte

Zuzendariak

Aitor Soroa eta Ander Barrena

Laburpena

Testu batetik hitz bat hartuta, hitz horrek zer esanahi duen edo zein den hitz horri dagokion adiera erabakiko duen ikaskuntza sakon eta sare neuronaletan oinarritutako sistema bat garatu da proiektu honetan. Sistema aurretik eginiko lan batetik hartu da. Sistema berrinplementatzea eta lan horretako esperimentuak erreplikatzea izan da proiektu honen hasierako helburua.

Hitz-adiera desanbiguazio ataza burutzeko, batetik, desanbiguatu nahi den hitzaren testuinguru lortzen da. Bestetik, hitz horri dagozkion adiera kandidatuen errepresentazioak lortzen dira hiztegi bat erabiliz. Azkenik, errepresentazioen arteko konparaketa bat egiten da eta adiera bat aurresaten da. Hasierako helburua gauzatu ostean, bi esperimentu berri egin dira sistemaren mugak zein diren ikusteko. Lehena, hiztegia ez erabiltzeak sistemarengan sor dezakeen galera ikustea izan da. Bigarrena, aldiz, hitzak adierekin konparatzeko neurri berri bat probatzea izan da.

Memoria honetan proiektua egiteko egin den ikasketa prozesua, inplementatu den algoritmoa eta egin diren ebaluazio desberdinak aurrerago azalduko dira beraien analisi eta ondorioak ateraz.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

1.1 Atazaren deskribapena

Hitz-adiera desanbiguazioa testuinguru jakin batean hitz polisemiko batek duen adiera edo esanahia antzematea da. Hitz polisemikoen esanahia zein den jakitea ataza zaila izan daiteke. Esate baterako, *baso* hitzak, Harluxet hiztegiaren arabera, adiera bat baino gehiago izan ditzake:

1. *iz. Zuhaitzez edota zuhaixkaz jantzitako lur-eremua, bereziki gizakiak landu ez duena.*

2. *iz. Edalontzia.*

3. *iz. Pareta definituak dituen hodi edo kanala. Bertan odola, linfa, landareen izerdia eta beste hainbat likido mugitzen dira.*

Hariari jarraituz, ondorengo esaldiak emanez gero, adibidez, zein izango litzateke esaldiko hitz bakoitzari dagokion adiera?

1. *Tabernan baso bat ur eskatu dut.*

2. *Euskal Herriko basoetan pagoa eta haritza da nagusi.*

Hizkuntza ondo ulertzen duen batek testuinguruan aintzat hartuz erraz antzeman dezake 1. esaldian *baso* hitzak *edalontzia* esan nahi duela eta 2. esaldian *zuhaitzez jantzitako lureremua*. Sistema automatiko batentzat, aldiz, lan neketsua izan daiteke horrelako adierak desanbiguatzea.

1.2 Prozesuaren azalpena

Proiektu honetan izenen adiera desanbiguazioaren inguruan lan egin da eta helburua burutzeko ikasketa automatikoko zein hizkuntzaren prozesamenduko hainbat teknika erabili dira. Lehenik eta behin, hitz-adiera desanbiguazioa gauzatzeko sistema bat garatu da.

Hitz-adiera desanbiguazio ataza burutzen duen sistema batean desanbiguatu nahi diren hitzak beraiei dagozkien adierekin lotu behar dira. Horretarako, hitzen eta adieren errepresentazioak kodetuko dira hizkuntza ereduak erabiliz.

Proiektua egiteko [Scarlini et al., 2020]-ren lanaz baliatu da. Lan horretatik hartu da SensEmBERT baliabidea, baita hitz-adiera desanbiguazioa burutzeko sistema ere. SensEmBERT hitzen adierak ordezkatzeko ezagutzan oinarritutako saiakera berria da. SensEmBERT baliabideak izenen adierak errepresentatzeko bi mota ditu, ezagutzan soilik oinarritutako errepresentazioak eta ezagutzaz aparte testu anotatuan oinarritutako errepresentazioak.

Hitz-adiera desanbiguazioa egiteko garatu den sistemak honela funtzionatzen du. Hasteko, adieren errepresentazioak lortzen dira hizkuntza eredu batean oinarrituz eta hiztegi baten laguntzaz xede-hitzaren kandidatuak filtratzen dira. Ondoren, testuingurutik desanbiguatu nahi den hitzaren errepresentazioa eskuratzen da hizkuntza eredu bera erabiliz. Amaitzeko, kandidatuen adieren errepresentazioak hitzaren testuinguru errepresentazioarekin konparatzen dira eta antzekotasunean oinarrituz adieren artean bat aukeratzen da hitzari adiera bat esleituz. Beraz, sistemak ongi funtziona dezan beharrezkoa da errepresentazio hauen kodeketa ona izatea.

Tamalez sistemaren kodea ez dago atzigarri eta ondorioz proiektuaren lehenengo helburua [Scarlini et al., 2020]-ren laneko sistema berrinplementatzea izan da. Berrinplementazioa egin ostean, esperimentu berriak proposatu dira eta lortutako emaitzak aurrekoekin konparatu dira.

Berrinplementazioa egin eta gero, bi esperimentu nagusi burutu dira. Lehenengo esperimentuan hitzak desanbiguatzeko garaian kandidatuak filtratzeko hiztegia ez erabiltzeak

suposa dezakeen galera jakiteko esperimentua izan da. Hizkuntza batzuetan hiztegia erabiltzea ez da posible lematizazioa arazoa izan daitekeelako, adibidez Euskara hizkuntzan arazoa da. Esperimentu hau egin ahal izateko hitzak desanbiguatzeko bi metodo garatu dira: hiztegiduna eta hiztegi gabea.

Adiera errepresentazioak xede-hitzen testuinguruko errepresentazioekin konparatzeko neurri bat baino gehiago daude. Normalean K-NN erabiltzen da eta bigarren esperimentuan, berriz, errepresentazioak konparatzeko neurriekin jokatu da. Esperimentu honetan CSLS ([Lample et al., 2018]) neurri berria probatu da bilaketa espazio handietan ondo funtzionatzen duelako.

Memorian proiektuaren nondik norakoak azaldu dira hainbat kapitulutan antolatuta. 2. kapituluan hitz-adiera desanbiguazioaren artearen egoera eta ikaskuntza sakonean oinarritutako teknika berrienak azaldu dira. 3. kapituluan [Scarlini et al., 2020]-ren artikuluko adieren errepresentazioak nola lortzen diren azaldu da. 4. kapituluan inplementatu diren metodoak azaldu dira. 5. kapituluan egin diren esperimentuak eta lortutako emaitzak azaldu dira. Azkenik, 6. kapituluan proiektutik ateratako ondorioak eta etorkizunean irekita geratu diren bideak azaldu dira.

2. KAPITULUA

Artearen egoera

Kapitulu honetan proiektuaren oinarrien artearen egoera azalduko da. Proiektuaren oinarriak hizkuntza prozesamendurako ikaskuntza sakonaren erabilera, *transformer*rrak, BERT hizkuntza eredu neuronala eta hitz-adiera desanbiguazioan erabiltzen diren baliabideak dira.

2.1 Ikaskuntza sakona Hizkuntza Prozesamendurako

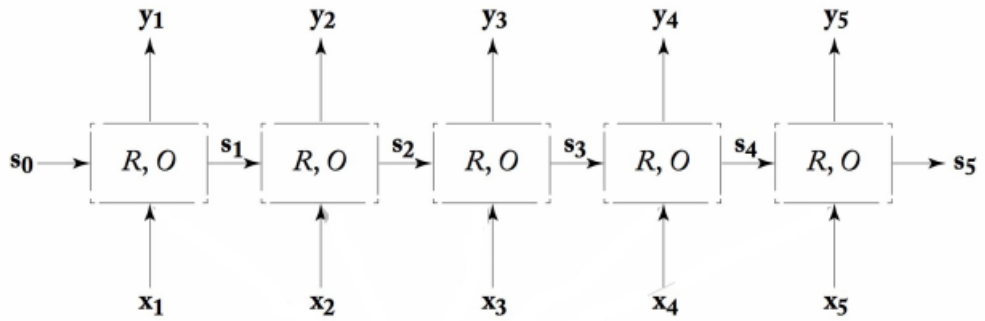
Hizkuntza prozesamendua gizakion hizkuntzak ordezkatzeko eta analizatzeko algoritmoen sorkuntzaz arduratzen da. Algoritmo hauen sorkuntza ez da gaur egungo kontua. 1950ez geroztik, Turing-ek Turing-en testa [1] sortu zuenetik, ari da gizakia hizkuntzak ordezkatzeko baliabide eta metodo berriak sortu nahian.

Azkeneko hamarkadan indarra hartu duen metodoa ikaskuntza sakona izan da. Aurretik geruza gutxiko sare neuronalak erabiltzen ziren, baina emaitzek ez zuten hobekuntza handirik izan. Geruza asko erabiltzen hasteak ekarri du ikaskuntza sakonaren gorakada. Horretaz gain, corpus handien sorrerak ere ahalbidetu du gorakada hein handi batean, sare neuronalen entrenamenduak esanguratsuagoak eta tamaina handiagokoak izan baitira. Kostu konputazionala ere areagotu egin da, baina software eta hardware hobeen sorkuntzak kostua txikiagotzen lagundu du proiektuan garrantzitsuak izango diren hitzen errepresentazioak lortzeko.

[1] https://en.wikipedia.org/wiki/Turing_test

Proiektu honetan, lehen aipatutako sare neuronal arkitektura jakin bat erabili da, *transformerra* hain zuzen ere. Baina hau azaltzen hasi aurretik bi kontzeptu azalduko dira transformer arkitekturak zertan datzan hobeto uler dadin.

Transformerren aurretik **sare neuronal errekurrenteak** erabiltzen ziren. Mota honetako sareek sarreren zatiak memorizatu eta informazio hori erabiliz aurresate zehatzak egiteko gai dira. Hitz-adiera desanbiguazio atazan esaterako, sare neuronal errekurrenteak 2.1 irudian bezala funtzionatuko luke. Hitz bakoitzaren errepresentazioa lortzeko aurrekoen informazioa erabiliko luke sare neuronal errekurrenteak. LSTMa adibidez, mota honetako sarean dago oinarrituta.



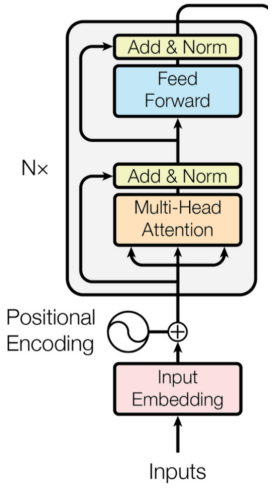
2.1 Irudia: Sare neuronal errekurrente baten egitura. $x_1, x_2, x_3...$ sarrerako esaldiko hitzak dira, $y_1, y_2, y_3...$ sarrerako hitzen errepresentazioak dira eta $s_1, s_2, s_3...$ aurreko hitzen informazioa gordetzen dutenak dira.

Transformerren aurretik azaldu beharreko beste kontzeptua **atentzioa** da. Atentzioa sarrerako zati batzuei garrantzi handiagoa eta beste zatiei txikiagoa emateko erabiltzen den teknika da. Modu honetara sarrerako hitzaren informazio bereizgarria kontzentratu egiten da. Hitz-adiera desanbiguazioan, beste era batera esanda, atentzioak agerpen berri baten aurrean esaldiko hitz garrantzitsuenengan arreta jartzeko balio du.

2.1.1 *Transformer*rak

Transformerra [Vaswani et al., 2017]-ren artikuluan aurkeztutako arkitektura berria da. Estilo honetako arkitekturek bezala, Transformerrek sekuentziatik sekuentziarako eraldaketak egiten dituzte kodetzaile eta dekodetzaileak erabiliz. Arkitektura honen berrikuntza kodetzaile eta dekodetzaileetan sare errekurrenteak ez erabiltzetik dator. Beraz, transformerra atentzio mekanismoak soilik dituen arkitektura da.

Baina zein da Transformerrek duten egitura? 2.2 irudian Transformerrek izan ohi duten egitura ikus daiteke. Irudian azaltzen dena kodetzaile bat da da, baina pilatu egin daiteke n aldiz bata bestearen gainean. Modeloaren zati garrantzitsu bat kodeketa posizionala (positional encoding 2.2 irudia) da, sare errekurrenterik ez baitu. Transformerrak kodeketa posizionala hitzen posizioak eta ordena kodetzeko erabiltzen du. Posizioa hitzen errepresentazioetan gehitzen da.



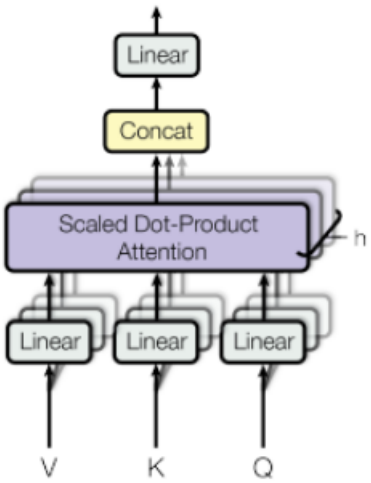
2.2 Irudia: Tranformerraren egitura [Vaswani et al., 2017]-ren artikulutik hartua. Hitzen testuinguruko errepresentazioa n geruzetan kodetzen da atentzioari esker eta hitzen errepresentazio estatikoa 1. geruzan kodetzen da.

Atentzioaren kalkulua

Atentzio mekanismoen bidez hitz bakoitzaren errepresentazioa kodetzen da gainerako hitzen informazioa kontuan hartuz. Hitz bakoitzari dagokion atentzioa kalkulatzeko lehenik eta behin hiru bektore sortzen dira, 2.3 irudian ikus daitekeen bezala:

- Galdera bektorea (**Q**)
- Gako bektorea (**K**)
- Balio bektorea (**V**)

Ondoren 2.1 ekuazioa aplikatzen da hitz bakoitzeko. Galdera bektorearen (Q) eta gako bektore (K) guztien artean biderketa eskalarra kalkulatzen da eta hau gako bektorearen



2.3 Irudia: Atentzioa kalkulatzeko egitura [Vaswani et al., 2017]-ren artikulutik hartua

dimentsioaren erro karratuaz ($\sqrt{d_k}$) zatitzen da. Lortutako bektoreari softmax funtzioa aplikatzen zaio eta amaitzeko, balio bektorearekin (V) biderketa eskalarra kalkulatzen da.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

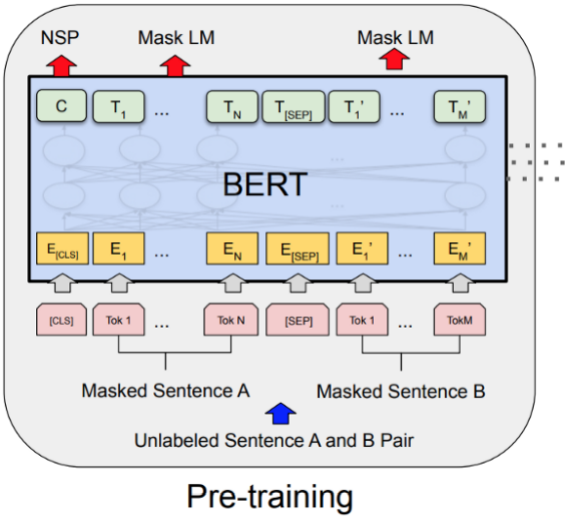
Hitz bakoitzeko bere atentzio balioa lortu ostean balio guztiak konkatenatu egiten dira eta hurrengo geruzara pasatu.

Honetaz gain, 2.3 irudian ikus daiteke atentzio mekanismoa paralelizatua izan daitekeela hainbat proiekziotan, honela, modeloak Q, K eta V-ren errepresentazio desberdinetatik ikas dezake. Hau Q, K eta V balioak pisu jakin batzuez (W) biderkatuz lortzen da. Gainera, Q, K eta V matrizeak desberdinak dira dekodetzaile bakoitzarentzat.

2.1.2 Hizkuntza eredu neuronalak: BERT

Lan honetan BERT erabili da testuingurutik hitzen errepresentazioak eskuratzeko. BERT Bidirectional Encoder Representations from Transformers-en akronimoa da Google-ek sortua. Izenak esaten duen bezala, Transformer motako arkitektura erabiltzen du eta atentzioa bi noranzkoetan (ezkerrera eta eskuinera) ezartzen du. BERT kodetzaileez (encoderrez) osatutako arkitektura da, hau da, input bezala sartutako esaldiak kodetuta itzultzen ditu hitz bakoitzeko errepresentazio bat itzuliz. Ez dauka dekodetzailerik barnean.

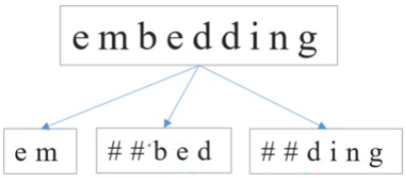
BERTen bidez hizkuntza ereduak entrenatu daitezke, ikasitako hitzen errepresentazioak, ondoren hizkuntza prozesamenduko ataza desberdinetan (hitz-adiera desanbiguazioa, galdera erantzunak, sentimendu analisia...) aplikatzeko. BERT hizkuntz errepresentazio gainbegiratu gabea da, entrenatzeko testu hutsa erabili baita. 2.4 irudian ikus daiteke entrenamendu-aurrearen egitura.



2.4 Irudia: BERTen entrenamendu-aurrearen egitura [Devlin et al., 2018]-ren artikulutik hartua. Bertan BERTek entrenamendu-aurrean input diren lehenengo esaldiaren hasieran CLS tokena eta bi esaldi banatzeko SEP tokena sartzen dituela ikus daiteke. *Output* bezala *input*-aren errepresentazio trinko bat itzultzen du.

WordPiece tokenizazioa

Tokenizazioa hizkuntza ereduen bokabulario tamaina finkatzeko erabiltzen da. Bokabulariorik erabiliko ez balitz, milioika hitz aldaera izango lirateke. BERT hizkuntza ereduak WordPiece tokenizazioa erabiltzen du eta hitzak tokenizatzeko garaian hitza azpitoken batzuetan bana dezake. BERTek 30.000 tokenetako bokabularioa dauka.



2.5 Irudia: *embedding* hitza tokenizatuta WordPiece tokenizazioa erabiliz.

Hitzen bat ez badago bere bokabularioan azpitoken batzuetan banatzen du. 2.5 irudian *embedding* hitza nola banatuko lukeen ikus daiteke. Hitza hiru azpitokenetan banatzen du. Hitzaren hasiera ez diren azpitoken guztiei bi # ikur jartzen dizkio. Azpitoken horien bi # ikurrak bokabularioko hitzen parte dira, ez dira formatuagatik jartzen zaizkion ikurrak, hots, bere esanahia dute. Hasierako azpitokena, aldiz, zegoen bezala uzten du askotan hasierako azpitokenak ez daukalako hurrengoekin zerikusirik. Adibidez, *wedding* hitza bi azpitokenetan banatuko luke, *wed* eta *##ding* azpitokenetan. *Wed* hitzak berak "ezkontza"esanahia ematen dio eta beste azpitokena aldaera bat da.

Proiektuan hitzen testuinguruko errepresentazioak lortzeko bi metodo erabili dira. Batetik, WordPieceko **azpitokenen batezbestekoa** eginez lortu da hitzaren errepresentazioa. Bestetik, WordPieceko **azken azpitokena soilik** erabili hitza errepresentatzeko.

Entrenamendu-aurrea

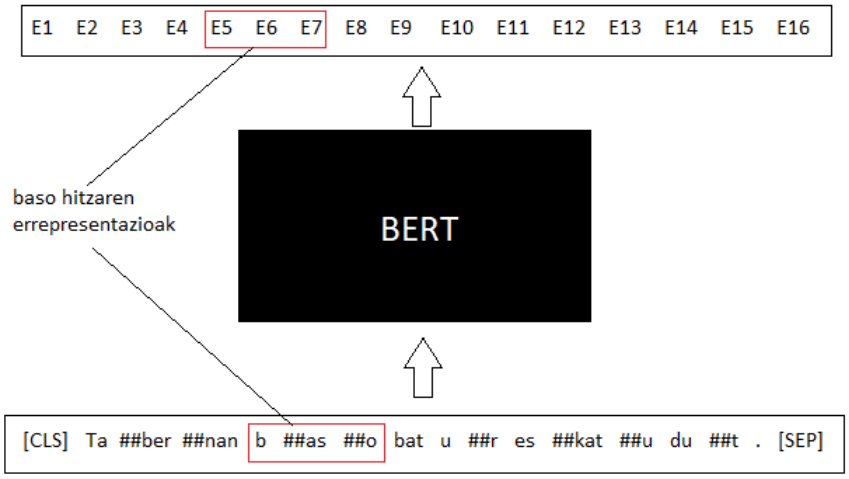
BERTek sarreran ingelesezko Wikipediako eta BookCorpuseko testu hutsa jasotzen du sekuentzietan banatuta. Entrenamendu-aurrean bi ataza desberdin burutzen dira.

Ezkutuko hizkuntza modelazioa (Masked Language Modeling): ataza honetan sekuentzia bakoitzeko hitzen %15 ezkutatu zuten ausaz. Helburua ezkutatutako hitz hauek aurresatea da. Horretarako ezkerreko eta eskuineko testuinguruak erabiltzen dira.

Hurrengo esaldiaren predikizoa (Next Sequence Prediction): hizkuntza ereduek ez dituzte jarraian datozen esaldien arteko erlazioak behar bezala kodetzen. Hau ekiditeko ikertzaileek entrenatzeko erabiltzen zituzten esaldiak binaka lotu zituzten. Binaka lotzean, bikoteen %50eko kasuetan bigarren esaldia benetan lehenaren hurrena zetorrena zen eta *IsNext* etiketa jartzen zioten. Gainerako %50ek aldiz, ez zuten zegokien hurrengo esaldia eta *NotNext* etiketa jartzen zieten.

Entrenamendu-aurrea egin ostean, BERT hizkuntza ereduak hitzen errepresentazio kontestualak ikasten ditu MLM eta NSPri esker. Proiektuan BERT hizkuntza eredua hitz bat testuinguru batean emanda hitz horren errepresentazioa kodetzeko erabili da. 1. kapituluko *baso* hitzaren adibidera itzuliz, *Tabernan baso bat ur eskatu dut.* esaldia 2.6 irudian bezala kodetzen du BERTek. Esaldia bera tokenizatzen du eta BERTek azpitoken bakoitzeko testuinguruaren araberako errepresentazio bat itzultzen du. *Baso* hitza hiru tokenetan banatzen du, 2.6 irudian gorriz markatuta daudela ikus daiteke. Hiru errepresentazio horiek osatzen dute hitzaren errepresentazioa. Proiektuan, bukaerako errepresentazioa lortzeko

bi aldaera erabili dira, azpitokenen errepresentazioen arteko batezbestekoa egitea eta azken azpitokenarenarekin gelditzea.



2.6 Irudia: *Tabernan baso bat ur eskatu dut.* esaldia BERT hizkuntza ereduak nola kodetuko lukeen erakusten duen adibidea.

2.2 Hitz-Adiera Desanbiguazioa

2.2.1 Zer da?

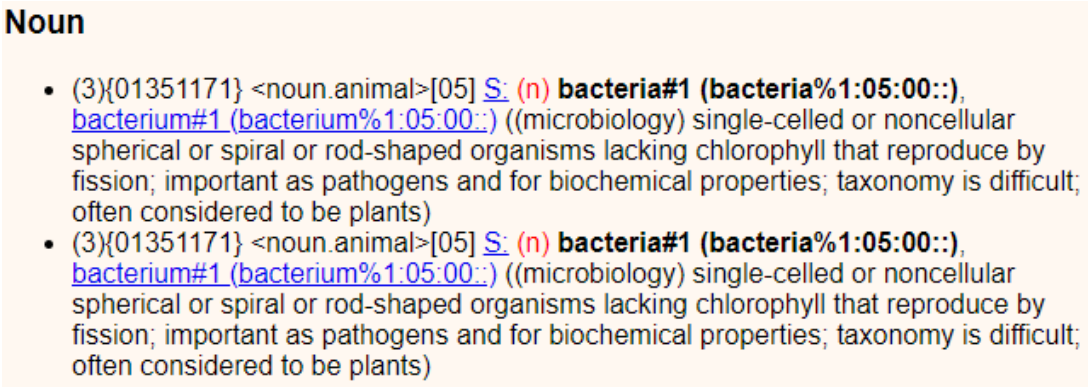
Hitz-adiera desanbiguazioa esaldi edo testuinguru batean hitzek duten esanahia edo adiera antzematean datza. Ataza hau hizkuntza prozesamenduaren barruan dagoen ataza da eta bere zailtasunak ditu.

Hitz-adiera desanbiguazioa atazak **gramatika kategoriak etiketatzearekin erlazio handia** du eta askotan bi atazak aldi berean egin ohi izaten dira. Lan honetan gramatikalki jada etiketatuta dagoen datu-multzoak erabili dira ebaluaziorako, beraz, gramatika etiketatze arazorik ez da izan.

Zailtasun horietako lehena **hiztegien arteko desberdintasuna** da. Hitz batek izan ditzakeen adiera posibleak gizakiok sortzen ditugu eta ondorioz, hitz jakin batzuentzako adiera desberdinak dituzten hiztegi desberdinak izan ditzakegu. Proiektuan WordNet-ek eskaintzen duen hiztegia erabili da.

2.2.2 WordNet

WordNet ezagutza base lexikaletik atera da proiektuan erabili den hiztegia. WordNet ingelesezko datu-multzo lexiko bat da [2] ([Oram, 2001]), hitzak sinonimo-taldeei lotuta daude adieren arabera. Datu-baseak 155,327 hitz dauzka 175,979 adieretan antolatuta. Adierak kategoria gramatikal hauetakoak izan daitezke: aditzak, adjektiboak, izenak edo aditzondoak. 2.7 irudian WordNeten adibide bat ikus daiteke.



2.7 Irudia: WordNet-en *bacteria* hitzari buruz dagoen informazioa.

Adierak erlazio semantiko desberdinen bidez lotuta daude. Sinonimia da erlaziorik garrantzitsuena, adierak berak hori adierazi nahi baitu, adiera bakoitzak kontzeptu bati egiten diolarik erreferentzia. Hala ere, beste erlazio mota batzuk ere badaude. Izenen artean hiperonimia, hiponimia, meronimia, holonimia eta termino koordenatuak ezarrita daude. Aditzen kasuan, aldiz, hiperinimia, troponimia, ondorio logikoak eta termino koordenatuak aurki daitezke. Hiperonimia, hiponimia eta erlazio semantikoa proiektuko adieren errepresentazioak sortzeko erabili dira.

Bi adieren artean adiera baten esanahia bestearena baino orokorragoa denean, orokorragoa den adiera beste adieraren **hiperonimoa** dela esaten da.

Adibidez, *altzari* adiera *lanpara* adieraren hiperonimoa da *lanpara* altzari mota bat delako, argia ematen duen altzaria hain zuzen ere.

Hiponimia, aldiz, hiperonimiaren kontrakoa da. Bi adiera emanda adiera baten esanahia bestea baino zehatzagoa denean adiera hori bestearen hiponimoa dela esaten da.

Aurreko adibidera itzuliz, *lanpara* adiera *altzari* adieraren hiponimoa izango zen lanpara

[2] http://wordnet.princeton.edu/

argia ematen duen altzari mota bat delako, hau da, *lanpara* adierak zehaztasuna ematen dio *altzari* adierari.

Bi adieren artean **erlazio semantikoa** dagoela esaten da bi adierek adierazten dituzten kontzeptuak bata bestearekin lotuta daudenean. Adibidez, *ordenagailu* adiera eta *ordenagailuko sagua* adiera lotuta daude eta bien artean erlazio semantiko bat dago.

WordNeteko adierei buruzko informazioa adieren errepresentazioak kodetzeko erabili da. Honetaz gain, WordNet proiektuan zehar erabili den hiztegia sortzeko ere erabili da.

2.2.3 Datu-multzoak

Hitz-adiera desanbiguazioa ebaluatzeko eskuz etiketatutako datu-multzoak erabiltzen dira eta gizaki bakoitzak anotatzeko irizpide desberdina izan dezake. Esaterako, Senseval-2 anotatzeko garaian anotatzaileak %85ean etorri ziren bat. Beraz, nahiz eta emaitzak asko hobetu beti izango ditu arazo hauek hitz-adiera desanbiguazioak.

Lan honetan ebaluaziorako erabili diren datu-multzoak [Raganato et al., 2017]en lanean garatutako ebaluazio esparru unifikatutik lortu dira. Lan honetan bost datu-multzo elkartu ziren hitzen adiera desanbiguaziorako ebaluazio esparru bat sortzeko, horrez gain, momentuko sistema desberdinak ebaluazio esparru honetan probatu ziren, hauen arteko konparaketa baliozkoak sortuz.

Datu-multzo guztiak ingelesez daude eta adiera guztiak WordNet 3.0 erabiliz daude anotatuak. Ebaluazio esparrua 6 datu-multzok osatzen dute, 5 desberdinak dira eta azkena aurreko bosten arteko konkatenazioa da.

- Senseval-2 [Preiss and Yarowsky, 2001]. Datu-multzo honek 2283 xede-hitz ditu. Xede-hitzen artean izenak, aditzondoak, izenondoak eta aditzak daude.
- Senseval-3 [Snyder and Palmer, 2004]. Datu-multzo honek 1850 xede-hitz ditu.
- SemEval-07 [Pradhan et al., 2007]. Datu-multzo hau txikiena da, 455 xede-hitzez osatua. Izenak eta aditzak daude.
- SemEval-13 [Navigli et al., 2013]. Datu-multzo honek 1644 xede-hitz ditu, izenak bakarrik.
- SemEval-15 [Moro and Navigli, 2015]. Datu-multzo honek 1022 xede-hitz ditu.

2.2.4 Gainontzeko baliabideak

Proiektu honetan WordNetez gain, beste baliabide batzuk ere erabili dira.

Wikipedia

Wikipedia [3] sarean dagoen entziklopedia eleanitza da. Edonork parte har dezake artikuluak argitaratuz. Irabazi asmorik gabeko erakunde batek sostengatzen du proiektua, Wikimedia Fundazioak. Wikipediako artikuluek kontzeptu zein entitate ugari deskribatzen dituzte. Wikipedia 6.000 milioi inguru hitzez osatutako corpusa da.

Lan honetan, Wikipediako orriak erabili dira adieren errepresentazioak sortzeko. Bertako ingeleseko artikuluetako testuez baliatu da kontzeptu ugariren inguruko informazioa bilduz.

Honetaz gain, BERTen entrenamendu-aurrea burutzeko corpusaren parte ere badira Wikipediako testuak, BookCorpusarekin batera.

BabelNet

BabelNet[4] [Navigli and Ponzetto, 2012] WordNet eta Wikipedia bezalako iturrietatik datorren informazioa konprimatzen duen sare semantiko eleanitza da. WordNet bezala sarea adieretan antolatuta dago eta adiera hauen artean erlazio desberdinak daude. Horretaz gain, hizkuntza desberdinetatik datozen hitzen mapaketa egiten da BabelNeten.

Lanean BabelNeteko hiru erlazio mota erabili dira adieren errepresentazioak kodetzeko garaian. Horien artean daude hiperonimia eta hiponimia erlazioa, erlazio semantikoa eta Wikipediako kontzeptuetara mapatzea.

NASARI bektore lexikoak

NASARI bektoreek [Camacho-Collados et al., 2016a] BabelNeteko kontzeptuen inguruko errepresentazioak ematen dituzte. Dimentsio bakoitza hitz batek duen espezifikazio lexikoa da ordezkatzen ari den adierarekiko. Hitz horiek adiera gehien bereizten dutenak dira. Adibidez, *sagu* hitzaren *animali* adieraren bektore lexikoak *katua, arratoia, animalia...* hitzen espezifikazio lexikoak izango ditu.

[3] https://eu.wikipedia.org
[4] http://babelnet.org

Proiektuan NASARI bektoreak adieren errepresentazioak kodetzeko erabili dira.

3. KAPITULUA

SensEmBERT adiera errepresentazioak

Hitz-adiera desanbiguazio ataza burutzeko, alde batetik, WordNeteko adieren errepresentazioak lortu behar dira. Bestetik, testuinguruko hitzei dagozkien errepresentazioak ere lortu behar dira. Errepresentazioen arteko konparaketa egin eta adiera bat aurresaten da. Proiektuaren helburuetako bat oinarritu den [Scarlini et al., 2020]-ren laneko emaitzak erreplikatzea izan da. Atal honetan, aipatutako artikuluan adieren errepresentazioak nola lortzen diren azalduko da eta eskaintzen dituen baliabideen deskribapena egingo da.

3.1 SensEmBERT adiera errepresentazioak lortzen

SensEmBERT hiztegiko hitzen adiera errepresentazioak dira eta sortzeko hainbat baliabide behar izan dituzte. Baliabide horiek Wikipedia (ezagutza base eleanitza), BabelNet (Wikipedia eta WordNet-eko informazioa konprimatzen duen sare semantikoa), NASARI bektore lexikoak (BabelNet-eko kontzeptuen errepresentazioak, erlazionatutako hitzen puntuazioak biltzen dituena) eta BERT (Transformerretan oinarritutako hizkuntza eredua) dira.

SensEmBERT ezagutzan oinarritutako adieren errepresentazio eleanitzak sortzeko saiakera da. Hitzen adiera bakoitzeko errepresentazio bat sortu da, hau da, *baso* hitzaren *edalontzi* adierak bere errepresentazioa du eta *zuhaitzez jantzitako lur-eremua* adierak beste errepresentazio bat. Gainera, *baso* hitzaren *edalontzi* adieraren errepresentazioa gertuago dago edalontzi, botila, eta abarren errepresentazioengandik eta *zuhaitzez jantzitako lur-eremua* adieraren errepresentazioa, aldiz, zuhaitz,landare, eta abarren adierenengan-

dik gertuago dago. Errepresentazio horiek sortzeko hiru pausu jarraitu dira: Wikipediatik adieren testuingurua erauzi, errepresentazio bihurtu eta WordNeteko informazioarekin hedatu.

3.1.1 Wikipediako testuinguruaren erauzpena WordNeteko adierentzat

Pausu honen helburua Wikipediatik adiera bakoitzarentzat bereizgarria den testuingurua erauztea da. Adieren testuinguruak biltzeko, adiera eta Wikipedia orrien arteko BabelNetek eskaintzen duen mapaketa erabili da.

Lehenik eta behin, s adiera jakin batekin erlazioa (hiponimia, hiperonimia edo semantikoa) duten adierak bilduko dira. Artikuluan s-rekin erlazionatutako adiera bilduma, R_s, honela definitu da. :

$$R_s = \{s' \mid (s,s') \in E\} \tag{3.1}$$

E hiponimo, hiperonimo eta erlazio semantikoen bilduma da.

R_s bildumaren egiazkotasuna areagotzeko bilduma filtratu egiten da Wikipediako orrietako informazioa erabiliz. Horrela R_s bilduman s-rekin erlazioa handia duten adierak bakarrik mantenduko dira. R_s-ko adiera bakoitzeko bere Wikipediako p_i orria hartzen da eta bere bektore lexikoa kalkulatzen da [Camacho-Collados et al., 2016b]-ren lanean egiten den moduan. Errepresentazio lexiko hauek Wikipedia orriak puntuatzeko erabiltzen dira eta horretarako Weighted Overlap (WO) neurria erabiltzen da, [Pilehvar et al., 2013]-koa. WO neurriak Wikipediako p_1 eta p_2 ren arteko antzekotasuna honela kalkulatzen du:

$$WO(p_1,p_2) = \left(\sum_{w \in O} \frac{1}{r_w^{p_1} + r_w^{p_2}}\right)\left(\sum_{i=1}^{|O|} \frac{1}{2i}\right)^{-1} \tag{3.2}$$

non O p_1 eta p_2-ren artean gainjartzen diren dimentsioen bilduma den eta $r_w^{p_i}$ p_i bektore lexikoan w hitzaren maila adierazten duen.

Ondoren, R_s-ko partaideak hiru partiziotan banatzen dira erlazioaren arabera. Hiponimoak alde batetik, hiperonimoak beste aldetik eta erlazio semantikoa dutenak, azkenik. Partizio bakoitzetik WO neurriaren arabera k altuenak hartzen dira ($k = 10$).

Honen ostean, oraindik gehiago konprimatzen da R_s bilduma. Kasu honetan R_s-ko partaide bakoitza (s') hartzen da eta R_s-ko beste partaideekin (s'') WO neurria kalkulatzen da. $WO(p_s, p_{s'}) < WO(p_{s''}, p_{s'})$ betetzen bada R_s-tik ezabatua izango da, horrela, testuinguru desberdina duten adierak gordeko dira R_s-en.

Amaitzeko, R_s bildumako adieren Wikipedia orrietako esaldiak BoC_s-en (Bag of Contexts) gordeko dira.

3.1.2 Wikipediako testuinguruetatik adieren errepresentazioetara

Pausu honen helburua BoC_s-ren barruan dauden esaldiak BERT erabiliz errepresentazio bihurtzea da. Helburu adieraren hitz garrantzitsuak bere NASARI bektorean azaltzen direnak dira.

Formalki, W_s s adieraren NASARI bektorean zero ez duten hitzen bilduma izango da. W_s bilduman dauden hitzak ez dituen BoC_s-ko esaldiak ez dira kontuan hartuko eta ondoren, $w \in W_s$ hitz bakoitzeko bere errepresentazioa kalkulatuko da BoC_s-ko agerpenetako BERTeko errepresentazioen batezbestekoa eginez. $w \in W_s$ hitz bakoitzeko bere v_w bektorea honela kalkulatuko da:

$$v_w = \frac{\sum_{c \in BOC_s^w} BERT(c,w)}{|BOC_s^w|} \tag{3.3}$$

non BoC_s^w BoC_s multzoan w hitza agertzen den testuinguru multzoa den eta $BERT(c,w)$ w hitzaren BERT errepresentazioa den c testuinguruan.

Adiera ongien bereizten duen informazioa lehenesteko W_s-ko hitzei pisuak ematen zaizkie adieraren NASARI bektorean duten mailaren arabera. Ondoren, adieraren azken errepresentazioa konputatzen da hitzen bektoreak konbinatuz. Formalki eginda, honela lortzen da adieraren errepresentazioa:

$$v_s = \frac{\sum_{w_i \in W_s} rank(w_i)^{-1} v_{w_i}}{\sum_{w_i \in W_s} rank(w_i)^{-1}} \tag{3.4}$$

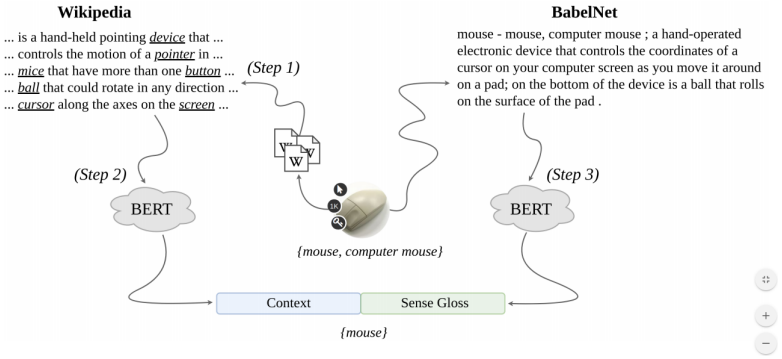
Puntu honetan, adiera bakoitzeko Wikipedian oinarritutako errepresentazio bat lortuko da.

3.1.3 Adiera errepresentazioak WordNeteko informazioarekin hedatzen

Lortu ditugun bektoreei adieren lema bakoitzari buruzko esanahiaren inguruko informazioa falta zaie. Beraz, hau lortzeko adieren lema, sinonimoak eta WordNeteko informazioa erabiliz sekuentzia bat sortzen da. Hasieran adierari dagokion lema jartzen da; jarraian, adierari lotutako sinonimoak txertatzen zaizkio lema bera eta guzti eta bukaeran, adierari dagokion esanahia (*glossa*) txertatzen zaio. Lortutako sekuentzia BERTen bidez kodetu egiten da eta token guztien batezbestekoa eginez adieraren errepresentazioa lortzen da.

3.1 irudiaren eskuinaldean ikus daiteke *mouse* adieraren *gloss*-a nola sortzen den. WordNeteko esaldiaren hasieran *mouse* adierarekin lotutako hitzak jarri zaizkio (*mouse* eta *computer mouse*) eta horren aurretik *mouse* hitza bera jarri zaio.

Amaitzeko, lehen zegoen v_s Wikipediatik erauzitako adieraren errepresentazioari WordNeteko esalditik lortutako errepresentazioa konkatenatzen zaio SENSEMBERT errepresentazioa lortuz. 3.1 irudian prozesu guztiaren azalpena ikus daiteke.



3.1 Irudia: *Mouse* hitzaren SensEmBERT$_{kb}$ adiera errepresentazioa lortzeko jarraitzen den prozesua [Scarlini et al., 2020]-ren lanetik hartua.

3.1.4 SENSEMBERT$_{sup}$

Orain arte azaldu diren SENSEMBERT errepresentazioak ezagutzan oinarritutako sistematik (KB) sortuak dira (SENSEMBERT$_{kb}$ hemendik aurrera), Wikipedia eta WordNetetik hain zuzen ere, baina proiektu honetan adiera errepresentazio horiez gain SENSEMBERT$_{sup}$ errepresentazioak ere erabili dira.

Adiera errepresentazio hauek lortzeko prozesuaren bukaeran aldaketa bat dago. v_s errepresentazioa eta WordNetetik lortutako errepresentazioa konkatenatu beharrean bien arteko batezbestekoa egiten da SENSEMBERT$_{sup}$ errepresentazioaren lehenengo zatia lortuz.

Bigarren zatia lortzeko SemCor [Miller et al., 1993] jada anotatuta dagoen testuingurua erabiltzen da. SemCor WordNeteko adierekin eskuz anotatutako 40K esaldiz osatutako korpusa da. Esaldi horietatik hitz-adiera (w,s) bikote bakoitzeko w hitza s adierarekin azaltzen diren $c_1,...c_n$ esaldi guztiak aukeratzen dira. BERTekin w hitzak c_i esaldian duen errepresentazioa $BERT(c_1,w),...,BERT(c_n,w)$ ateratzen da, esaldi guztietarako prozesua errepikatuz. SENSEMBERT$_{sup}$ adieren errepresentazioen bigarren zatia $BERT(c_1,w)$, $...,BERT(c_n,w)$ errepresentazioen batezbestekoa kalkulatuz lortuko da. Adieraren bat ez

bada SemCorreko esaldietan agertzen, bigarren zati hori WordNetetik lortutako errepresentazioak ordezkatuko du.

3.2 Baliabidearen deskribapena

SensEmBERT adiera errepresentazioak izenentzat bakarrik daude, NASARI bektoreak behar baitira adiera errepresentazioak sortzeko eta hauek izenek bakarrik baitituzte. Denera 146312 adiera errepresentazio daude eta adiera errepresentazio bakoitzak 2048 dimentsio dauzka. Aurreko atalean aipatu den bezala, 1024 dimentsio Wikipediatik erauzitako informaziotik lortu dira eta beste 1024 dimentsioak WordNetetik.

4. KAPITULUA

Diseinua eta inplementazioa

Kapitulu honetan diseinuan hartu diren erabakiak beraien justifikazioekin, inplementazio garaian egin den lana eta emaitzak lortzeko hartu diren erabakiak azalduko dira. Proiektuan zehar bi metodo desberdin garatu dira, hiztegiduna eta hiztegi gabea. Bi metodo hauek hasieratik bukaerara azalduko dira eta bukaeran emaitzak ebaluatzeko erabili den teknika azalduko da.

4.1 Diseinua

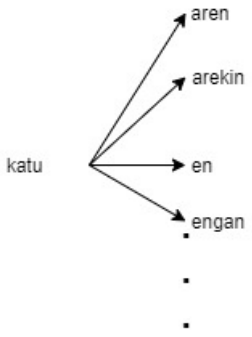
Orain arte SensEmBERT hizkuntza errepresentazioaz baliabide gisa hitz egin da, baina [Scarlini et al., 2020]-ren lanean hitz-adiera desanbiguazio ataza burutzeko baliabideaz gain sistema bat ere garatu da. Sistema horrek hitz baten testuinguruko errepresentazio bat emanda hitz horri dagozkion kandidatuak bilatzen ditu SensEmBERT adiera errepresentazioen artean. Bilaketarako WordNetek eskaintzen duen hiztegia erabiltzen du. Ondoren, hitzaren testuinguru errepresentazioa kandidatuen errepresentazioekin konparatzen ditu neurri baten bidez, kosinu antzekotasuna hain zuzen ere. Kosinu antzekotasuna bi errepresentazioen arteko angeluan oinarritzen da eta -1 eta 1 artean balio bat itzultzen du. Zenbat eta balio handiagoa izan, orduan eta antzekoagoak dira bi errepresentazioak. Amaitzeko, kosinu antzekotasun handiena (K-NN) duena aukeratzen du eta adiera hori esleitzen dio hitzari.

Proiektu honetan ordea, sistema hori berrinplementatzeaz gain bi esperimentu berri ere egin dira. Lehenengo esperimentua kandidatuak filtratzeko hiztegirik ez erabiltzea izan

da eta horretarako metodo berri bat garatu behar izan da, hiztegi gabea. Bigarren esperimentuan errepresentazioak konparatzeko K-NNa egin beharrean CSLS neurria erabili da.

4.1.1 Hiztegi gabearen justifikazioa

Hiztegien erabilera ahalbidetzen duena hitzen lemak eskuragarri izatea da. Hizkuntza batzuetan hitzen lematizazio prozesua zaila izan daiteke. Euskararen kasuan lematizazioa arazoa da hitz batek deklinabide atzizki asko izan ditzakeelako bata bestearen atzetik. Horren ondorioz, hitz bat era askotara idatz daiteke eta zaila da bere lema zein den identifikatzea. 4.1 irudian ikus daiteke adibide txiki bat.



4.1 Irudia: *katu* hitzak izan ditzakeen atzizkien adibidea

Aurretik aipatu den moduan lan honetan erabili den hiztegia WordNetetik sortu da eta hitzak desanbiguatzeko garaian bi metodo desberdinez baliatu da, hiztegiduna eta hiztegi gabea, hain zuzen ere.

Metodo hiztegidunean adiera posible guztietatik adiera kandidatuak filtratzeko hiztegi bat erabili da, horrela asmatzea errazagoa da, baina aipatu behar da hiztegiaren menpe dagoela metodo hau, hiztegian ez badago hitzaren adiera jakin bat, ezingo baita adiera hori aurresatea lortu.

Metodo Hiztegi gabean, aldiz, kandidatu posibleak bilatu beharrean adiera posible guztien artean bat aukeratu da. Hau hiztegia lortzeko aukerarik izango ez balitz, sistemak zenbat sufrituko lukeen ikusteko egin da, kasu batzuetan hiztegia ezin baita erabili, adibidez, lematizazioa arazoa denean.

4.1.2 CSLS

Orain arte K-NN sailkatzailea erabili da hitzaren adiera aurresateko hiztegi gabeko metodoan, baina K-NN sailkatzailea asimetrikoa da izatez. K-NNak x hitzarentzat y adiera aurresaten duenean, ez du esan nahi x hitzaren adieren artean y adiera dagoenik, horregatik da asimetrikoa. Dimentsio askotako espazioetan K-NN sailkatzailea erabiltzerakoan honek arazo bat sor dezake, hau da, bektore batzuek probabilitate handiagoa dute auzokide gertukoena izateko beste batzuek baino. Bektore hauei *hubs* deitzen zaie eta probabilitate txikia dutenei *anti-hubs*.

Hub hauen eragina txikiagotzearren, K-NN ordez CSLS erabiltzea erabaki da, eman ditzakeen emaitzak hobeak izan daitezkeelako. [Lample et al., 2018]-ren lanetik hartu da neurri berria, hona hemen CSLS kalkulatzeko azalpena:

$$CSLS(W_{x_s}, y_t) = 2cos(W_{x_s}, y_t) - r_t(W_{x_s}) - r_s(y_t) \tag{4.1}$$

x hitzaren eta y adieraren arteko CSLS neurria kalkulatzeko, lehenik eta behin, hitzaren testuinguru errepresentazioaren eta adiera errepresentazioaren arteko kosinua kalkulatzen da eta hau bikoiztu ($2cos(W_{x_s}, y_t)$). Kosinuaren bikoitzari bi balio kentzen zaizkio hurrena. Lehena, x hitzaren K adiera gertuenen kosinuen batezbestekoa da eta bigarrena, y adieraren K adiera gertuenen kosinuen batezbestekoa.

4.2 Hiztegidun metodoa

Aurretik aipatu bezala emaitzak lortzeko bi metodo desberdin inplementatu dira lan honetan. Bi metodoen arteko desberdintasun nagusia hiztegidun metodoan hiztegi batez baliatu dela SensEmBERT adiera errepresentazio guztietatik kandidatu posibleak aukeratzeko izan da; hiztegi gabekoan, berriz, ez da dira kandidatuak filtratu eta adiera errepresentazio guztiekin egin da konparaketa.

Hona hemen hiztegidun metodoak aurresateak egiteko jarraitzen duen prozesuaren sasi-

kodea: 1

Data: hiztegia, sensEmbeds, sarreraEsaldia, xedeHitza
Result: agerpen berriari aurresaten zaio adiera
xedeHitzarenAdierak = lortuAdierak(hiztegia,xedeHitza)
kandEmbeds = aukeratuKandidatoak(sensEmbeds,xedeHitzarenAdierak)
testEmbeds = BERT(sarreraEsaldia)
testEmbed = agerpenarenTestEmbedLortu(testEmbeds,xedeHitza)
aurresatea = knn(testEmbed,kandEmbeds)
return aurresatea

Algorithm 1: Hiztegidun metodoaren sasikodea

Adibidez, hiztegidun metodoan sarrera *"The dog is in the garden"* esaldia bada eta desanbiguatu nahi den hitza *dog*, hau izango litzateke burutuko zatekeen prozesua:

- Lehenik eta behin, *dog* hitzaren adierak filtratuko lirateke hiztegiaz baliatuz eta SensEmBERT errepresentazioetatik kandidatu posibleak lortuz.
- Ondoren, sarrerako esaldia BERT ereduari pasatuko litzaioke eta hemendik *dog* hitzari dagokion testuinguruko errepresentazioa hartu beharko litzateke.
- Azkenik, *dog*-en adieren errepresentazioak testuingurutik lortutako errepresentazioarekin konparatu beharko lirateke eta kosinu antzekotasun altuena duena aukeratu kandidatu posibleetatik.

Sarrera fitxategiek esaldi ugari dituzte eta esaldi batek xede-hitz (*target*) bat baino gehiago izan ditzake. Xede-hitz bakoitzeko aurreko prozesua behin eta berriz errepikatu behar da instantziei adierak aurresateko.

4.2.1 Hiztegia

Hiztegiaren erabilerarekin kandidatu posibleak filtratzeko aukera lortzen da. Baina hiztegia erabili ahal izateko sarrera fitxategiko hitzak lematizatuta eduki behar dira edo lematizazio prozesu bat egin beharra dago. Lan honetan ebaluaziorako erabili diren datumultzoak (WSD Evaluation Framework) jada lematizatuta ditu hitz guztiak, beraz, alde horretatik ez da arazorik izan.

Kandidatuak filtratuta bi hobekuntza desberdin lortzen dira. Batetik, denbora gutxiago beharko da aurresateak egiteko, kosinu antzekotasunen konparaketak multzo txikiago baten gainean egiten baitira. Bestetik, aurresatea multzo txikiago baten gainean egiten de-

nez, aurresaten den adiera egokia izateko aukera gehiago egongo dira eta emaitza hobeak lortuko dira ondorioz.

4.2.2 Sarrera fitxategia

Aurreko atalean aipatu bezala, WSD Evaluation Framework [Raganato et al., 2017]-tik lortu dira sarrera fitxategiak. Fitxategiak kargatzean esaldiak, xede-hitzak eta xede-hitzen indizeak kargatzen dira. Xede-hitzen indizeak beharrezkoak dira esaldiak tokenizatzeko garaian hitzak token askotan banatu daitezkeelako eta hauen hasierako eta bukaerako indizeak gorde behar dira, ondoren adiera errepresentazioak behar bezala aukeratu ahal izateko.

Sarrera fitxategitik kategoria gramatikala izena duten xede-hitzak bakarrik hartu dira SensEmBERT fitxategian izenen adieren errepresentazioak bakarrik daudelako. Hau egiteko arazorik ez da izan [Raganato et al., 2017]-ren lanean sortutako datu-multzoek postagging ataza egina dutelako dagoeneko.

4.2.3 SensEmBERT adiera errepresentazioak

Hiztegidun metodoan ez dira SensEmBERT adiera errepresentazio guztiak kargatzen, xede-hitzak aurresateko beharrezkoak direnak bakarrik baizik. SensEmBERT adiera errepresentazioak kargatzeko hiztegitik lortutako adieren lista erabiltzen da, hau da, hiztegitik kandidatu guztien IDak biltzen dira eta ID horien errepresentazioak bakarrik kargatzen dira 1. sasikodeko *sensEmbeds* parametroan.

4.2.4 Testuingurutik lortutako errepresentazioak

Testuingurutik hitzen errepresentazioak lortzeko sarrera fitxategitik kargatu diren esaldiak BERT hizkuntza ereduari pasatu behar zaizkio. BERTek esaldika egiten du lana, hau da, sartutako esaldi bakoitzeko erantzun independente bat itzultzen du. Hitzaren errepresentazioa eskuratzeko esaldi horren errepresentaziotik hitz jakin horri dagokion errepresentazioa aurkitu behar da, esaldi guztia hartzeak ez baitu balio.

Xede-hitzari dagokion testuinguru errepresentazioak aurkitzean bi aukera daude: xede-hitza azpitoken batzuetan banatuta egotea tokenizazioaren ondorioz eta xede-hitzaren errepresentazioa bakarra izatea. Bigarren kasuan xede-hitzari dagokion errepresentazioa

hartzearekin nahikoa da, baina lehengoan aukera bat baino gehiago daude. Proiektuan aztertu diren aukera horiek **azpitokenen arteko batezbestekoa** egitea eta **azpitokenen artean azkena** hartzea izan dira.

4.2.5 Aurresateak

Behin testuingurutik ateratako hitzen errepresentazioak eta kandidatuak lortu eta gero, xede-hitzen aurresateekin hasten da. Xede-hitz bakoitzari dagozkion kandidatuen errepresentazioak aukeratzen dira kargatutako SensEmBERT adiera errepresentazioetatik. Aurresatea K-NN sailkatzailearen bidez egiten da, hitzen testuinguru errepresentazioak kandidatuen errepresentazioekin konparatuz. Kosino antzekotasun altuena duen kandidatua aukeratzen da aurresate bezala.

Aurresate guztiak egitean hauek itzuli eta fitxategi batean idazten dira urre patroian (*gold standard*) bezala ondoren lortutako emaitzak ebaluatu ahal izateko.

4.3 Hiztegi gabeko metodoa

Hiztegi gabeko metodoa sinpleagoa da, ez baitago hiztegia erabili beharrik kandidatuak lortzeko. 2. sasikodean ikus daiteke kodea motzagoa dela, baina sarrera fitxategia, tokenizazioa eta aurresateko sailkatzailea (K-NN) hiztegidun metodoan erabiltzen den bera da. Aldatzen den bakarra aurresateko garaian kandidatuen SensEmBERT adiera errepresentazioekin bakarrik konparatu beharrean SensEmBERT adiera errepresentazio guztiekin konparatzen dela da.

Data: sensEmbeds, sarreraEsaldia, agerpenBerria
Result: agerpen berriari aurresaten zaio adiera
testEmbeds = BERT(sarreraEsaldia)
testEmbed = agerpenarenTestEmbedLortu(testEmbeds,agerpenBerria)
aurresatea = knn(testEmbed,sensEmbeds)
return aurresatea

Algorithm 2: Hiztegi gabeko metodoaren sasikodea

Aurreko adibidera itzuliz hiztegi gabeko metodoan jarraitu beharreko metodoa ondorengoa izango litzateke:

- Lehenik eta behin, BERTetik pasatu beharko litzateke *"The dog is in the garden"* esaldia eta *dog* hitzari legokion testuinguru errepresentazioa hartu.

- Ondoren, testuingurutik lortutako errepresentazioa SensEmBERT errepresentazio guztiekin konparatu beharko litzateke eta kosinu antzekotasun altuena duena aukeratu.

Lortutako emaitzen tratamendua hiztegidun metodoan egiten denaren berdina da. Emaitzak okerragoak izango dira orokorrean, baina hiztegirik ez badago edo instantzien lematizazioa egiteko baliabiderik ez badago, ez da beste aukerarik geratzen eta hiztegi gabeko metodoa erabili beharra dago.

4.4 Emaitzen ebaluazioa

[Raganato et al., 2017]-ren laneko datu-multzoetatik lortutako emaitzak ebaluatu edo puntuatzeko hiru neurri erabili dira. Neurri hauek zehaztasuna (P), estaldura (R) eta F1 neurria izan dira. Hiru neurri hauek kalkulatzeko ebaluazio esparruak berak eskaintzen duen baliabidea erabili da, *scorer*ra hain zuzen ere.

*Scorer*ra java lengoaian eginiko sistema puntuatzaile bat da, urre patroia eta emaitzen fitxategia pasatuta zehaztasuna, estaldura eta F1 neurria inprimatzen dituena.

		Aurreikuspena	
		Negatiboa	Positiboa
Egia	Negatiboa	Negatibo egiazkoa	Positibo faltsua
	Positiboa	Negatibo faltsua	Positibo egiazkoa

4.2 Irudia: Zehaztasuna, estaldura eta F1 azaltzeko negatibo eta positibo taula.

4.4.1 Zehaztasuna

Ebaluazio esparru bakoitzean aurreikusi beharreko xede-hitz asko daude eta kasu batzuetan aurreikusi nahi den xede-hitza berria ezezaguna da. Kasu horietan baliteke aurreikuspenik ez itzultzea, zehaztasunarekin adierazi nahi dena aurreikuspena itzuli den kasuetan

zenbat asmatu diren da kasu ezezagunak kontuan hartu gabe. 4.2 irudiko terminoak erabiliz hau izango litzateke zehaztasuna kalkulatzeko formula:

$$Zehaztasuna = \frac{Positibo\ egiazkoa}{Positibo\ egiazkoa + Positibo\ faltsua} \tag{4.2}$$

4.4.2 Estaldura

Estaldurak zehaztasunak kontuan hartzen ez dituen xede-hitz ezezagunak aintzat hartzen ditu, hau da, estaldurak adierazten duena xede-hitz guztietatik zenbat asmatu diren da. Ondorioz, estaldurak ezezagunak ez diren kasuak ere kontuan hartuko ditu. Hona hemen 4.2 irudian oinarrituz estalduraren formula:

$$Esataldura = \frac{Positibo\ egiazkoa}{Positibo\ egiazkoa + Negatibo\ faltsua} \tag{4.3}$$

4.4.3 F1 neurria

Estaldura eta zehaztasuna ez dira neurri berdinak, baina bi neurri hauen oreka bilatu nahi bada, F1 neurria erabiltzen da. F1 neurria estaldura eta zehaztasunaren artean sortzen da eta bi neurriek pisu berdina dute F1 sortzeko garaian. Beraz, F1 altua bada, estaldura eta zehaztasunak ere altuak izango dira. Hona hemen F1 neurriaren formula:

$$F1 = 2 * \frac{Zehaztasuna * Estaldura}{Zehaztasuna + Estaldura} \tag{4.4}$$

4.2.2. atalean aipatu da sarrera fitxategietatik izenak diren xede-hitzak bakarrik hartu direla kontuan. Emaitzak ebaluatzeko garaian ere gauza bera egin da, hau da, urre patroia filtratu egin da izenei dagozkien instantzien emaitzak dituen fitxategi berri bat sortuz ebaluaziorako datu-multzo bakoitzeko.

5. KAPITULUA

Esperimentuak eta emaitzak

Kapitulu honetan garatu den sistemarekin eginiko esperimentuak zein izan diren eta esperimentuak egiteko erabili diren ezarpenak azalduko dira.

Hau amaitzean egin diren esperimentuak zein izan diren eta hauen emaitzak erakutsiko dira. Ondoren lortutako emaitzen analisia egingo da.

5.1 Esperimentuko ezarpenak

Lanaren lehenengo esperimentua [Scarlini et al., 2020]-ren artikuluan egiten den esperimentua erreplikatzea da, beraz, bertan erabiltzen diren ezarpen berdinak erabili dira.

5.1.1 BERTen ezarpenak

Esperimentuan BERTen 1024 dimentsiotako *cased* ingeleseko modelo aurre entrenatua erabili da, *original word masking*-a egiten duena. Hitzen testuinguru errepresentazioak lortzeko azken lau geruzen arteko batura egin da, sekuentziaren luzera maximoa 128koa jarri da, *batch*-aren tamaina 32 jarri da eta hitzak ez dira minuskuletara pasatu, hau da, $do\_lower\_case = False$ jarri da.

5.2 Emaitzak

5.2.1 Batezbestekoa vs azken tokena

Esan bezala, esperimentazioan hitzen testuinguru errepresentazioak bi eratara sortu dira: azpitoken guztien batezbestekoa eginez eta azken azpitokena hartuz. Esperimentu honetan hiru emaitza konparatu dira. Alde batetik, [Scarlini et al., 2020]-ek argitaratutako laneko emaitzak hartu dira; bestetik, lan honetan eginiko bi esperimentuen emaitzak ere hartu dira.

Datu-multzoa	[Scarlini et al., 2020]-ren emaitzak	Batezbestekoa	Azken azpitokena
Senseval-2	**%83.7**	%83.2	%83.2
Senseval-3	%79.7	%79.7	**%80.1**
Semeval-2007	**%79.9**	%78.6	%78.6
Semeval-2013	**%78.7**	%78.3	%77.9
Semeval-2015	%80.2	%80.2	**%81.0**
ALL	**%80.4**	%80.1	%80.1

5.1 Taula: F1 neurrian konparaketak [Raganato et al., 2017]-ren ingeleseko WSD datu-multzoetako kategoria gramatikala izena duten xede-hitzak bakarrik hartuz. ALL datu-multzoa aurreko guztien baturatik sortua da, azken finean aurreko emaitza guztien arteko batezbestekoa da.

5.1 taulako emaitzak ikusiz [Scarlini et al., 2020]-ren laneko sistema berrinplementatzea lortu dela esan daiteke. Lortu diren emaitzak [Scarlini et al., 2020]-renak baina pixka bat txarragoak dira, dezima batzuen aldea dago. Honetaz gain, hitzaren testuinguru errepresentazioak lortzeko garaian subtokenen batezbestekoa edo azken subtokena erabiltzeak ez dakar aldaketa handirik. Emaitzak oso parekoak dira batean edo bestean. Datu-multzo batzuetan batek emaitza hobeak ematen ditu besteak baino eta alderantziz, baina datu-multzo guztien bilduran (ALL) emaitza berdina itzultzen dute.

Hemendik aurrerako esperimentuetan hitzen testuinguruak sortzeko azpitokenen batezbestekoa egin da [Scarlini et al., 2020]-ren lanean metodo hau erabiltzen baita eta aztertu diren bi metodoen artean ez baitago ezberdintasunik.

Datu-multzoa	SensEmBERT$_{sup}$		SensEmBERT$_{kb}$	
	[Scarlini et al., 2020]	Gureak	[Scarlini et al., 2020]	Gureak
Senseval-2	**%83.7**	%83.2	%80.6	%81.1
Senseval-3	**%79.7**	%79.7	%70.3	%70.6
Semeval-2007	**%79.9**	%78.6	%73.6	%73.0
Semeval-2013	**%78.7**	%78.3	%74.8	%74.6
Semeval-2015	%80.2	%80.2	%80.2	**%80.6**
ALL	**%80.4**	%80.1	%75.9	%76.1

5.2 Taula: F1 neurrian konparaketak sistema gainbegiratuaren eta gainbegiratu gabearen artean [Raganato et al., 2017]-ren ingeleseko WSD datu-multzoetako kategoria gramatikala izena duten xede-hitzak bakarrik hartuz.

5.2.2 SensEmBERT$_{kb}$ vs SensEmBERT$_{sup}$

[Scarlini et al., 2020]-ren lanean bi SensEmBERT eredu sortu dira, ezagutzan soilik oinarritutakoak (SensEmBERT$_{kb}$) eta horrez gain anotatutako datuak erabiliz sortutakoak (SensEmBERT$_{sup}$). [Raganato et al., 2017]-ren laneko ebaluazio esparruan bi hizkuntza errepresentazio eredu desberdin horiek probatu dira eta lortutako emaitzen arteko konparaketak egin dira esperimentu honetan. 5.2 taulan ikus daitezkeen emaitzetatik atera daitekeen ondorioa SensEmBERT$_{sup}$ adiera errepresentazioak erabilita orokorrean emaitza hobeak lortzen direla da. Hala ere, Semeval-2015 datu-multzoan ezagutzan soilik oinarritutako adiera errepresentazioek emaitza hobea eman dute eta bien arteko diferentzia ALL datu-multzoan (datu-multzo guztien bilduran) %5ekoa da. Beraz, hemendik ondoriozta daiteke SemCorreko datu anotatuak erabiliz emaitza hobeak lortzen direla.

Proiektuan lortu diren emaitzak SensEmBERT$_{sup}$ adiera errepresentazioak erabiliz dezima batzuk txarragoak dira [Scarlini et al., 2020]-ren emaitzekin konparatuz. SensEmBERT$_{kb}$ adiera errepresentazioak erabiliz, aldiz, proiektuan emaitza hobeagoak lortu dira.

5.3 Hiztegi gabearen analisia

Orain arte egin diren esperimentu guztietan hiztegiaz baliatu da adiera posibleak filtratzeko. Hemendik aurrera hiztegi gabeko metodoarekin egin dira proba berriak. Kanpotik begiratuta hiztegi gabeko metodoa probatzeak zentzugabekeria dirudi, hiztegia erabiltzeak lana aurrezten baitu. Zergatik utzi hiztegiaren erabilera alde batera emaitza desegokiak filtratzen laguntzen baldin badu eta prozesua azkartzen baldin badu?

Esperimentu hauen helburua hiztegia erabili gabe sistemarengan zenbateko eragina duen

ezagutzea da. Horretarako hainbat proba desberdin egin dira aurrerago hobeto azalduko direnak.

5.3.1 SensEmBERT hiztegiarekin vs hiztegirik gabe

Atal honetan hiztegidun metodoaren eta hiztegi gabearen arteko konparaketak egin dira. Hiztegidun metodoan hiztegiaz baliatu da adiera kandidatuak aukeratzeko eta hiztegi gabean SensEmBERT adiera errepresentazio guztiekin konparatu dira hitzen testuinguru errepresentazioak.

Datu-multzoa	Hiztegiarekin		Hiztegi gabe	
	CSLS	KNN	CSLS	KNN
Senseval-2	%82.8	**%83.2**	%52.0	%44.2
Senseval-3	%79.1	**%79.7**	%57.1	%50.0
Semeval-2007	%78.6	**%78.6**	%53.5	%48.4
Semeval-2013	%77.9	**%78.3**	%44.6	%38.5
Semeval-2015	%78.9	**%80.2**	%49.9	%42.6
ALL	%79.5	**%80.1**	%49.9	%43.2

5.3 Taula: F1 neurrian konparaketak hiztegidun metodoaren eta hiztegi gabearen artean [Raganato et al., 2017]-ren ingeleseko WSD datu-multzoetatik kategoria gramatikala izena duten xede-hitzak bakarrik hartuz. CSLS neurria KNN-arekin ere konparatzen da taula honetan bi metodoetan.

5.3 taulan ikus daitezke lortu diren emaitzak. Hiztegiarekin lortu diren emaitzak ia-ia bi aldiz hobeak dira hiztegi gabekoekin konparatuz gero. Beraz, emaitza horietatik hiztegiaren erabilerak abantaila handia dakarrela ondoriozta daiteke.

Hiztegirik gabeko metodoan xede-hitzen adiera aurresateko garaian baliteke xede-hitzari ez dagokion adiera bat aurresatea adiera errepresentazio guztiekin konparatzen baita. Adibidez, Senseval2 datu-multzoan *tower* hitza 5 alditan agertzen da xede-hitz bezala eta hiztegi gabe 5 horietan *clock_tower%1:06:00::* adiera aurresaten du sistemak. Ez du asmatzen beste hitz bati dagokion adiera baita aurresaten duena. Hiztegia erabiliz, aldiz, 5 kasuetan *tower%1:06:00::* adiera aurresaten du eta asmatu egiten du. Hiztegia ez erabiltzeak kandidato askoz gehiago izatea dakar. Adibide honetan esaterako, kandidato posibleak SensEmBERT-eko adiera guztiak izatetik hiru (*tower* hitzak 3 adiera posible dauzka WordNet-en izenak bakarrik kontuan hartuz) bakarrik izatera pasatuko bailitzateke.

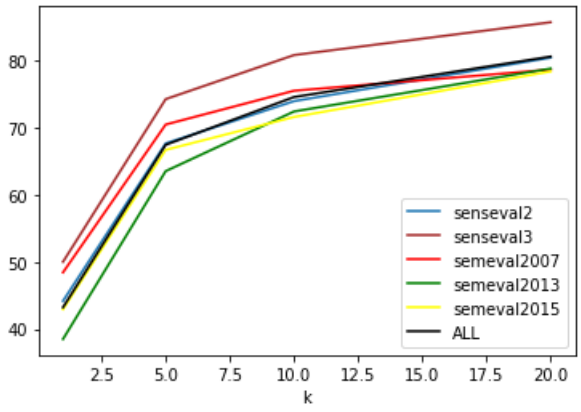
5.3.2 KNN ordez CSLS

5.3 taulan ikus daitezke CSLS neurria erabiliz lortu diren emaitzak. Hiztegirik gabe emaitzak asko hobetu dira, %7 inguru. Hobekuntza lehen aipatutako *hub* horiek bereizi direlako lortu da. Esaterako, Senseval2 datu-multzoan askotan aurresaten den adiera *kirk%1:06:00::* da. Baina CSLS neurria erabiliz KNN-aren ordez, horrelako adierak ez aurresatea lortu da.

Hiztegiarekin, aldiz, emaitza txarragoak lortu dira. Hiztegia erabiltzearekin kandidatuak filtratu egiten dira *hubness* handiko adiera errepresentazioak baztertuz. Horregatik ez dira hobetu KNN-an lortutako emaitzak.

5.3.3 Kosinu altuenen artean adiera egokia bilatzen (1, 5, 10 eta 20)

Esperimentu honetan kosinu altuena duen adiera aukeratu beharrean, 1, 5, 10 eta 20 altuenen artean bilatu da sistemak aukera gehiago aztertuta emaitzak zenbateraino hobetu ditzakeen ikusteko. Proba guztiak hiztegirik gabe egin dira. 5.1 irudiko grafikoan ikus daitezke esperimentuak eman dituen emaitzak.



5.1 Irudia: Hiztegirik gabeko metodoarekin 1, 5, 10 eta 20 kosinu altuenak aukeratuz [Raganato et al., 2017]-ren lanean sortutako datu-multzoetan lortzen diren F1 emaitzen grafika

5.1 irudiko grafikoan ikus daiteke bakarra aukeratuz emaitzak %40 – %50 tartean daudela. 5 altuenak aukeratuz emaitzak %65 – %75 tartera hobetzen dira . 10ekin hobekuntza txikiagoa da, baina hala ere, emaitzak %70 – %80 tartera igo dira eta 20 altuenak aukeratuz ,berriz, %80 – %85 tartera.

Kontuan hartuz hitzaren testuinguru errepresentazio bakoitza 146312 adiera errepresentaziorekin konparatzen dela eta 20 altuenak hartuz emaitzak %80 inguruan dabiltzala sistema ona dela esan daiteke.

6. KAPITULUA

Ondorioak eta etorkizuneko lanak

Kapitulu honetan proiektuan zehar egin den lanaren ondorioak eta ondorio pertsonalak azalduko dira. Azkenik, proiektuak irekita utzi dituen ateen inguruko hausnarketa egingo da.

6.1 Ondorioak

Proiektuan hitz-adiera desanbiguazio ataza burutuko duen sistema bat berrinplementatzea lortu da. Hortik abiatuz, ondoren, [Scarlini et al., 2020]-ren laneko esperimentuak errepli-katzea lortu da eta esperimentu berri batzuk egin dira. Lortutako emaitzak beraien artean konparatu dira.

6.1.1 Proiektuaren ondorioak

Proiektuan garatu den sistemari aldaera batzuk egin zaizkio eta hiztegia erabili gabe beste azterketa desberdin batzuk egin dira hiztegirik gabe sistemak zenbateraino sufritzen duen jakiteko. Hitzen lemak lortzeko ez da arazorik izan lanean erabili diren datu-multzoak [Raganato et al., 2017] anotatuak direlako; beraz, hiztegiaren erabileran ez da arazorik izan.

Proiektuan hitz-adiera desanbiguazioa gauzatzeko bi metodo desberdin garatu dira hiztegiduna eta hiztegi gabea. Hiztegidun metodoarekin eta [Scarlini et al., 2020]-ren laneko

adiera errepresentazioekin, [Raganato et al., 2017]-ren datu-multzoetan emaitza oso onak lortu dira. Emaitza onak izateak ateak irekita uzten ditu sistema honekin etorkizunean hitz-adiera desanbiguaziorako proiekturen bat egiteko.

Honetaz gain, hiztegi gabeko metodoan errepresentazioak konparatzeko KNN-a erabili ordez CSLS neurria erabiltzeak emaitzak hobetzen dituela ikusi da. Kosinua soilik erabilita agerpen berriaren errepresentazioren antzekoena bilatzen da, baina askotan antzekotasun hori ez da simetrikoa. CSLSak antzekotasuna simetrikoa izatea bilatzen du eta hiztegi gabeko metodoan emaitzak dezente hobetu ditu. Hiztegidunean, aldiz, eragina ez da positiboa izan konparazioak errepresentazio gutxi batzuen artean egiten baitira.

6.1.2 Ondorio pertsonalak

Maila pertsonalean esperientzia aberasgarria izan da, tamaina honetako proiektu baten garapenerako lan handia baitago eta gauza asko ikasteko balio izan baitu arlo desberdinen inguruan. Ikasketa prozesu etengabea izan da proiektu honen garapena.

Graduko hainbat irakasgaitan ikasitako kontzeptu ugari oso baliagarriak izan zaizkit proiektuaren garapenean. Adibidez, *Hizkuntza Prozesamendua* irakasgaian landu nuen hitz-adiera desanbiguazioa oso baliagarria izan zait zerotik ez hasteko, jada oinarri bat baineukan. Horretaz gain, antolakuntzarako *Proiektu Kudeaketa* irakasgaian ikasitako teknikak ere aplikatu ditut proiektu honetan eta *Estatistikan* zein *Datu Meatzaritzan* ikasitako kontzeptu asko ere erabili ditut.

Sare neuronalen artearen egoerari buruz ere asko ikasi dut. Aurretik irakasgai askotan landu nituen sare neuronalak eta banekien zertarako erabiltzen ziren, baina behin ere ez nituen erabili proiektu batean. Lan honetan transformer arkitekturako sare bat erabili dut eta hauen mugak zein diren eta zenbateraino laguntzen duten ikasi dut.

Proiektuan zehar arazo asko izan ditut berrinplementazioa gauzatzeko garaian eta ikerkuntzaren alde txarra ezagutu dut, asteak joan eta etorri emaitzak ez baitziren hobetzen. Gogorra izan da, baina esperientzia positibotzat jotzen dut lana eginez arazoak konpondu daitezkeela ikasi baitut.

Gainera, ikerkuntzarekin lotutako proiektuetan zaila izaten da lortuko diren emaitzak aurreikustea eta esperimentu ugari alperrik egin daitezke. Beraz, proiektuaren aurreko planifikazio zehatz bat egitea oso zaila da.

Laburbilduz, proiektuan eginiko lana esperientzia aberasgarria izan dela onartu nahi dut,

formakuntza pertsonala eta akademikoa egiten lagundu didalako eta etorkizuna argitzen lagundu didalako.

6.2 Etorkizuneko lanak

Mota honetako proiektuetan bukaera finkatzea oso zaila da, beti baitago hobekuntzak egiteko aukera eta gauza berriak probatzeko aukera. Horretaz gain, hobekuntzek bide berriak irekitzen dituzte.

Gure proiektuan sistema baten berrinplementazioa egitea lortu da eta honekin proba batzuk egitea ere bai, hala ere, etorkizunean proiektuak har ditzakeen jarraipen lerro edo bide berri posible hauek bururatu dira:

- Proiektuan erabili diren adieren errepresentazioek Ingelesa ez den beste hizkuntzetarako ere balio dute. Euskara hizkuntza horien artean dago eta posible izango litzateke sistema hau euskarazko datu-multzoekin probatzea. Euskaraz hitzen adierak desanbiguatzen dituen sistemarik ez da sortu, beraz, euskarazko hitz-adiera desanbiguazio ataza burutzen duen lehenengo sistema izango litzateke.
- Horretaz gain, lan honetan oinarrituz jarraitu daitekeen bidea sistemak sortzen dituen emaitzak hobetzeko teknika berrien inplementazioa da. Bide honetan saiakera ugari egin daitezke proba asko eginez. Lanean zehar otu den saiakera bat sistemaren bukaeran emaitzak kodetuko dituen sare neuronal bat probatzea da, K-NNa erabili beharrean sare neuronala erabiliz emaitzak hobetu ahalko bailirateke.

Eranskinak

A. ERANSKINA

Proiektuaren helburuen dokumentua

Eranskin honetan proiektuaren helburuen dokumentua garatuko da. Bere barne proiektua egiteko garaian egindako plangintza, irismena, faseak eta arriskuak izango ditu.

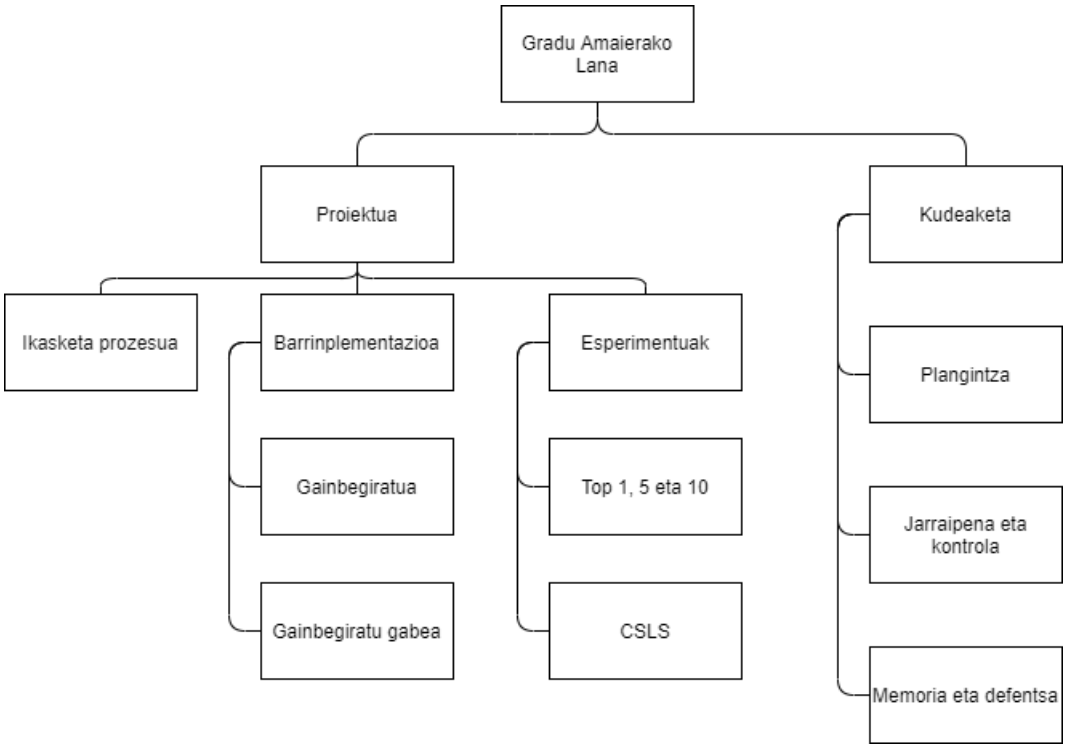
A.1 Irismena

Proiektuaren helburua hitzen adiera desanbiguazioan erabiltzen den sistema bat berrinplementatzea da; horretarako, [Scarlini et al., 2020]-ren lanean oinarritu da. Sistema berrinplementatzearen helburua sistemaren funtzionamendua eta oinarriak ulertzea eta egin daitezkeen hobekuntzak aztertzea da.

A.2 Proiektuaren plangintza

A.2.1 LDE diagrama

A.1 irudian ikus daiteke proiektuan zehar egin diren lanen deskonposaketa LDE diagramaren bitartez.



A.1 Irudia: LDE diagrama

A.2.2 Lan-paketeak

Gradu amaierako lana bi pakete nagusitan banatu da: proiektua eta kudeaketa izena eman zaie lan-pakete nagusiei.

Proiektua lan-paketea beste hiru paketetan banatu da, ikasketa prozesua, berrinplementazioa eta esperimentuak lan-paketeetan, hain zuzen ere. Ikasketa prozesuak ez dauka beste lan paketerik; berrinplementazioa, aldiz, gainbegiratua eta gainbegiratu gabea lan-paketeetan banatu da. Eta esperimentua paketea beste bi paketeetan banatu da Top 1,5 eta 10 eta CSLS paketeetan.

Kudeaketa hiru lan-paketeetan banatu da: plangintza, jarraipena eta kontrola eta memoria eta defentsa.

A.2.3 LDE diagrama

A.1 irudian ikus daiteke proiektuan zehar egin diren lanen deskonposaketa LDE diagramaren bitartez.

Lan-paketeen azalpena eta bakoitzari eskainiko zaion denbora (A.1 taula):

- Proiektua:

  IP (*Ikasketa prozesua*). Lan-pakete honetan proiektua aurrera atera ahal izateko egin diren ikasketa guztiak sartzen dira, inplementazioa hasi aurretik, tartean eta ondoren egin diren ikasketa guztiak.

  Berrinplementazioa:

  G (*Gainbegiratua*). Lan-pakete honetan inplementazioan sistema gainbegiratua garatzeko egin behar izan diren ataza guztiak sartzen dira.

  GG (*Gainbegiratu gabea*). Lan-pakete honetan inplementazioan sistema gainbegiratu gabea egiteko egin behar izan diren ataza guztiak sartzen dira.

  Esperimentuak:

  T (*Top 1,5 eta 10*). Lan-pakete honetan sistema ez gainbegiratuan eginiko proben emaitzak lortzeko egin diren ataza guztiak sartzen dira.

  CSLS (*CSLS*). Lan-pakete honetan CSLS neurriarekin eginiko proben emaitzak lortzeko eginiko ataza guztiak hartuko dira kontuan.

- Kudeaketa:

  P (*Plangintza*). Lan-pakete honetan proiektuaren plangintza egiteko eginiko ataza guztiak hartuko dira kontuan.

  JK (*Jarraipena eta kontrola*). Lan-pakete honetan proiektua jarraipena egiteko eginiko bilerak eta kontrolak hartuko dira kontuan.

  MD (*Memoria eta defentsa*). Lan-pakete honetan memoria idazteko eta defentsa preparatzeko eginiko ataza guztiak hartuko dira kontuan.

A.2.4 Emangarriak

Atal honetan proiektuan garatu beharreko emangarriak zein diren azalduko dira.

- **Memoria**
- **Defentsaren aurkezpena**

Emangarria	**Denbora(ordutan)**
Proiektua	350
Ikasketa prozesua	50
Berrinplementazioa	150
Gainbegiratua	100
Gainbegiratu gabea	50
Esperimentuak	150
Top 1,5 eta 10	75
CSLS	75
Kudeaketa	250
Plangintza	30
Jarraipena eta kontrola	20
Memoria eta defentsa	200
GUZTIRA	600

A.1 Taula: Lan-pakete bakoitzari eskainiko zaion denbora orduetan aurreikusten duen taula

A.2.5 Mugarriak

Atal honetan proiektuaren emangarrien entrega datak adieraziko dira A.2 taularen bidez.

Lan-paketeak	**Data**
Memoria	2020/06/21
Aurkezpena	2020/06/1-12

A.2 Taula: Lan-pakete bakoitzaren entrega datak adierazten dituen taula

A.2.6 Gantt diagrama

Atal honetan lan-paketeak eta mugarriak zein datatan egingo diren azalduko da *Gantt diagrama* baten laguntzaz. Ikus A.2 irudia.

Lan-paketea			Hasiera	Bukaera	2020						
					1	2	3	4	5	6	7
Proiektua	Ikasketa prozesua		2020/01/30	2020/06/05							
	Berrinpleme -ntazioa	Gainbegiratua	2020/02/15	2020/03/15							
		Gainbegiratu gabea	2020/03/15	2020/04/03							
	Esperiment uak	Top 1,5 eta 20	2020/04/5	2020/04/25							
		CSLS	2020/04/25	2020/05/05							
Kudeaketa	Plangintza		2020/01/27	2020/02/05							
	Jarraipena eta kontrola		2020/02/01	2020/07/10							
	Memoria eta defentsa		2020/05/05	2020/07/10							

A.2 Irudia: Gantt diagrama

A.3 Arriskuak eta prebentzioak

Atal honetan proiektua garatzeko garaian sor daitezkeen arazoak zein izan daitezkeen eta hauei eman daitezkeen balizko irtenbideak adieraziko dira.

A.3.1 Arriskuak

- **Esperimentua erreplikatzean arazoak izatea.** Esperimentua artikulu bateko informazioan oinarrituz erreplikatu beharra dago, baina baliteke artikuluan behar adina datu ez izatea erreplikatzeko edo emaitza desberdinak lortzea.

- **Hiztegi gabeko esperimentuetan denbora arazoak izatea.** Hiztegi gabeko sistema erabiliz emaitzak lortzeko denbora asko behar da konparaketa ugari egin behar direlako.

A.3.2 Prebentzioa

- **Esperimentua erreplikatzean arazoak izatea.** Esperimentua erreplikatzeko gai ez banaiz irtenbide batzuk bilatu beharko ditut. Horien artean, artikuluaren jabeari laguntza eskatzea dago, baina hau aurrera eraman ahal izatea artikuluaren jabearen esku dago.

- **Esperimentu gainbegiratu gabeetan denbora arazoak izatea.** Denbora murrizteko esperimentua paraleloki exekutatzea izan daiteke arazo honi eman daitekeen soluzioa.

A.4 Jarraipena eta kontrola

Jarraipena egiteko proiektuaren hasieratik bilerak egin dira tutoreekin. Bileren helburua aste bakoitzean egin beharreko lana egin den ala ez ziurtatzea da eta egiteko gai izan ez denean galderak egitea izan diren arazoen inguruan. Horrez gain, bileretan hurrengo asterako egin behar denaz ere hitz egin da.

Bilerez gain, tutoreekin harremana posta elektroniko bidez ere egin da eta drive-ko karpeta bat sortu da denen artean partekatua.

Amaitzeko, proiektuaren garapenaren tartean COVID19 pandemiak eragin zuzena izan du. Bilerak egiteko baliabideak aldatu behar izan dira presentzialak izatetik bideokonferentzi bidez egitera. Honetaz gain, fakultateko baliabideak ere ezin izan dira erabili kodea exekutatzeko eta beti egon da proiektua atzeratzeko arriskua gaixotasunaren eraginen ondorioz.

Bibliografia

[Camacho-Collados et al., 2016a] Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016a). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

[Camacho-Collados et al., 2016b] Camacho-Collados, J., Pilevar, M. T., and Navigli, R. (2016b). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240.

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

[Lample et al., 2018] Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *International Conference on Learning Representations*.

[Miller et al., 1993] Miller, G., Leacock, C., Tengi, R., and Bunker, R. (1993). A semantic concordance. pages 303–308.

[Moro and Navigli, 2015] Moro, A. and Navigli, R. (2015). SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

[Navigli et al., 2013] Navigli, R., Jurgens, D., and Vannella, D. (2013). SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

[Navigli and Ponzetto, 2012] Navigli, R. and Ponzetto, S. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Oram, 2001] Oram, P. (2001). Wordnet: An electronic lexical database. christiane fellbaum (ed.). cambridge, ma: Mit press, 1998. pp. 423. *Applied Psycholinguistics*, 22(1):131–134.

[Pilehvar et al., 2013] Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1341–1351, Sofia, Bulgaria. Association for Computational Linguistics.

[Pradhan et al., 2007] Pradhan, S., Loper, E., Dligach, D., and Palmer, M. (2007). SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

[Preiss and Yarowsky, 2001] Preiss, J. and Yarowsky, D., editors (2001). *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France. Association for Computational Linguistics.

[Raganato et al., 2017] Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

[Scarlini et al., 2020] Scarlini, B., Pasini, T., and Navigli, R. (2020). SensEmBERT:Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of AAAI*, New York, USA. AAAI.

[Snyder and Palmer, 2004] Snyder, B. and Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

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