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

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Adiera desanbiguazioa euskararako ikasketa sakona erabiliz

Egilea

Tasio Aguirre Blanco

Zuzendariak

Eneko Agirre eta Ander Barrena

Orwell warns that we will be overcome by an externally imposed oppression. But in Huxley's vision, no Big Brother is required to deprive people of their autonomy, maturity and history. As he saw it, people will come to love their oppression, to adore the technologies that undo their capacities to think.

- Neil Postman, *Amusing Ourselves to Death*

Laburpena

Adiera-desanbiguazioan eta hizkuntzaren prozesamenduko beste hainbat atazetan, ikasketa sakonean oinarritutako metodoek hobekuntza asko ekarri dituzte azken urteotan. Ikasketa sakoneko metodoak erabiltzean, ordea, arazo batekin topatzen gara: gizakiek eskuz etiketatutako datu kopuru handiak beharrezkoak dira ikasketa faserako, eta ingelesa bezalako hizkuntza batekin alderatuz, euskara bezalako hizkuntza txikietan ez da erraza izaten datu kopuru handiak biltzea.

Proiektu honen helburua, ingelesezko datuak erabiliz entrenatutako sistema bat eraldatzeko prozesu bat garatzea izan da. Hizkuntzen artean kontzeptuak mapatzeko informazioa erabiliz, euskarazko testuak desanbiguatzeko balio duen sistema eleanitz bat sortu da. Proiektuan sistema konkretu bat erabili da oinarri bezala eta honen gainean lan eginez lortu diren emaitzak aztertzen dira, bai garapen fasean, baita euskarazko datu-multzo baten gainean ere.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Gure hizkuntza anbiguoa da, hitz baten agerpen batek zentzu guztiz desberdina izan dezake testuinguruaren arabera. Gu hizkuntzaren ezaugarri honetara ohituta gaude, eta oso erraza egiten zaigu hitzen adiera desberdinak identifikatzea; gehienetan kontzeptu desberdinei erreferentzia egiteko hitz berdina erabiltzen dugula ohartu ere egin gabe. Honen adibide da ingelesezko (eta gaztelaniazko) *second* hitza, "bigarren" eta "segundo" denbora unitatea kontzeptuak adierazten dituela.

Hitzen adiera egokia zein den erabakitzea beharrezkoa da hainbat atazatan, adibidez, hizkuntzen artean itzulpenak egiteko, interneteko bilatzaile batetik emaitza desegokiak ezabatzeko edo galdera bat modu egokian erantzuteko. Ataza hau gizakiek egin dezaketen arren, konputagailu batek modu automatikoan egiteari hitzaren adiera-desanbiguazio (HAD) automatikoa deritzo.

Hitzaren adiera-desanbiguazioa hizkuntzaren prozesamenduko lehen atazetako bat izan zen, 40. hamarkadan sortua. Urte askotan zehar ez zen arrakasta handirik lortu, HAD sistemen zailtasuna agerian geratuz, eta ez zen izan 80. hamarkada arte aurrerapen handiak lortu zirela: hiztegi elektronikoak ugaritu ziren, metodo estatistikoak nabarmendu ziren eta lehen lehiaketak sortu ziren ebaluazio bateratuak egiteko. Azken urteotan ikasketa sakonean oinarritutako metodoetan aurrerapen asko garatu dira, HADn hobekuntza asko eskainiz.

Proiektu hau garatzeko ikasketa sakoneko metodoak erabiliz garatutako sistema batetik abiatzen da[1]. Sistema hau garatzeko ikasketa gainbegiratua erabiltzen da, hau da, eskuz

[1] *Language Modelling Makes Sense* (4 kapituluan azaltzen da)

etiketatutako ingelesezko corpus bat erabiltzen da adiera desberdinen errepresentazioak ikasteko. Jarraian, datu-base lexikal bat erabiliz, corpusean agertu ez diren adieren errepresentazioak lortzen dira kontzeptuen arteko erlazioak erabiliz.

Proiektu honetan sistema honetatik abiatuz, euskaraz erabili daitekeen sistema eleanitz bat garatzen da, ingelesez lortutako ezagutza baliatuz euskarazko HAD burutu ahal izateko. Jatorrizko sistema eta egindako aldaketak ulertu ahal izateko beharrezkoa den oinarri teorikoa azaltzen da, eta lana burutzeko hartu diren erabakiak azaltzen dira. Euskarako HAD ebaluazio datu-multzoa antolatu da, existitzen den Euskal SemCor corpusean oinarrituta. Azkenik, ingelesezko eta euskarazko ebaluazioa burutzen da, emaitzak aztertuz eta etorkizuneko hobekuntza posibleak proposatuz.

Memoria honetan proiektuaren nondik norakoak azaltzen dira, kapitulu desberdinetan antolatuta. Lehenik proiektuan adiera-desanbiguazioaren azalpen orokor bat egiten da 3 kapituluan, hala nola, erabili diren baliabide eta metodoen azalpena. 4 kapituluan oinarri bezala erabili den artikuluaren aurkezpena egiten da, bertan burutzen den prozesuaren azalpena eginez pausoz-pauso. 5 kapituluan sistema eleanitz bihurtzeko eginiko lana azaltzen da, eta sistema berria ebaluatu ahal izateko datu-multzoetan egin beharreko aldaketak azaltzen dira. 6 kapituluan proiektuaren garapenean zehar lortutako emaitzak aurkezten dira, eta euskarazko azken emaitzen azterketa burutzen da. Azkenik, 7 kapituluan proiektutik ateratako ondorioak eta etorkizunen egin daitezkeen hobekuntza posibleak jasotzen dira.

2. KAPITULUA

Proiektuaren Helburuen Dokumentua

Kapitulu honetan, eginiko proiektua deskribatu eta helburuak azaltzeaz aparte, lanaren deskonposaketa eredua (LDE) eta egingo diren atazak, aurreikusitako denboraren kudeaketa eta Gantt diagrama azalduko dira. Gainera, arriskuen analisia burutuko da eta hauek ekiditeko prebentzio-plana aurkeztuko da.

2.1 Proiektuaren deskribapena eta helburuak

Proiektu honen helburu nagusia euskarazko adiera-desanbiguazioko sistema bat garatzea izan da, [Loureiro and Jorge, 2019] artikuluan (**LMMS** hemendik aurrera) aurkezten den sistemaz baliatuz. Sistema hau ikasketa sakoneko tekniketan oinarritzen da, ingelesezko testuak erabiliz hitz desberdinen adieren errepresentazioak ikasteko. Gure lan nagusia, jatorrizko sistemaren emaitzak erreplikatzea, eta ingelesez ikasitako ezagutza hau euskara-ra eramatea izango da, euskarazko testuen gainean adiera-desanbiguazioko atazak burutu ahal izateko. Euskarazko adiera-desanbiguazioa burutzeko datu-multzorik eskuragarri ez dagoenez, eginiko lana ebaluatzeko datu-multzo bat garatu beharko da eskuragarri dauden baliabideez baliatuz.

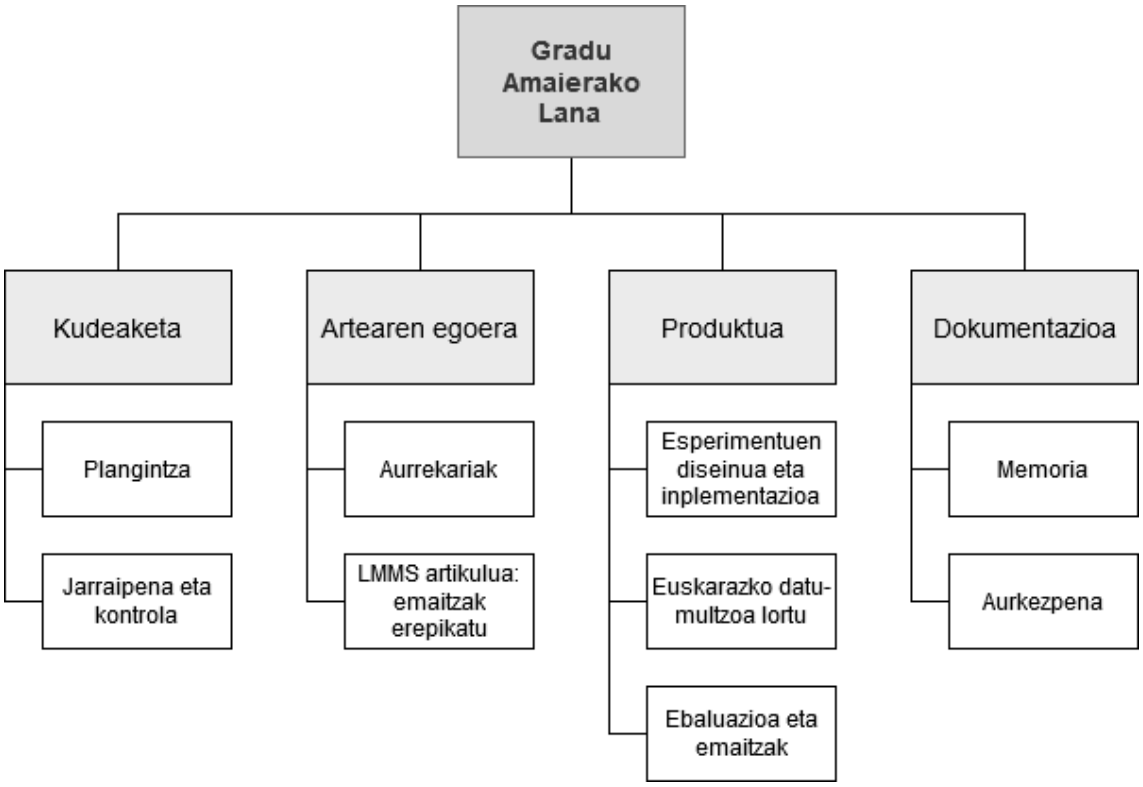
Lehenik, artearen egoera aztertuko da, erabiliko diren baliabide desberdinak azalduz. LMMS artikuluko sistema nola sortu den, eta sistema hori euskara emateko burutuko diren pausuak azalduko dira. Azkenik, lortutako emaitzak aztertu eta ondorioak erauziko dira.

2.2 Plangintza

Azpiatal honetan proiektua garatzeko eginiko plangintza azalduko da. Proiektua aurrera eramateko planteatutako faseak aurkeztuko dira, fase bakoitza burutzeko aurreikusitako denbora, eta egutegia.

2.2.1 LDE diagrama

Proiektuan zehar burutu beharreko lanaren deskonposaketa agertzen da 2.1 irudian, LDE diagrama erabiliz.



2.1 Irudia: Lanaren Deskonposaketa Eredua.

2.2.2 Lan-paketeak

LDE diagraman agertzen diren lan-paketeen deskribapena egiten da azpiatal honetan. Ataza bakoitzari egokitutako ordu kopurua 2.1 taulan agertzen da.

Ikerkuntzako proiektuetan ez da erraza izaten ordu kopurua zehazki aurreikustea, batez ere ezezaguna den gai baten egiten den lehen proiektua izanik; hori dela eta ordu kopurua zehaztasunik gabeko estimazioa besterik ez da. Aurreikusitako ordu kopurutik asko urruntzen garela nabarituz gero, proiektuaren helburuetan gehiago edo gutxiago sakontzeko aukera egongo da, ahal den neurrian guztizko 300 orduen inguruan mantenduz.

Kudeaketa

- **Plangintza:** Ataza honetan proiektuaren planifikazioa garatuko da; helburuak, atazak eta lan-metodologia definitu, eta proiektuaren bideragarritasuna eta arriskuen analisia burutuko da.
- **Jarraipena eta kontrola:** Ataza honetan proiektuaren helburuak betetzen direla bermatuko da, astero zuzendariekin egingo diren bilerei esker bizitza-ziklo osoan zehar.

Artearen egoera

- **Aurrekariak:** LMMS artikuluarekin lanean hasi aurretik artearen egoera eta erabiliko diren baliabideak aztertuko dira. Orokorrean, hizkuntza eredu neuronalak, hitz eta esaldien errepresentazioak eta eskuz etiketatutako corpusen formatua aztertuko da proba desberdinak eginez.
- **LMMS artikulua:** Ataza honetan LMMS artikulua sakonean aztertuko da, bertan azaltzen den prozesua ondo ulertzeko. Proiektuan zehar erabiliko den ingurunea martxan jarriko da, eta pausoz-pauso kodea exekutatuz, artikuluko emaitzak erreplikatuko dira hauen egiazkotasuna bermatzeko.

Produktua

- **Esperimentuen diseinua eta inplementazioa:** Ataza honetan euskaraz erabil daitekeen sistema eleanitza sortuko da LMMS-n aurkezten den lanetik abiatuz. Proze-

su osoan zehar esperimentuak egingo dira bide onetik goazela ikusteko, eta emaitza onenak eskainiko dituzten erabakiak hartzeko.

- **Euskarazko datu-multzoa lortu:** Ataza honetan euskarazko sistema ebaluatu ahal izateko beharrezkoa den datu-multzoa lortuko da. Horretarako jada existitzen den datu-multzo bat lortu, edo berri bat sortu beharko da. Azkenik, datu-multzoa eraldatu beharko da, ingelesezko multzoen formatu berdinera bihurtzeko.
- **Ebaluazioa eta emaitzak:** Ataza honetan esperimentuetan lortutako emaitzak aztertuko dira, sistemaren eraginkortasuna ebaluatzeko. Ondorioak aterako dira, etorkizunean egin daitezkeen hobekuntzak proposatuz.

Dokumentazioa

- **Memoria:** Proiektuaren azalpen guztiak eta lortutako emaitzen analisia biltzen dituen dokumentua garatuko da.
- **Aurkezpena:** Proiektuaren defentsarako aurkezpen bat prestatu beharko da, eginiko memoria oinarri bezala hartuz, atal garrantzitsuenak azalduko dira gardenkien bidez.

Lan-paketea	**Iraupena (orduak)**
Kudeaketa	**30**
Plangintza	10
Jarraipena eta kontrola	20
Artearen egoera	**30**
Aurrekariak	15
LMMS artikulua	15
Produktua	**140**
Esperimentuen diseinua eta inplementazioa	60
Euskarazko datu-multzoa lortu	60
Ebaluazioa eta emaitzak	20
Dokumentazioa	**100**
Memoria	80
Aurkezpena	20
Guztira	**300**

2.1 Taula: Proiektuko lan-pakete bakoitza garatzeko beharko den denboraren estimazioa.

2.2.3 Emangarriak eta mugarriak

2.2 taulan proiektuan zehar garatu beharreko emangarrien emate datak agertzen dira.

Emangarria	Data
Memoria	2020-06-21
Aurkezpena	2020-06-29
	2020-07-10

2.2 Taula: Proiektuko emangarrien datak.

2.2.4 Gantt diagrama

2.2 irudian ikus daitezke LDE diagramako ataza bakoitzak izango duen hasiera eta bukaera datak *Gantt* diagramaren bidez. Proiektua orokorrean nahiko lineala izatea espero da, mota honetako proiektuetan ez delako ohikoa izaten ataza desberdinak paraleloan burutzen joatea.

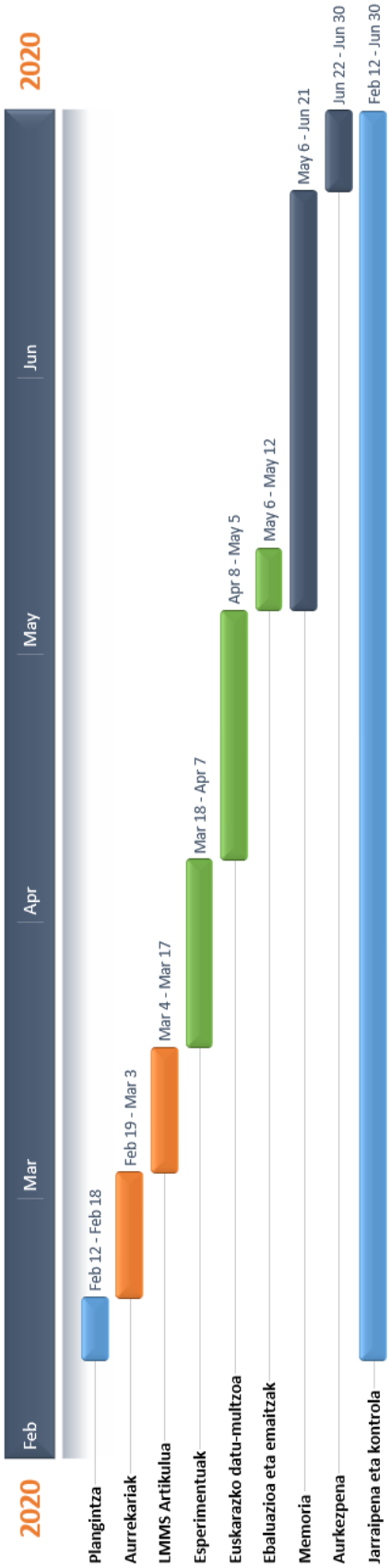
2.3 Lan Metodologia

Proiektua garatzeko etxetik lan egingo da ordenagailu pertsonala erabiliz. Ez da ordutegi finkorik ezarriko; beharrezkoak diren orduak modu egokienean banatuko dira, aldi berean burutzen ari diren irakasgaien lan karga kontuan izanik, hauek izanik lehentasuna orokorrean.

2.3.1 Bilerak

Ikasle eta zuzendarien arteko komunikazioa astero ezarritako bileren bidez burutuko da. Bilera hauetan jarraipena eta kontrola burutuko da, hala nola garapenean zehar sortutako zalantzak argitu. Bilerak normalean fakultatean bertan burutuko dira, bideokonferentzia bidez egiteko aukera izanik ezinbesteko arrazoi bat dela eta.

Posta elektronikoa ere erabiliko da komunikazio kanal bezala, bileren ordua zehazteko, edo noizean behingo zalantzak argitzeko.



2.2 Irudia: Proiektuko kronograma. Gantt diagrama.

2.4 Arriskuak eta prebentzioa

Proiektuan zehar egon daitezkeen arriskuak identifikatzen dira atal honetan. Arrisku bakoitzarentzat prebentzio-plana sortuko da.

Informazio galera

- **Deskribapena:** Esperimentuak ordenagailu pertsonalean egingo direnez, posible da arazo teknikoengatik datuen galera sufritzea.
- **Prebentzioa:** Egiten den lana bi disko gogorretan mantenduko da, datuen tamaina handiak ezinezkoa egiten baitu lainoan babes kopiak izatea. Hala ere, garatzen diren programen kopia mantenduko da USB memoria batean.

Konputazio behar handiak

- **Deskribapena:** Posible da inplementazioa burutzeko beharrezkoa den konputazio ahalmena handiegia izatea modu lokalean exekutatu ahal izateko.
- **Prebentzioa:** Honelako kasu bat antzematen bada, Google Colab edo IXA taldearen zerbitzariak erabiltzeko aukera izango dugu. Lana urruneko zerbitzaritan exekutatuko da, etxetik atzituz.

Zailtasun maila handiegia

- **Deskribapena:** Mota honetako proiektuetan ez da erraza izaten proiektuak izango duen zailtasun maila aurreikustea. Gerta liteke egin beharreko atazaren bat gradu amaierako lan batek eskatzen duen ezagutza mailatik haratago geratzea, edota ezarritako helburuak urruti gelditzea.
- **Prebentzioa:** Jarraipen eta kontroleko bileretan zailtasunaren inguruko analisia burutuko da proiektuaren lehen asteetan. Zuzendarien eta IXA taldeko kideen laguntza jasotzeko aukera egongo da ataza konkretuak burutzeko, beti mantenduz helburuak egokitu, edo irismena aldatzeko aukera.

3. KAPITULUA

Artearen egoera

Kapitulu honen lehen atalean hitzen adiera-desanbiguazioari buruzko azalpen sakonago bat eskaintzen da; ataza hau ebazteko erabiltzen diren baliabideak aurkeztuz. Bigarren atalean, azken urteotan hizkuntzaren prozesamenduan arrakasta handia izan duen ikasketa sakona [Otter et al., 2018] zer den azaltzen da.

3.1 Hitzen adiera-desanbiguazioa

Sarreran aipatu den bezala, hitzak anbiguoak dira: hitz berak interpretazio desberdinak izan ditzake testuinguru desberdinetan. Konputazio-metodoak erabiliz hitz baten agerpen bati adiera egokia emateari hitzaren adiera-desanbiguazioa (HAD) deritzo. Atal honen helburua, adiera-desanbiguazioari buruzko informazio orokorra eta lana gauzatzeko erabili diren baliabideak aurkeztea da. Eginiko lana zehazki azaltzen hasi aurretik beharrezkoa den informazio eta kontzeptu garrantzitsuenak aurkeztuko dira.

3.1.1 Zer da HAD?

HAD zer den ulertu ahal izateko modu errazena adibide baten bidez izan ohi da. Horregatik, eta oso ondo azaltzen duelako, [Lopez de Lacalle and Agirre, 2010] artikuluan azaltzen den adibidea hona ekarriko dugu.

Hitz batek agertzen den testuinguruaren arabera hainbat interpretazio izan ditzakeela ikusteko, *banku* hitzaren ondoko bi adibideak aztertuko dira:

i: Parkeko *bankuan* eseri nintzen egunkaria irakurtzera.

ii: Dirua *bankuan* gorde dut.

Hiztegi baten *banku* hitza bilatuz honako definizioak aurkitu daitezke[1]:

banku 1: Eserleku luzea, bizkarduna nahiz bizkarrik gabea, hainbat lagun batera esertzeko aukera ematen duena.

banku 2: Bezeroen diru-gordailuak onartu eta kreditu-eragiketak egiten dituen enpresa publiko edo pribatua.

Nahiz eta hizkuntza ezagutzen duen pertsona batek naturalki identifikatu ahal duen testuinguru bakoitzean hitz bakoitzak hartzen duen adiera, makinentzat ez da lan erraza izaten. HAD sistema automatiko batek, lehenengo adibidea esertzeko altzariari buruzkoa dela (*i: banku 1*), eta bigarrena banketxeari buruzkoa dela (*ii: banku 2*) aukeratu beharko luke. Era berean, testuinguruko gainontzeko hitzak desanbiguatuko lituzke HAD sistemak. Adiera hauek datu-base lexikaletan biltzen dira, adibidez, WordNet-en.

3.1.2 WordNet

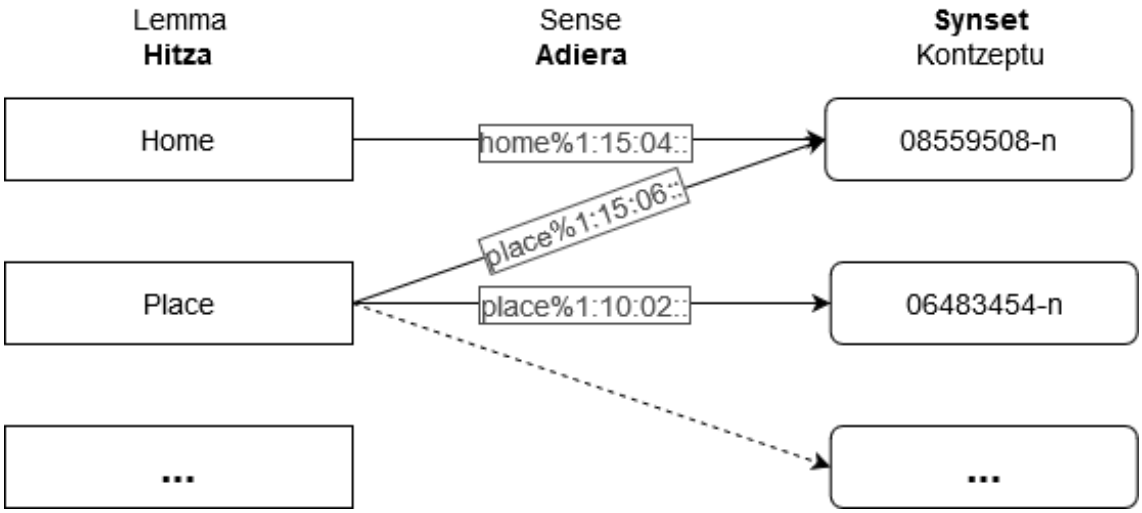
WordNet [Fellbaum, 1998] ingelesezko eskuz etiketatutako datu-base lexikal handi bat da, izen, aditz, adjektibo eta adberbioen arabera antolatuta datu-multzo desberdinetan. Datu-multzo bakoitzak **hitz** kopuru bat du, hauetako bakoitzak **synset** bat edo gehiago izanik lotuta. Synsetak sinonimo multzoak dira (*synonym set*), kontzeptu lexikal edo adiera bati erreferentzia egiten diotenak.

Normalean hainbat hitz egon daitezke synset berdinari lotuta, kasu honetan, hitz hauek elkarren artean sinonimoak direla esaten da eta ale lexikal hauei *variant* deitzen zaie. Adibidez, ingelesezko *home* eta *place* hitzak synset baten *variant*ak izango dira kontzeptu bera adierazten dutenean. Hau erraz ikus daiteke esaldi berdinean trukatzen badira bi hitz hauek: *"Deliver the package to my place"* eta *"Deliver the package to my home"* esaldiek zentzu bera dutela ikus daiteke.

[1] EusWordNeten agertzen diren definizioak (EusWordNeten azalpena 3.1.3 atalean).

Hitzen eta synset-en arteko erlazio hau hobeto uler daiteke 3.1 irudian. Synset bakoitzak adierazten duen kontzeptua, normalean, **glosa** (edo definizio) baten bidez adierazten da; horrela, hau izango litzateke aurreko adibidearen glosa: *'where you live at a particular time'*. Glosaz aparte, kategoria sintaktiko eta taldekatze logikoa ere definitzen da, *lexname*[2] erabiliz. *Home* hitzaren *lexname*-a *'noun.location'* izango litzateke adibidean.

Hala ere, hitzen eta synseten erlazioei buruz hitz egitean, beste elementu garrantzitsu bat erabiltzen da WordNeten: **adierak** (*sense* ingelesez). 3.1 irudian ikusten den bezala, adierak, erlazio konkretuak identifikatzeko erabiltzen dira *sense_key* izeneko kode baten bidez etiketatuz[3]. Oso baliagarriak izaten dira hitz baten adiera konkretu bati buruz informazioa lortu nahi denean, gainontzeko adierak edo hitzak kontuan izan gabe.



3.1 Irudia: WordNeten hitzen eta synseten arteko erlazioaren ilustrazio bat. Hitz batek hainbat synset izan ditzake, eta alderantziz. Erlazio hauetako bakoitzari adiera deritzo.

Glosak synsetei lotuta daudenez, synsetak dira WordNeteko sarrerekin erlazionatzen diren oinarrizko unitateak, eta horregatik dira synsetak WordNeteko gainontzeko erlazio gehienetan parte hartzen duten elementuak; ez hitzak edo adierak. Beraz, WordNeten erlazio semantiko garrantzitsuenetako bat sinonimia da, baina hau ez da bakarra.

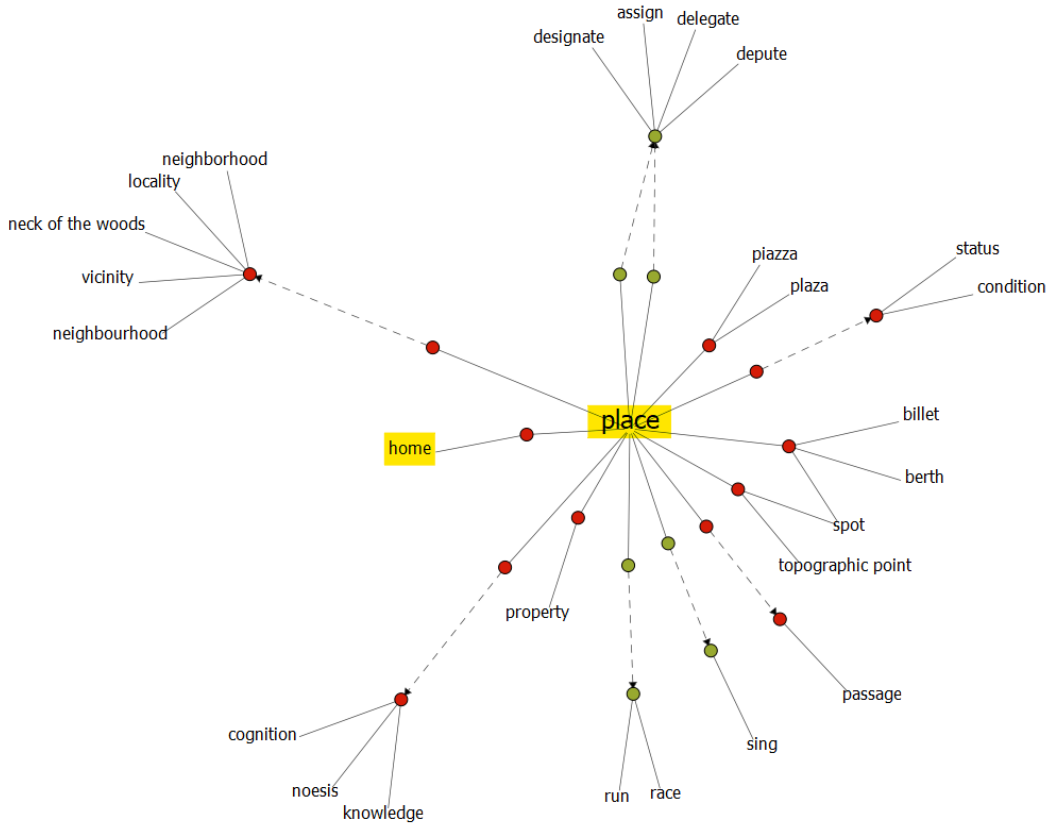
Synsetak elkarren artean erlazionatuz sortzen da WordNeten oinarrizko egitura, grafo bat izango balitz bezala. Honen adibidea ikus daiteke 3.2 irudian, non synseten arteko loturak erabiltzen dira hitzak elkarren artean lotzeko[4]. Hitzak eta aditzak antolatzeko sortzen den hierarkia egiteko beste erlazio semantiko batzuk erabiltzen dira, garrantzitsuenetako bat

[2] https://wordnet.princeton.edu/documentation/lexnames5wn

[3] Kode honek kodetzen duen informazioa hemen ikus daiteke: https://wordnet.princeton.edu/documentation/senseidx5wn

[4] Grafoa sortzeko tresna hau erabili da: https://code.google.com/archive/p/synonym/

hiperonimia-hiponimia erlazioa izanik. Erlazio hau erabiliz, synset orokorrenak synset zehatzagoekin lotzen dira, kate luzeagoak sortzeko aukera izanik, kontzeptu orokorretatik zehatzagoetara joanez. (Honen adibidea ikus daiteke hurrengo kapituluko 4.2 irudian)



3.2 Irudia: WordNeten *place* hitzaren sinonimia bidezko loturak, lerro etenaz hiperonimia-hiponimia erlazioa. *place* eta *home* grafoan lotuta agertzen dira synset berdinaren bidez erlazionatuta daudelako (ikusi 3.1 irudia).

Lan honetan WordNet 3.0 bertsioa erabili da. Bertsio honek 117,798 hitz, 11,529 aditz, 21,479 adjektibo eta 4,481 adberbio ditu[5]. Batez bestean hitz bakoitzak 1.23 adiera eta aditz bakoitzak 2.16 adiera ditu; guztira, 117,659 synset eta 206,941 adiera ezberdin izanik. Nahiz eta WordNeten 3.1 online bertsio berriagoa existitu, gaur egun 3.0 bertsioa zabalki erabilia da oraindik; proiektu hau garatzeko erabili diren baliabide asko bezala.

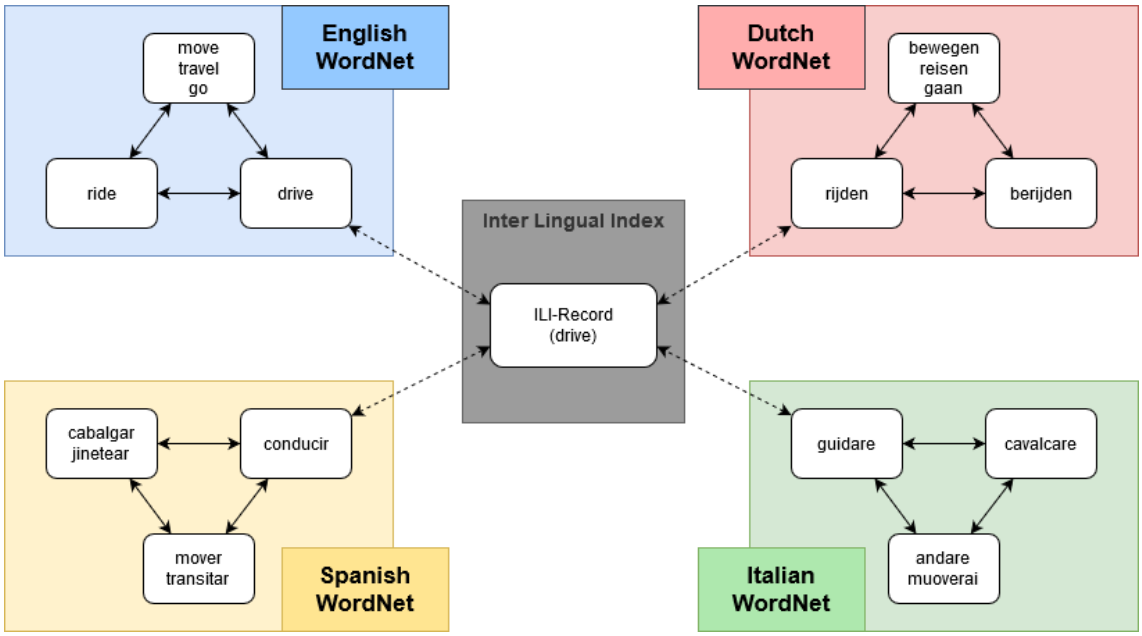
3.1.3 WordNet Eleanitzak eta EusWordNet

Ingelesaz aparte, hizkuntza desberdin askotarako ere garatu dituzte WordNetak. Global WordNet Asoziazioak (GWA) WordNet desberdinen estandarizazioa sustatzen du, hau

[5]Datu eguneratuak hemen: https://wordnet.princeton.edu/documentation/wnstats7wn

da, ahal den neurrian kontzeptu bera adierazten duten hizkuntza desberdinetako synsetek kode bera izan dezaten, ala ez balitz, synset kode desberdinen artean mapaketa eskuragarri egon dadin. Eskuragarri dauden hizkuntza desberdinetako WordNet bertsioak GWA-ren web orrian aurkitu daitezke[6]. Eleaniztasun honek, hizkuntzaren prozesamenduko hainbat ataza desberdinetarako oso baliagarriak egiten ditu WordNeteko bertsio desberdinak; ez bakarrik hitzen adiera-desanbiguaziorako.

EuroWordNet proiektuak [Vossen, 1998] Europako hainbat hizkuntzatako WordNetak garatu ditu, elkarrekin lotuz Inter-Lingual-Index (ILI) deritzona erabiliz. ILI-ren inguruan ez gara azalpen askotan sartuko, baina ideia orokor bat izateko 3.3 irudian azaltzen da modu orokor batean nola erlazionatzen diren hizkuntza desberdinetako synsetak ILI erabiliz.



3.3 Irudia: EuroWordNet datu basearen arkitektura, ILI erabiliz kontzeptuak lotuz. Irudi hau [Vossen, 1998] liburuan ILI azaltzeko erabiltzen den ilustrazioaren adaptazio bat da.

Multilingual Central Repository (MCR)[7], WordNeten osagarri den datu-base bat da; ingelesa, espainiera, katalana, euskara, galiziera, eta portugesa biltzen ditu EuroWordNeten egitura berdina erabiliz. EusWordNet (edo EusWN), beraz, MCR 3.0-ren barruan banatzen den euskarazko WordNet bertsio bat da, non euskarazko kontzeptuak ingelesezko WordNeteko synset kode berdinak erabiliz etiketatu diren. Lan honetarako oso baliagarria da, sistema eleanitz bat sortzea asko errazten baitu eta ingelesezko synseten informazioa

[6] http://globalwordnet.org/resources/wordnets-in-the-world/
[7] https://adimen.si.ehu.eus/web/MCR

euskarara eramateko aukera eskaintzen baitu. Gure kasuan, EusWN erabili da euskarazko hitz batek izan ditzakeen synset posibleak identifikatzeko.

3.1.4 Anotutatuko Corpusak: SemCor eta EuSemcor

Anotatutako Corpusak bi mota ezberdinetan bereiz daitezke: *All-words* (hitz guztiak) motako corpusetan hainbat testu aukeratzen dira, eta testu horietako izen, aditz, adberbio eta adjektibo guztiak etiketatzen dira. *Lexical sample* (hitzen lagina) motako corpusetan hitz jakin batzuk aukeratzen dira, eta hainbat esaldi ala testuinguru ematen dira hitz bakoitzeko.

SemCor semantikoki etiketatutako corpus bat da [Agency., 1993]. WordNet sortu zuen talde berdinak landu du corpus hau eta 200,000 agerpenez baino gehiago dago osatua; izen, aditz, adjektibo eta adberbioak kategoria gramatikalekin automatikoki etiketatuta, eta WordNeteko adierekin (*sense_key* kodeen bidez) eskuz etiketatuta egonik. Corpus hau *all-words* motakoa da. Nahiz eta hasiera baten WordNet 1.6 bertsioa erabiliz etiketatu bazen, lan honetan erabiltzen den bertsioa 3.0 bertsiora egokitua izan da.

EuSemcor euskarazko SemCor corpusa da [Pociello et al., 2011], hau da, adierak eskuz etiketatuta dituen euskarazko corpusa. Euskarazko 407 izen ohikoenek osatzen dute, izen hauen 42.615 agerpen[8] eskuz etiketatuta daudelarik EusWordNeteko synsetekin. SemCor ez bezala, corpus hau *lexical sample* motakoa da.

EuSemcor-en azkenengo bertsioa 2006 urtekoa denez, EusWN 1.6 bertsioa erabili zen hitzak etiketatzeko unean; hori dela eta, egindako lanaren parte bat izan da synset hauek 3.0 bertsiora mapatzea (ikusi 5.3 atala).

3.1.5 Datu-multzoak

Lan honetan [Raganato et al., 2017] artikuluan azaltzen diren datu-multzoak erabiltzen dira. Guztira bost *all-words* datu-multzo aurkezten dira, Senseval eta SemEval lehiaketetako ataza desberdinetatik jaso, eta formatu berdinera bihurtuak. Datu-multzo hauek WordNet 3.0 bertsioa erabiliz etiketatuta eskaintzen dira.

[8] Datu eguneratuak hemen: `http://ixa2.si.ehu.es/mcr/estatistikak.html`

3.2 Ikasketa sakona hizkuntzaren prozesamenduan

Neurona-sareen hasiera 40. hamarkadan eman zen, konputagailuek ataza adimentsuak gauzatzeko irrikak bultzatuta. Hasieran ez zuten arrakasta handirik izan; konputagailuak ez ziren behar bezain ahaltsuak eta ez zen existitzen oraindik datu-multzo handirik. Hurrengo hamarkadetan bi arlo hauetan eginiko hobekuntzek neurona-sareei ikasketa sakonaren bidez etekin hobeagoa ateratzea ahalbidetu zuten.

Adimen artifiziala, ikasketa automatikoa eta ikasketa sakona gaur egun asko entzuten diren kontzeptuak dira, baina ez da beti argi egoten kontzeptu hauek zertan bereizten diren. Adimen artifiziala izaki bizidunen prozesu arrazional eta deduktiboak makina baten bidez burutzeko gaitasuna da; hau lortzeko bide posible bat ikasketa automatikoa izanik. Ikasketa automatikoa, beraz, adimen artifizialaren adar bat da, esperientziatik ikasteko gai diren konputagailu programak garatzea helburu duena, ataza konkretu bat ebazteko programatuta egon gabe. Ikasketa sakona ikasketa automatikoko teknika multzo bat da, nerbio-sistema biologikoetan informazioaren prozesamendua nola gertatzen den imitatzen duen eredu konputazionaletan oinarrituta, hau da, neurona-sare artifizialetan.

Ikasketa sakonak hasiera batean gizakion burmuina simulatzea helburu zuen arren, gaur egun printzipio orokorrago bati egiten dio erreferentzia, non estatistika eta matematika aplikatuaren ezagutza erabiltzen den oinarri gisa. Alor askotan lortu du ikasketa sakonak bestelako metodoak gainditzea, hala nola, konputagailu bidezko ikusmenean, bioinformatikan, patroien errekonozimenduan, hizkuntzaren prozesamenduan...

Hizkuntzaren prozesamenduan, hizkuntza-ereduek probabilitate bat esleitzen diote hitzen sekuentzia bati, antzeko hitz eta esaldien artean bereizteko. Normalean probabilitate hau erabiltzen da aurreko hitzak jakinik, hurrengoa zein izango den iragartzeko; baina ez du beti horrela izan beharrik. Hizkuntza-ereduak oso erabilgarriak dira hizkuntzaren prozesamenduko hainbat atazatan, batez ere testua sortzean oinarritzen direnetan.

Neurona-sareetan oinarritutako hizkuntza-ereduak (Hizkuntza-Eredu Neuronalak) testu asko erabiliz entrenatzean, ezagutzen diren hitz desberdinen kopurua, hiztegia, handituz joaten da. Hiztegia handitzean, sor daitezkeen esaldi desberdin kopurua esponentzialki handitzen da. Honek arazoak ekar ditzake dimentsio handiko espaziotan[9]. Neurona-sareetan arazo hau saihesteko hitzen errepresentazioak erabiltzen dira.

[9] https://en.wikipedia.org/wiki/Curse_of_dimensionality

3.2.1 Hizkuntza-ereduen errepresentazioak: Embeddingak

Embeddingak hitzei (edo esaldiei) errepresentazio abstraktu bat esleitzen dioten zenbaki errealez osaturiko bektoreak dira[10]. Bektore hauek hitz bakoitza hainbat dimentsiotako bektore-espazio baten kodetzen dute, mapa bateko koordenatuak izango balira bezala, hitz bakoitzak posizio konkretu bat hartzen du. Honek, hitzen arteko antzekotasunak eta erlazioak modu erraz batean neurtzea baimentzen du eragiketa matematikoen bidez.

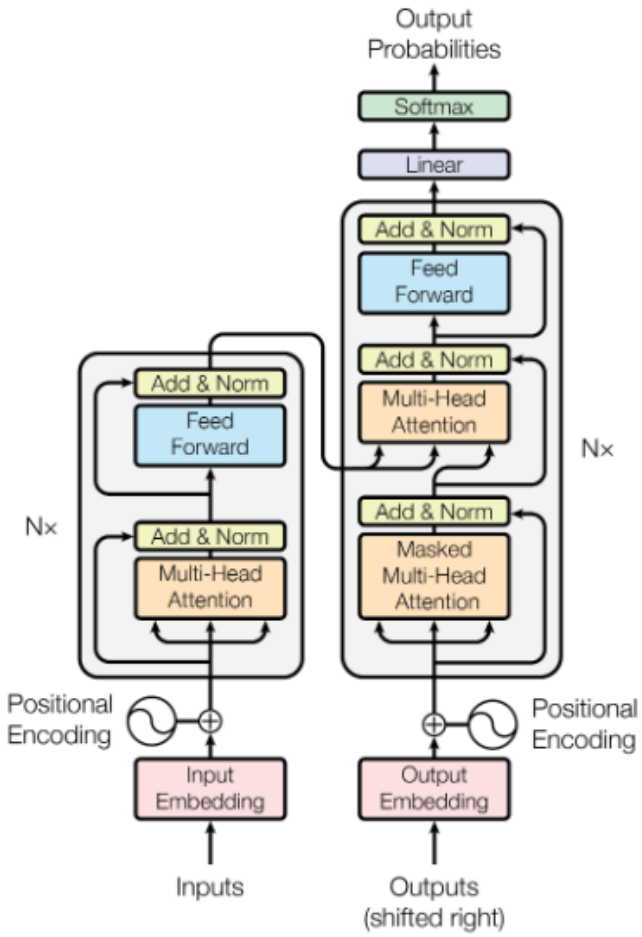
Embeddingak hizkuntzaren prozesamenduko hainbat ataza ebazteko erabili daitezke. Hori dela eta hainbat embedding eskuragarri daude, hala nola fastText [Caliskan et al., 2017], non hitz baten agerpen guztiak kontuan hartzen dira (normalean Wikipedia eta Common Crawl[11] erabiliz) embeddingak sortzean, hitzek izan ditzaketen adiera desberdinetan erreparatu gabe. Hauei embedding estatikoak (*Static Word Embeddings*) deritze.

Embedding estatikoak hainbat ataza burutzeko oso erabilgarriak diren arren, ez dute balio hitz baten adiera desberdinak bereizteko, hori dela eta HADn gehienbat testuinguruaren menpeko embeddingak (*Contextual Embedding*) erabiltzen dira. Mota honetan, hitz baten adiera desberdinak kontuan hartzen dira embeddingak sortzean, testuinguru desberdinek hitz bakoitzarentzat errepresentazio desberdinak sortuz. Testuingurua aipatzen denean, desanbiguatu nahi den hitza erdian duen luzera finkoko testu bati buruz hitz egiten da. Eskuz etiketatutako corpus batetik embeddingak sortu ahalko balira ere, ez da erraza existitzen diren hitz guztien adiera guztientzat embedding onak sortzea, adiera bakoitzarentzat adibide kopuru handia beharrezkoa baita, eta ohikoak ez diren agerpen asko ez direlako behar beste agertzen.

3.2.2 Transformerrak

Transformerra [Vaswani et al., 2017] lanean aurkeztutako eredua da. Datu sekuentzia ordenatuak maneiatzeko diseinatuta daude, adibidez, lengoaia naturalean idatzitako esaldi bat. Neurona-sare errepikakorrak (ingelesez *RNN - Recurrent Neural Network*) ez bezala, transformerrak erabiliz ez da beharrezkoa sekuentziak ordenan prozesatzea, hau da, prozesatu beharreko sekuentzia esaldi bat balitz, esaldiaren amaiera prozesatzeko ez litzateke beharrezkoa izango hasiera prozesatu izana. Honi esker transformerrak oso erraz paralelizatu daitezke, entrenamendu kostua nabarmenki gutxituz.

[10] https://eu.wikipedia.org/wiki/Word_embedding
[11] https://commoncrawl.org/



3.4 Irudia: Transformerraren arkitektura. [Vaswani et al., 2017].

Neurona-sare errepikakorrak erabiliz sekuentzia luzeak tratatu behar direnean, askotan irteerak ez du mantentzen sententziaren hasierako elementuen informazioa. Hau arazo handia izan daiteke, adibidez, hizkuntza desberdinen artean itzulketa egitean, jatorrizko esaldiaren lehen hitzaren garrantzia handia izaten baita itzulketa egiteko. Arazo hau konpontzeko *atentzio* izeneko mekanismo bat gehitzen da, sarrerako sekuentziako elementu garrantzitsuenak identifikatzeko aukera eskainiz.

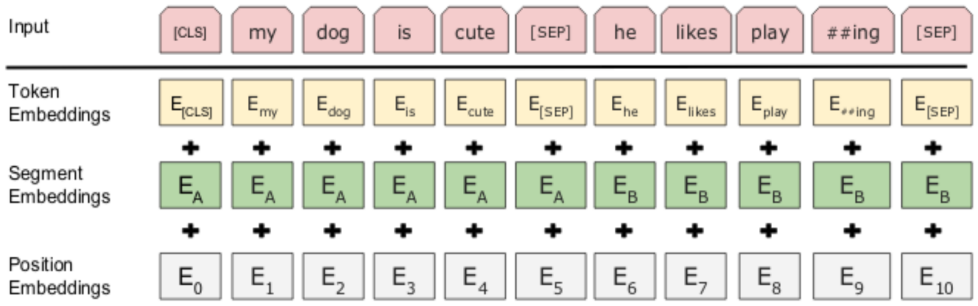
Transformerrak atentzio-mekanismoetan bakarrik oinarritzen dira, sei geruza dituen kodetzaile-deskodetzaile arkitektura erabiliz, 3.4 irudian ikus daitekeen bezala. Hasiera batean itzulpen automatikorako proposatua izan bazen ere, ataza desberdinetara hedatu da azken urteetan.

3.2.3 Hizkuntza-eredu neuronalak: BERT

BERT [Devlin et al., 2018] transformerren kodetzailearen arkitekturan oinarritutako hizkuntza-eredua da. Bi norabideko atentzio mekanismoak erabiliz, hau da, ezker eta eskuineko informazioa kontuan izanik, hizkuntzaren errepresentazioak aurre-entrenatzeko diseinatuta dago. Behin errepresentazioak lortuta, hizkuntzaren prozesamenduko ataza desberdinetara egokitu daiteke, egungo sistemak gaindituz arkitekturan aldaketa handirik egin gabe.

Sarrerako datuak

Eredu honen sarrera **token**, **segmentu** eta **posizio** embedingen baturaz osatzen da, esaldi bat edo bi adierazteko.



3.5 Irudia: BERT-en sarrerako embeddingen errepresentazioa, token, segmentu eta posizio embeddingak batuz. [Devlin et al., 2018]

Token embeddingak (3.5 irudiko *Token Embeddings*) WordPiece [Wu et al., 2016] embeddingak erabiliz adierazten dira. WordPiece embeddingak hitzak zati txikiagoetan banatuz lortzen dira, muga berezi batzuk gehituz jatorrizko esaldia berreskuratu ahal izateko. Adibidez, euskarazko *'etxera'* hitza *'etxe'* eta *'##ra'* tokenetan bihurtzen da banatu ondoren. Horrela, *'etxera'* hitzaren errepresentazioa bi tokenen batezbestekoa eginez lortu daiteke.

Bi token berezi ere erabiltzen dira: [CLS] sekuentziaren hasiera markatzeko, eta [SEP] sarrerako bi esaldiak banatzeko.

Segmentu embeddingetan, A sententzia embeddinga gehitzen da lehenengo esaldiaren token bakoitzeko eta B embeddinga bigarren sententziako token bakoitzeko. Sententzia bakarreko kasuetan A bakarrik erabiltzen da. (3.5 irudiko *Segment Embeddings*)

Posizio embeddingetan token bakoitzaren posizioa adierazten dute, 512ko luzera maximoa izanik. (3.5 irudiko *Position Embeddings*)

Aurre entrenamendua

BERT bi norabideko atentzio mekanismoetan oinarritzen denez, iraganeko zein etorkizuneko informazioa dauka. Esaldi baten hurrengo tokena zein den iragarri nahiko balu, jada token horren informazioa jasota izango luke. Hau dela eta, autoreek gainbegiratu gabeko bi iragarpen ataza proposatzen dituzte entrenamendua gauzatzeko: Maskaradun Hizkuntza-Eredua (*Masked Language Model*) eta Hurrengo Esaldiaren Iragarpena (*Next Sentence Prediction*).

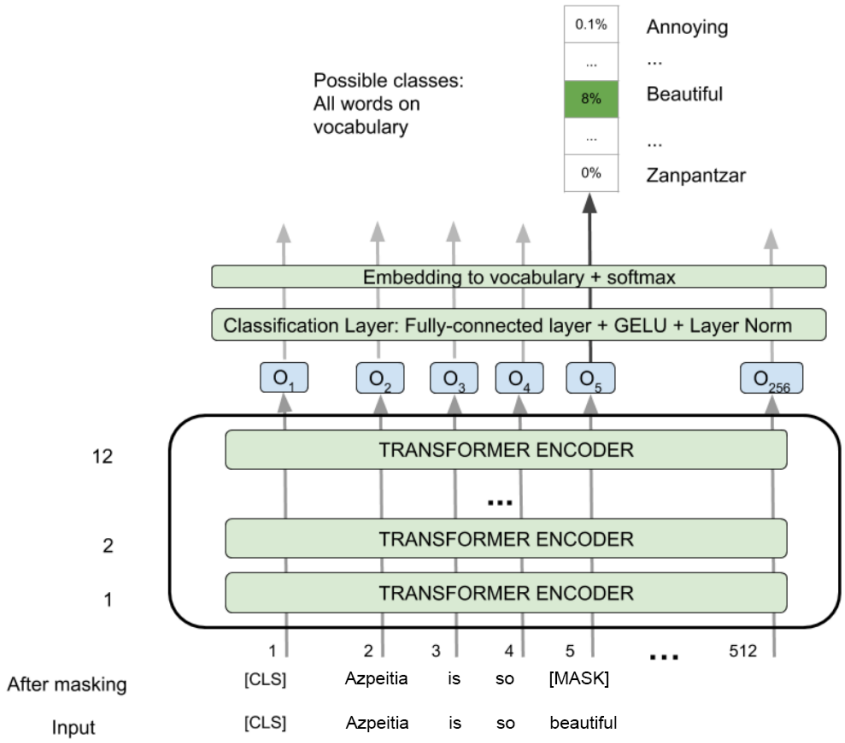
Maskaradun Hizkuntza-Eredua atazan, sarrerako tokenen %15 ezkutatzen dira, eta sistemak jatorrizko tokena zein zen iragarri behar du. Tokenak ezkutatzeko [MASK] erabiltzen da, baina token hau ataza honetan bakarrik erabiltzen denez desdoikuntza sortuko luke ezkutatzen den tokena beti [MASK] erabiliz ordezteak. Arazo hau konpontzeko autoreek konponbide hau iradokitzen dute:

- Kasuen %80-an [MASK] tokena erabiltzen da. Adibidez 'Azpeitia is so beautiful' 'Azpeitia is so [MASK]' bihurtzen da.
- Kasuen %10-ean ausazko hitz batekin ordezten da: 'Azpeitia is so potato'.
- Kasuen %10-ean ez da hitza aldatzen.

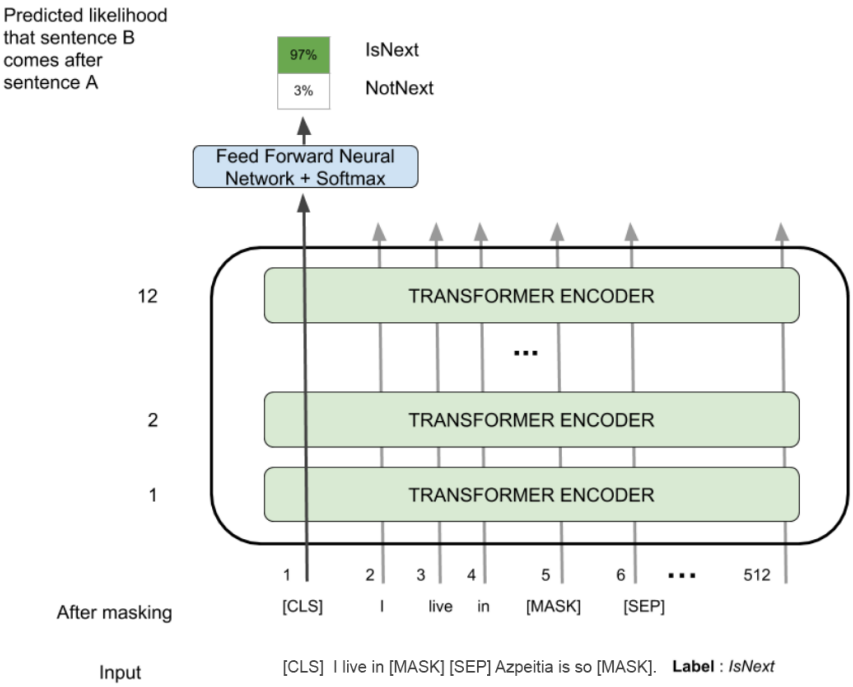
Konponbide hau erabiliz, sistemak ez daki zein hitz iragarri beharko duen, eta sarrerako token guztien testuinguruaren menpeko errepresentazioak mantentzera behartuta geratzen da. Ataza honen irudikapena ikus daiteke 3.6 irudian.

Hurrengo Esaldiaren Iragarpena atazan esaldien arteko erlazioa erakutsi nahi da. Hau lortzeko bi esaldi aukeratzen dira corpusetik, kasuen %50-ean jarraian doazen esaldiak aukeratzen dira, gainontzeko kasuetan elkarrekin erlaziorik ez duten ausazko bi esaldi aukeratzen dira. Sistemak, jasotako esaldiak elkarren jarraian doazen edo ez iragarri behar du. Honen irudikapena ikus daiteke 3.7 irudian.

```
Input = [CLS] I live in [MASK] [SEP] Azpeitia is so [MASK] [SEP]
Label = IsNext
Input = [CLS] [MASK] live in Azpeitia [SEP] I [MASK] trains [SEP]
Label = NotNext
```



3.6 Irudia: Maskaradun Hizkuntza-Eredua atazaren ilustrazio bat aurre entrenamenduaren atazan.



3.7 Irudia: Hurrengo Esaldiaren Iragarpena atazaren ilustrazio bat aurre entrenamenduan.

3.2.4 BERT eleanitzak: mBERT eta BERTeus

Ereduen tamainari dagokionez bi BERT bertsio daude eskuragarri:

- **BERT-Large:** 24 transformer kodetzaile geruza, 1024 dimentsio eta 16 atentzio buru ditu. Eredu honek dituen parametro kopurua guztira 340M da.
- **BERT-Base:** 12 transformer kodetzaile geruza, 768 dimentsio eta 12 atentzio buru ditu. Eredu honek dituen parametro kopurua guztira 110M da.

Tamaina handiagoko bertsioek emaitza hobeak eskaintzen dituzte, baina konputazio behar handiagoak ere eskatzen dituzte. Ingelesezko ereduak hainbat tamaina desberdinetan daude eskuragarri, BERT eleanitzak, aldiz, BERT-Base bertsioan bakarrik aurkitu daitezke. Eredu eleanitzei dagokionez, hainbat hizkuntza bektore espazio berean dituzten BERT ereduak dira, hau da, esanahi bera duten hizkuntza desberdinetako bi esaldi posizio berean (edo behintzat oso gertu) egongo lirateke. Lan honetan bi eredurekin egin da lan.

mBERT (*Multilingual BERT*) 104 hizkuntzatan entrenatu den BERT eredua da[12], eta BERT-Base bertsioan bakarrik dago eskuragarri. Eredu hau entrenatzeko munduko 100 Wikipedia handienak erabiltzen dira, hauen artean euskara.

BERTeus euskaraz erabiltzeko doituta dagoen BERT eredua da [Agerri et al., 2020]. Eredu hau euskal Wikipediaz aparte, hainbat aldizkari eta egunkari erabiliz entrenatu da, mBERT-ek euskaraz dituen token kopurua 35 milioitik 224 milioira eramanez. Entrenamendu corpuseko hobekuntzaz aparte, tokenizazioan ere hobekuntzak egin dira, euskarazko testuen gaineko errendimendua nabarmenki hobetuz.

Lan honetan erabili den eredua BERTeus-en bertsio berriago bat da, non, euskaraz aparte ingelesa eta gaztelera ere gehitu dira [Otegi et al., 2020]. Ingeles eta gaztelera gehitzeko Wikipedia erabili da, corpusak orekatuz hauen tamainan desoreka dela eta. Gainera, tokenizatzeko erabiltzen den hiztegia handitu da 50 milatik 112 mila tokenetara. Bertsio honi **mBERTeus** (*Multilingual BERTeus*) izenaz egingo zaio erreferentzia.

[12] Datu eguneratuak hemen: https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages

4. KAPITULUA

Hizkuntza-ereduetan oinarritutako desanbiguazio-sistema: LMMS

Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation [Loureiro and Jorge, 2019] 2019ko ekainean argitaratutako artikulua da, proiektu hau egiteko oinarri bezala erabili dena. Kapitulu honetan artikuluan aurkezten den lana azaltzen da, aurreko kapituluko informaziotik jarraituz.

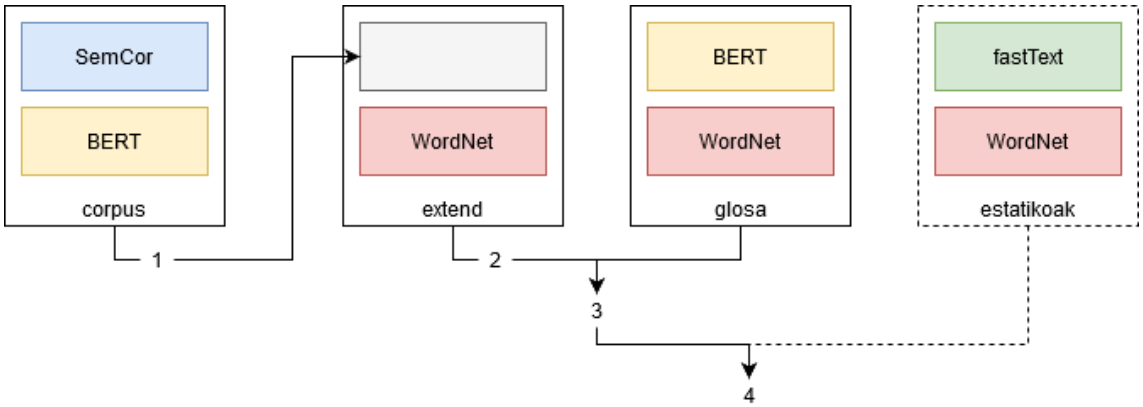
4.1 Adiera-Embeddingak sortzen

Artikuluan aurkezten den helburu nagusia, testuinguruaren menpeko adiera-embeddingak sortzea da (ikusi 3.2.1 atala), WordNeteko adieren estaldura osoa lortuz (200,000 baino gehiago). Prozesu hau lau atazatan banatzen da 4.1 irudian ikus daitekeen bezala. Artikuluaren egileek duten GitHub-ean[1] azaltzen da pausoz pauso nola exekutatu prozesu hau; eta exekuzio hau jarraituz BERT eredu ezberdinak erabiliz lor ditzakegu embeddingak.

4.1.1 **1. ataza:** Corpusetik embeddingak entrenatu

Lehenik, SemCor erabiltzen da bertan agertzen diren adiera desberdinekin embeddingak sortzeko. BERT erabiliz, adiera bakoitzaren agerpen bakoitzeko testuingurua emanik

[1] https://github.com/danlou/lmms



4.1 Irudia: Adiera-Embeddingak sortzeko prozesua, lau atazatan banatua: 1) Corpusetik entrenatu, 2) WordNet erabiliz zabaldu, 3) glosak erabiliz hobetu, eta 4) embedding estatikoak gehitu. Azkenengo ataza hautazkoa da.

bektore bat itzultzen du, eta bektore hori dagokion adiera kodea (*sense_key*) erabiliz etiketatzen da; hainbat aldiz agertzen diren adieren kasuan bektoreen batezbestekoa eginez. Modu honetan 33,360 adiera desberdin estaltzen dira, WordNeten daudenen %16 inguru. Adiera hauek orokorrean gehien agertzen direnak dira, ondorioz, zenbait ataza burutzeko nahikoak izan daitezke.

4.1.2 **2. ataza:** Embeddingen estaldura zabaldu WordNet osora

Ataza honetan, WordNeten egitura erabiltzen da embeddingen estaldura zabaltzeko. Prozesu honetan ez dira aurreko azan lortutako embeddingak aldatzen. Hiru fase desberdin burutzen dira estaldura osoa lortzeko helburuarekin, erlazio zehatzetatik orokorretara joanez, fase bakoitzean aurrekoan lortutako embeddingetatik abiatzen da. WordNeteko falta diren adierak $\hat{s} \in W$ izanik, hauek dira kontuan hartzen diren erlazioak:

- **Synsetak:** Fase honetan synset berari erreferentzia egiten dieten adieren batezbestekoa lortzen da, falta diren gainontzeko adiera sinonimoei batezbestekoa esleituz. S erabiltzen da synset batek biltzen duen adiera multzoa adierazteko.

$$if |S_{\hat{s}}| > 0, \qquad \vec{v}_{\hat{s}} = \frac{1}{|S_{\hat{s}}|} \sum \vec{v}_s, \forall \vec{v}_s \in S_{\hat{s}}$$

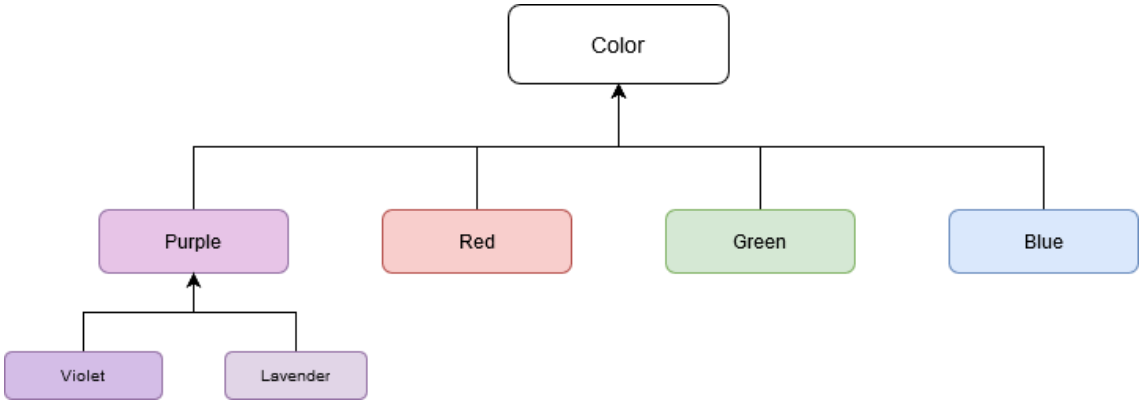
- **Hiperonimoak:** WordNeten egitura sortzeko hiperonimo-hiponimo erlazioa oso garrantzitsua da (ikusi 3.1.2 atala). Kasu honetan erlazio hau erabiltzen da embeddingetan agertzen ez diren synseten errepresentazioak lortzeko, hauen hiponimoen

informazioa erabiliz. Prozesu hau kontzeptu zehatzetatik orokorretara burutzen da, informazioa galdu ez dadin. Adibidez, *kolore* hitzarentzat errepresentaziorik izango ez bagenu, kolore desberdinen errepresentazioen batezbestekoa erabiliko genuke, 4.2 irudian ikusten den bezala. H erabiltzen da hiperonimoen multzoa adierazteko.

$$if|H_{\hat{s}}| > 0, \qquad \vec{v}_{\hat{s}} = \frac{1}{|H_{\hat{s}}|}\sum \vec{v}_{syn}, \forall \vec{v}_{syn} \in H_{\hat{s}}$$

- ***Lexname*ak:** WordNetek eskaintzen duen taldekatze logikoa erabiliz[2], talde hauen barne diren (eta falta diren) synseten errepresentazioak lortu daitezke, gainontzeko synseten batezbestekoa eginez. L erabiltzen da *lexname*-en multzoa adierazteko.

$$if|L_{\hat{s}}| > 0, \qquad \vec{v}_{\hat{s}} = \frac{1}{|L_{\hat{s}}|}\sum \vec{v}_{syn}, \forall \vec{v}_{syn} \in L_{\hat{s}}$$



4.2 Irudia: Hiperonimo-homonimo erlazioa *kolore* hitza erabiliz. Gaineko hitzek azpikoen esanahia bere baitan gordetzen dute.

Hiru fase hauek amaitu ostean, WordNeteko adiera guztien embeddingak lortzen dira. Estaldura ikusteko, 4.1 taulan ikus daiteke fase bakoitzean lortzen den ehunekoa.

Fasea	Estaldura
SemCor	%16.11
+ synset	%26.97
+ hiperonimo	%74.70
+ lexname	%100

4.1 Taula: WordNeteko adieren estaldura Adiera-Embeddingak sortzeko prozesuan.

[2]Talde hauen adibide izan daitezke: animaliak, gorputzeko atalak, posizioa adierazteko kontzeptuak...

4.1.3 **3. ataza:** Hobekuntza glosak erabiliz

Ataza honetan WordNetek eskaintzen duen informazio gehigarria erabiltzen da embeddingak hobetzeko. 3.1.2 atalean azaldutakoa gogoratuz, WordNeten synset bakoitzak kontzeptu bati erreferentzia egiten dio, sinonimoak diren hitzak identifikatzeko aukera emanez, eta glosa bat izanik. Informazio hau esaldi bat sortzeko kateatzen da, eta BERT-i esaldi hau emanik, honen embeddinga itzultzen du. Synset berdina duten hitz sinonimoek embedding berdina izan ez dezaten (adieraren menpeko embeddingak nahi ditugu), hitza bera gehitzen da esaldiaren hasieran. Sortzen diren esaldien egitura ulertzeko, artikuluaren egileek GitHub-en 4.3 irudiko taula honen bidez azaltzen dute.

Sensekey (sk)	Embedded String (sk's lemma, all lemmas, tokenized gloss)
earth%1:17:00::	earth - Earth , earth , world , globe - the 3rd planet from the sun ; the planet we live on
globe%1:17:00::	globe - Earth , earth , world , globe - the 3rd planet from the sun ; the planet we live on
disturb%2:37:00::	disturb - disturb , upset , trouble - move deeply

4.3 Irudia: WordNeteko hitzak eta glosak erabiliz, adiera bakoitzerako sortzen diren esaldiaren adibide batzuk.

Modu honetan, aurreko atazako embedding bakoitzeko bektore berri bat lortzen da espazio berdinean. Emaitza onenak lortzeko helburuarekin, jatorrizko embeddinga eta berria kateatzen dira, adiera bakoitzeko dimentsio bikoitza duen bektore bat utziz.

4.1.4 **4. ataza:** Embedding estatikoak gehitu

Normalean HAD burutzeko, desanbiguatu nahi den hitzaren adiera posibleen artean bakarrik egiten da aukeraketa. Honek embedding guztien artean aukeratzeak baino emaitza hobeagoak eskaintzen ditu, kandidatu kopurua asko murrizten delako. Hala ere, kasu batzutan hau egitea ez denez posible, embedding guztien artean aukeratu behar da adiera egokia, hitza eta kategoria gramatikalaren informazioa erabili gabe. Ataza honi *Uniformed Sense Matching (USM)* deitzen zaio. Kasu hauetan laguntzeko, aukerakoa den ataza honetan fastTextek eskaintzen dituen aurrez entrenatutako embeddingak kateatzen dira aurreko atazako embeddingei. Gure lanean hitz baten adierak bakarrik hartzen direnez kandidatu bezala, eta embedding estatiko hauek hizkuntzaren menpe daudenez, ez da ataza hau burutu embeddingak sortzean.

4.2 Desanbiguazio metodoa

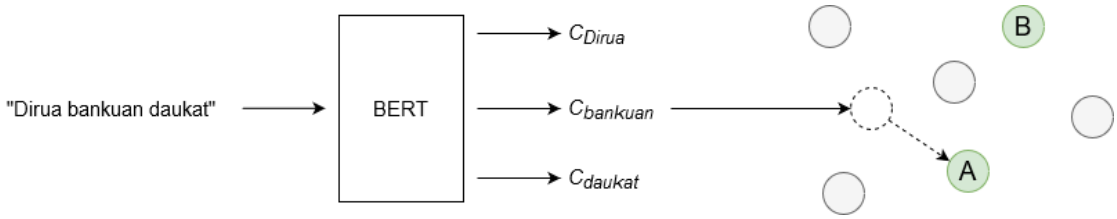
Behin adieren embeddingak sortu eta gero, ebaluazioan hitz baten agerpen bat desanbiguatu nahi denean, testuinguruan oinarrituta errepresentazio bektorea kalkulatzen de BERT hizkuntza-eredua erabilita. Ondoren, aurreko atalean sortu diren adiera-embeddingekin konparatzen da dagokion adiera esleitzeko. Prozesu hau ikus daiteke 4.4 irudian.

Adiera posibleen artean aukeratzeko k Nearest Neighbours (kNN) sailkatzailea erabiltzen da; $k = 1$ izanik balio lehenetsia, hau da, gertuen dagoen adiera aukeratzen da emaitza bezala. Gainontzeko bektoreetara dagoen distantzia kalkulatzeko, kosinu-antzekotasunaren formula erabiltzen da, bi bektorek (x eta y) sortzen duten angeluaren araberako distantzian oinarritua:

$$cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{||\boldsymbol{x}||\cdot||\boldsymbol{y}||}$$

Kontuan izan behar da 4.1.3 atazan sortu diren embeddingen dimentsioa bikoitza dela; kasu honetan, desanbiguatu nahi den hitzaren bektorea bikoiztu beharko da, bektorea berriz kateatuz bukaeran distantzia kalkulatu ahal izateko.

4.1.4 atalean aipatu den bezala, ataza honetan ez da ebaluazioa embedding guztien artean burutzen. Kandidatu posibleen kopurua gutxitzeko, eta emaitza hobeak lortzeko, hitz horrek izan ditzakeen adierak, eta hauen artean kategoria gramatikal bera duten embeddingak bakarrik aukeratzen dira. Embeddingak identifikatzeko erabiltzen den *sense_key* kodeak hitza eta kategoria gramatikalaren informazioa kodetzen duenez, oso modu errazean aukera daitezke kandidatuak.



4.4 Irudia: kNN-ren erabilera HAD ataza burutzeko. Berdez *banku* hitzak izan ditzakeen adieren bektoreak agertzen dira.

5. KAPITULUA

HAD eleanitza

Kapitulu honetan LMMS artikuluan eginiko lana beste hizkuntza desberdinekin erabili ahal izateko egin diren moldaketak aurkezten dira. Helburua, ingelesezko testuak erabiliz entrenatu diren adiera-embeddingetaz baliatuz, beste hizkuntzatan HAD burutzea da[1].

Lehenengo atalean, adiera-embeddingak hizkuntza baten menpe ez dauden Synset-Embeddingetan bihurtzeko prozesua azalduko da. Bigarren atalean, ebaluatzeko erabiliko den euskarazko datu-multzoan, EusSemcor, egin behar izan diren aldaketak komentatzen dira; hauek egiteko arrazoia eta prozesua azalduz.

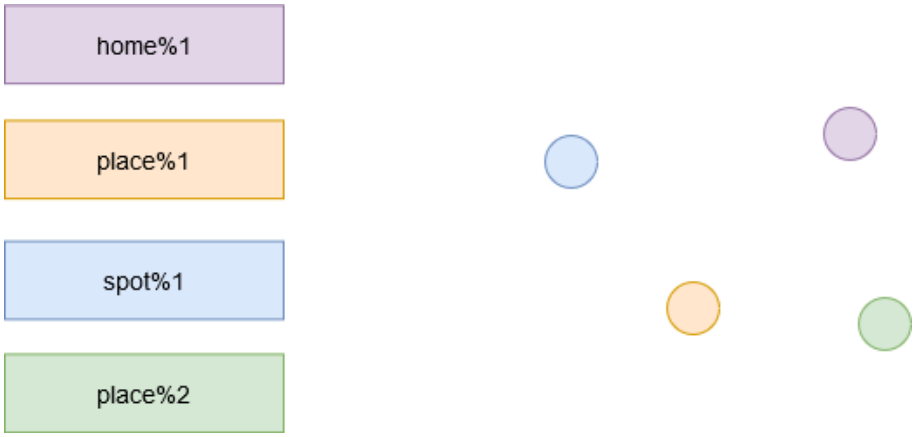
5.1 Synset-Embeddingak lortzen

LMMS artikulua jarraituz lortu diren adiera-embeddingak ezin dira zuzenean erabili beste hizkuntza bateko testua desanbiguatzeko. Honen arrazoia, adiera-embeddingen formatuan datza, 5.1 irudian ikusten den bezala, hitz bakoitzaren adiera bakoitzeko bektore bat dago. Honek asko errazten du ingelesezko testuetan hitza erabiliz adiera posibleak aukeratzea, baina ezinezkoa egiten du beste edozein hizkuntzatan erabiltzea, adierak hizkuntzaren menpe daudelako, synsetak ez bezala.

Eginiko aldaketak hobeto ulertzeko demagun 5.1 irudiko embeddingak bakarrik ditugula. Kasu hipotetiko honetan lau embedding desberdin daude, *home* eta *spot* hitzentzat bana, eta *place* hitzarentzat bi. *"Deliver the package to my **place**"* esaldia jasotzekotan, Em-

[1] Erabiltzen den BERT bertsioaren arabera; mBERT-en kasuan 104 hizkuntza desberdinetan gehienez.

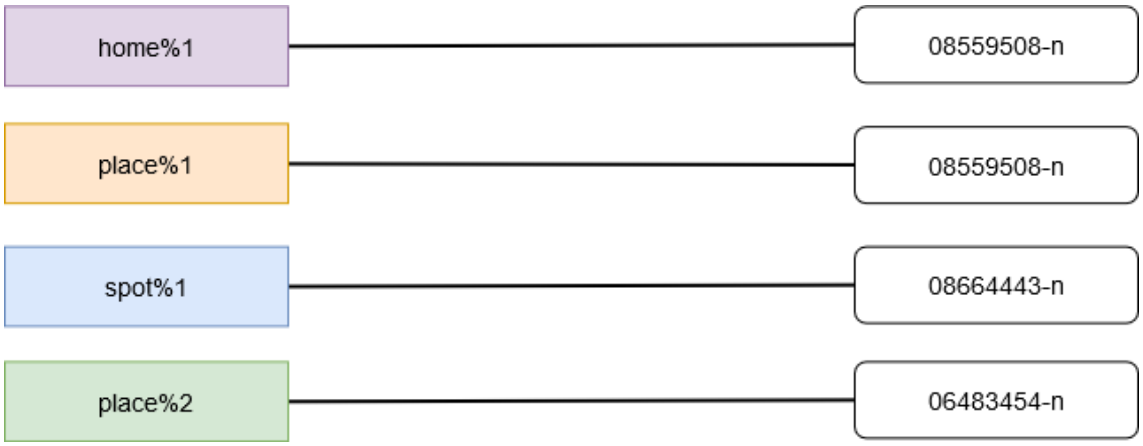
beddingen zerrenda iteratuko litzateke *place%1* eta *place%2* aukeratuz desanbiguatzeko orduan kandidatu bezala.



5.1 Irudia: Adiera-embeddingen formatuaren ilustrazio ba. Hitzen adiera bakoitzeko bektore desberdin bat.

5.1.1 Adierak synsetetara mapatu

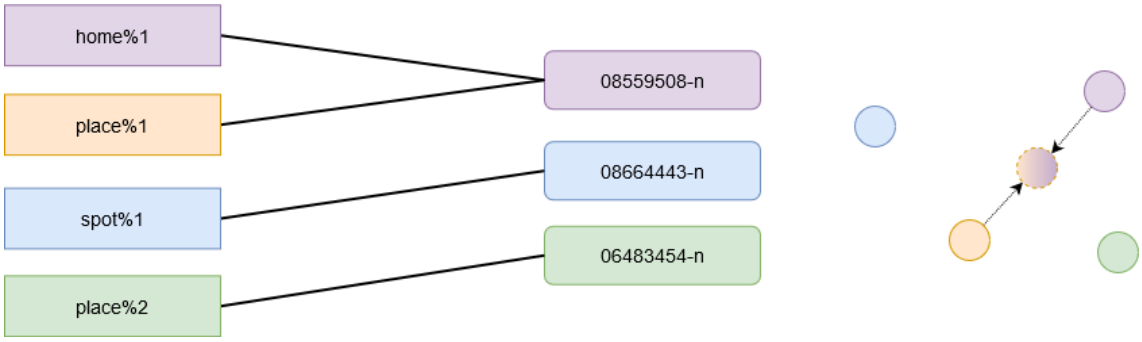
Embedding eleanitzak lortzeko, adiera bakoitza dagokion synsetarekin lotu da. 3.1.2 atalean komentatutako gogorarazteko, synsetak kontzeptuak bezala ikus daitezke; hainbat hitz desberdinek izan dezakete synset bera (hitzak sinonimoak dira), eta adierak ez bezala, hizkuntza desberdinen artean lotu daitezke. Hau lortzeko WordNeten *sense index* [2] fitxategia erabili da, 5.2 irudian agertzen den moduan.



5.2 Irudia: WordNeten sense index erabiliz adiera bakoitza synset bati mapatzen zaio.

[2] https://wordnet.princeton.edu/documentation/senseidx5wn

Ikus daitekeen bezala, orain adiera bat baino gehiago izan dezakegu lotuta synset bati. Adibidean *home%1* eta *place%1* synset berdinari geratzen dira lotuta, hau da, kontzeptu berari egiten diote erreferentzia. Synset bakoitzak izango duen bektore berria kalkulatzeko, dagokien adieren bektoreen batezbestekoa erabiliko da 5.3 irudian irudikatzen den moduan.

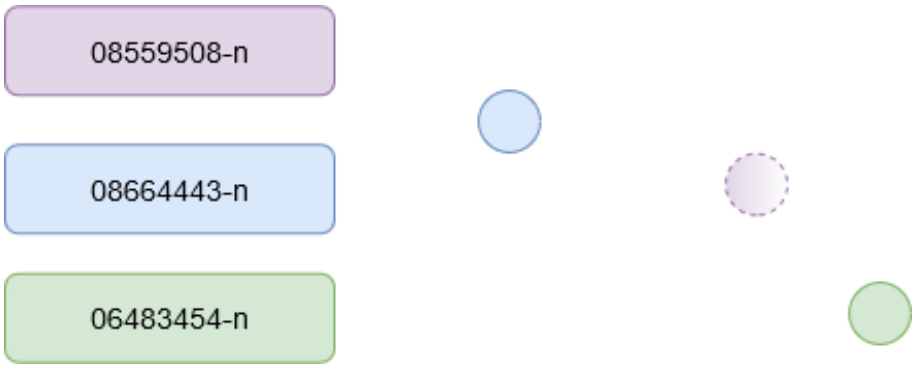


5.3 Irudia: Hainbat adierek izan dezakete synset bera. Honen balioa kalkulatzeko, adieren batezbestekoa egiten da.

Batezbestekoa kalkulatzeko orduan aukera desberdinak kontuan hartu dira, ahalik eta emaitza onenak lortzeko asmoarekin. Lehen aukera batezbesteko aritmetikoa egitea izango litzateke. Hau beharbada ez da konponbide onena, ez delako kontuan hartzen adierek izan ditzaketen agerpen kopurua desberdina dela, eta ondorioz oso gutxitan erabiltzen diren hitzek garrantzia handia izango dutela bektore berria kalkulatzeko unean.

Synsetak mapatzeko erabili den WordNet *sense index* fitxategiak adiera desberdinen agerpen kopuruari buruzko informazioa ematen du. Zenbaki honek adiera bakoitza hainbat testutan zenbat aldiz etiketatu den adierazten du. Informazio honetaz baliatuz, batezbesteko aritmetiko haztatua egitea izango litzateke bigarren aukera posiblea. Hala ere, adiera askotan gertatzen da ez dela inoiz etiketatu, agerpen kopurua zero izanik, eta ondorioz batezbestekoa egitean informazio hau galduko litzateke; honen eragina aztertzeko hirugarren aukera bat gehitu da, non adiera-embeddingetan agertzen diren adieren agerpen kopurua gehi bat leunketa bidez kalkulatu den.

Emaitza onenak lortzeko, eta estaldura handiena lortzeko helburuarekin, hirugarren aukera erabili da adiera-embeddingak bihurtzeko. Behin Synset-Embeddingak kalkulatuta 5.4 irudian agertzen den moduko zerbait lortzen da, synset bakoitzarentzat bektore bat izanik.

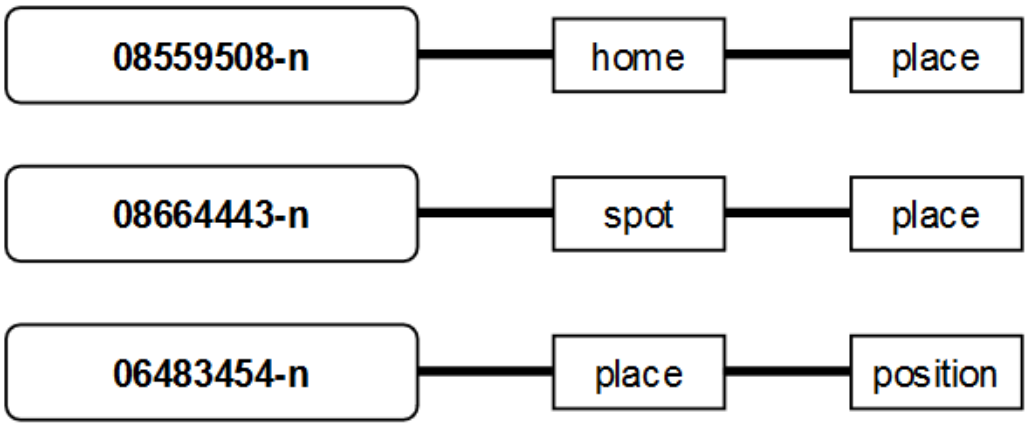


5.4 Irudia: Synset-Embeddingen formatuaren ilustrazio bat. Synset bakoitzeko bektore bat. Lerro etenaz irudikatutako bektorea batezbestekoa eginez kalkulatu da.

5.1.2 Synseten hiztegiaren sorkuntza

Orain arte egindakoak ez du konpontzen, ordea, hasieran azaldu den arazoa: hitzaren arabera kandidatuak aukeratzeko orduan ez dakigu zein synset dagozkion jaso dugun hitzari. Hau egin ahal izateko erabili nahi den hizkuntza bakoitzeko hiztegi bat eraikiko da, hitz bakoitzak izan ditzakeen synset posible guztiak zein diren jakiteko.

Hiztegi hau sortzeko MCR 3.0 erabili da, ingelesezko eta euskarazko bertsioetarako, nahiz eta ez litzatekeen arazorik egongo MCR-k onartzen dituen gainontzeko hizkuntzekin lan egiteko (espainiera, galiziera, katalan eta portugesa).



5.5 Irudia: Ingelesezko hiztegiaren ilustrazio bat. Synset bakoitzak izan ditzakeen hitzen zerrenda gordetzen du.

Hiztegiaren erabilerak emaitzetan eragina izango duela aurreikusi daiteke, aukeratzen den kandidatu kopurua handitu daitekeelako kasu batzutan. Lehengo adibidera itzuliz, *"Deliver the package to my **place**"* esaldian 'place' hitza desanbiguatu nahiko balitz, 5.5 hiztegira joan, eta 'place' hitza duten synset guztiak hartuko lirateke kandidatu bezala. Nahiz eta jatorrizko adiera-embeddingetan 'place' hitzarentzat bi kandidatu bakarrik izan, orain hiru izango genituzke, 'spot' hitzaren sinonimo gisa erabil daitekeelako. Honek ez luke eraginik izango 4.1.2 atazan lortutako embeddingetan, WordNeten estaldura guztia baitute.

5.2 Ingelesezko datu-multzoen moldaketa

Synset-Embeddingak lortu ondoren, hauek erabiltzeko beharrezkoa da orain arte erabili diren datu-multzoak moldatzea ebaluatu aurretik. Datu-multzoen urre patroia adierak erabiliz etiketatuta dagoenez ezin dugu erabili Synset-Embeddingekin. Synsetak erabiliz etiketatzeko 5.3 irudian azaldu den prozesu bera errepikatu behar da, hau da, adibide bakoitzak duen adiera dagokion synsetari mapatuz. Kontuan izanik adibide bakoitzak adiera bat edo gehiago izan ditzakeela.

Eginiko esperimentuak ebaluatu ahal izateko, eta proiektua eleanitz bihurtzeak duen eragina konparatu ahal izateko, 3.1.5 atalean azaltzen diren datu-multzo guztiak moldatu dira. Emaitza hauek hurrengo atalean xehetasun gehiagorekin azaltzen diren 6.1 eta 6.3 tauletan ikus daitezke.

5.3 EuSemcor: datu-multzoa sortzen

EuSemcor, 3.1.4 atalean azaldu den bezala, euskarazko SemCor corpusa da, hau da, adierak eskuz etiketatuta dituen corpusa. Euskarazko izen bakoitzeko XML fitxategi bat dago, agerpenak EusWordNet 1.6-ko synsetekin etiketatuta.

EuSemcor erabili ahal izateko, [Raganato et al., 2017] artikuluan eskaintzen diren datu-multzoen formatura doitzea erabaki da, ebaluatzeko programaren kodean aldaketarik egin beharrik ez izateko. Datu-multzo hauek SemEval-13 all-words atazako XML eskema erabiltzen dute formatu bezala. Formatu aldaketa egin aurretik, ordea, datu-multzoaren aurreprozesaketa egin behar izan da, akats batzuk konpondu, banaketa egin eta synsetak WordNet 3.0 bertsiora mapatzeko.

5.3.1 Garbiketa, esaldi errepikatuak kendu

Datu-multzoaren begirada bat egin ondoren, datuetan bistakoak diren akats batzuk ikus daitezke; hala nola, esaldi batzuen hasieran agertzen den testua[3], komatxoak adierazteko modu desberdinak[4], edo puntuz banatutako esaldi hutsak. Orokorrean akats hauek ez dute desanbiguazioan eraginik, hala ere, formatu aldaketa egitean gehienak zuzendu ditugu.

Etiketatutako corpusetan ohikoa da hitz batentzat adiera bat baino gehiago zuzenak direla erabaki izana. EuSemcor-en kasu hauetan XML fitxategietan hainbat instantzia errepikatu sortzen dira, 5.6 irudian ikusten den bezala, non *herri* hitzak bi adiera izan ditzakeen.

```
<instance id="herri.IZE.60" docsrc="eebs.450640531.txt.xml" topic="eebs" sentsrc="82" positsrc="1" sentn="2" positn="0">
<answer instance="herri.IZE.60" senseid="06382213"/>
<context>
Badira beste itxura bat eta mihi gehiago dituzten karrakak ere .
Hauek , noski , zarata handiagoa egiten dute .
<head>Herri</head> batzuetan eliza berak , bere zerbitzurako izan ditu bere jabegoko diren halako karrakak .
IGURTZITAKO BOTILAK .
Goilare edo bestelako tresnaz igurtzitako azal zimurra duten botilak , toki askotan , erritmoa markatzeko soinutresna gisa erabili izan dira .
</context>
</instance>

<instance id="herri.IZE.61" docsrc="eebs.450640531.txt.xml" topic="eebs" sentsrc="82" positsrc="1" sentn="2" positn="0">
<answer instance="herri.IZE.61" senseid="06383813"/>
<context>
Badira beste itxura bat eta mihi gehiago dituzten karrakak ere .
Hauek , noski , zarata handiagoa egiten dute .
<head>Herri</head> batzuetan eliza berak , bere zerbitzurako izan ditu bere jabegoko diren halako karrakak .
IGURTZITAKO BOTILAK .
Goilare edo bestelako tresnaz igurtzitako azal zimurra duten botilak , toki askotan , erritmoa markatzeko soinutresna gisa erabili izan dira .
</context>
</instance>
```

5.6 Irudia: EuSemcor-eko XML fitxategien zati bat. Hitz batek hainbat adiera dituenean instantzia errepikatuta agertzen da.

Honek sor ditzake arazo gehien datu-multzoaren entrenamendu eta test banaketa egiteko unean, esaldi bera banatuta geratu daitekeelako bi partiziotan, synset desberdinak izanik soluzio bezala. Arazo hau konpontzeko, oraingoz, esaldi errepikatuak identifikatu eta gainontzeko agerpenak ezabatu dira, esaldi bakoitzeko adibide bakarra utziz hainbat synsetekin etiketatuta. Etorkizunean EuSemcor-en bertsio berri bat sortuko balitz, egokiena hau kontuan hartzea izango litzateke XML fitxategiak sortzean, aurreprozesaketako pausu hau ekidin ahal izateko. Soluzioari dagokionez, | ikurra erabili da synsetak banatzeko. 5.7 irudian ikus daiteke nola geratzen den 5.6 irudiko kasua esaldi errepikatuak ezabatu ondoren.

```
<instance id="herri.IZE.60" docsrc="eebs.450640531.txt.xml" topic="eebs" sentsrc="82" positsrc="1" sentn="2" positn="0">
<answer instance="herri.IZE.60" senseid="06382213|06383813"/>
<context>
Badira beste itxura bat eta mihi gehiago dituzten karrakak ere .
Hauek , noski , zarata handiagoa egiten dute .
<head>Herri</head> batzuetan eliza berak , bere zerbitzurako izan ditu bere jabegoko diren halako karrakak .
IGURTZITAKO BOTILAK .
Goilare edo bestelako tresnaz igurtzitako azal zimurra duten botilak , toki askotan , erritmoa markatzeko soinutresna gisa erabili izan dira .
</context>
```

5.7 Irudia: Adiera anitz dituen hitz bat, errepikatutako agerpenak ezabatu ondoren.

[3] @0007@ 0046 5 moduko testua ageri da esaldi batzuen hasieran.

[4] Komatxoak adierazteko " eta « erabiltzen dira. Ingelesezko esaldietan “ erabiltzen dira.

5.3.2 Entrenamendu/garapen/test banaketa

HAD atazarako garrantzitsua da datu-multzoaren banaketa egoki bat edukitzea, bai entrenamendu faserako erabili nahiko balitz, bai lortu dugun sistemaren errendimendua ebaluatzeko garapen eta test partizioen bidez. Banaketa aurreko pausuan emaitza moduan geratzen den EuSemcor-en bertsioaren gainean egin da, honako ezaugarriak kontuan izanik:

- Hitzak 100 agerpen edo gehiago izatea.
- Hitzak 2 adiera edo gehiago izatea.
- 50, 25, 25 adibide proportzioak entrenamendu, garapen eta test partizioetan hurrenez hurren.
- Estratifikatuta, klase minimoak gutxienez 3 adibide baditu. Bestela, estratifikatu gabe.

Banaketa ondoren geratzen diren hitzen zerrenda, eta datu-multzo bakoitzean dagoen adibide kopurua eranskineko A.1 taulan ikus daiteke.

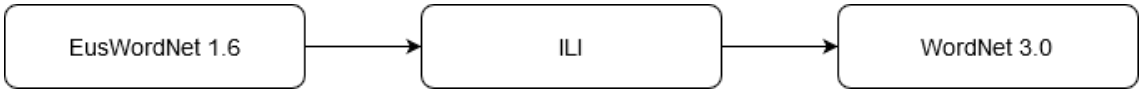
5.3.3 Synsetak WordNet 1.6-tik 3.0-ra mapatu

EuSemcor EusWordNet 1.6 erabiliz etiketatu zen, baina aurreko atalean azaldu den bezala, ebaluatzeko erabiltzen ditugun Synset-Embeddingak WordNet 3.0 bertsioa erabiltzen dute kontzeptu desberdinak adierazteko. Synsetak ez dira bi bertsio hauen artean trukagarriak; ez hizkuntza desberdina delako (EusWordNetek eta ingelesezko WordNetek synsetak partekatzen dituzte), Wordnet 1.6 eta 3.0 artean synsetak desberdinak direlako baizik; eta beraz beharrezkoa da EuSemcor-eko synsetak WordNeten 3.0 bertsiora mapatzea.

Posible izango litzateke ere Synset-Embeddingak mapatzea 1.6 bertsiora, baina kontuan izanda ingelesezko datu-multzoak 3.0 bertsioan daudela, eta 3.0 dela gaur egun deskargatu daitekeen WordNeten azken bertsioa, bide hau ez jarraitzea erabaki da. WordNetek bere bertsioen artean synsetak mapatzeko fitxategiak eskaintzen ditu[5], baina fitxategi hauek bertsio konkretu batzuetara mugatuta daude; hauen artean ez dago 1.6.

[5] https://wordnet.princeton.edu/documentation/sensemap5wn

Mapaketak egiteko *Collaborative Interlingual Index* (ILI) erabili da. ILI-k hizkuntza desberdinen artean kontzeptuak mapatzeko aukera eskaintzen du, hizkuntza desberdineko WordNetak lotzeko erabiltzen baita. WordNeteko synset desberdinak ILI kode konkretu bati estekatuta daude, eta WordNeten bertsio guztiak mapatuta daudenez, gure atazarako aukera bikaina eskaintzen du. 5.8 irudian ikus daiteke mapatzeko prozesua nola gauzatzen den.



5.8 Irudia: WN 1.6-tik 3.0-ra synsetak mapatzeko egiten den prozesua.

Hala ere bide hau ez da beti egingarria, eta batzutan ezinezkoa suertatzen da mapaketa burutzea. Hau gertatzeko arrazoiak hainbat izan daitezke: euskarazko synset batek ingelesezko synset baliokide bat ez edukitzea, WordNeten bertsio berriagoetan synset bat ezabatu edo aldatu izana, ILI-n kontzeptu hori ez agertzea, gaizki etiketatu egotea... Orokorrean hau ez da askotan gertatzen; banaketa egin ondoren sortu den datu-multzoko pare bat adiera geratu dira mapatu gabe arrazoi hauetako batengatik. Hitz hauek A.2 taulan kontsultatu daitezke.

6. KAPITULUA

Esperimentuak eta emaitzak

Kapitulu honetan LMMS artikuluan lortzen diren emaitzetatik abiatuz, proiektua garatu ahala ebaluatu ahal izateko hartu diren neurriak azaltzen dira, eta ingelesezko datu-multzoen gainean lortutako emaitzak aurkezten dira. Azkenik, EuSemcor datu-multzoan (ikusi 5.3) ebaluatu da sistema.

6.1 Ebaluatzeko neurriak

Sistema bat ebaluatzeko hainbat neurri erabil daitezke, normalean gehien erabiltzen direnak doitasuna (*precision*), estaldura (*recall*), eta F-neurria (*F-score*) izanik. Doitasunak sistemak iragarri ditzakeen artean asmatutako elementuen proportzioa adierazten du.

$$doitasuna = \frac{asmatutakoak}{iragarritakoak}$$

Estaldurak, berriz, urre patroiko elementuen artean asmatutako proportzioa adierazten du.

$$estaldura = \frac{asmatutakoak}{urrepatroia}$$

F-neurria doitasunaren eta estalduraren batezbesteko harmonikoa kalkulatuz lortzen da; oso erabilia izanik hizkuntzaren prozesamenduko atazetan.

$$F-neurria = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura}$$

Adiera-embeddingek WordNeteko adieren estaldura osoa dutenez, ebaluatzeko garaian hiru neurri hauek berdinak dira. Hala ere, emaitzetan WordNeten estaldura osoa ez duten corpusetik entrenatutako embeddingak gehitu direnez (ikusi 4.1.1 atala), F-neurria erabili da beti emaitzak kalkulatzeko.

6.2 Euskarazko oinarri lerroa: adiera usuena

Euskarazko datu-multzoaren gainean probak egitean, emaitzak konparatzeko oinarri lerro bat ezartzea komenigarria izaten da. HAD-en oinarri lerro bezala erabili ohi den metrika adiera usuena edo *Most Frequent Sense (WSD)* izan ohi da, non hitz batek izan ditzakeen adieren artean agerpen gehien dituena aukeratzen da.

Adiera usuena zein den kalkulatzeko EuSemcor-en entrenamendu partiketako agerpenak kontatu dira, agerpen gehien dituen adiera erabiliz test partiketa ebaluatzean. Hitz bakoitzaren adiera usuena zein den A.1 taulan ikus daiteke.

6.3 Emaitzak

6.3.1 Garapen faseko emaitzak

Garapen fasean ingelesezko datu-multzoen gainean egin dira probak (ikusi 3.1.5 atala), eginiko aldaketen eragina aztertzeko. Bost datu multzoen bildura ere gehitu da tauletan ALL izenarekin.

Hizkuntza ezberdinetan dauden testuak desanbiguatu ahal izateko egin beharreko lehen gauza hizkuntza-eredua aldatzea da, gure kasuan BERT ereduaren bertsio eleanitz bat erabiltzea. Hau egitean emaitzek okerrera egitea espero daiteke, eta okerragotze hau gertatzeko arrazoi bat baino gehiago egon daitezke. Alde batetik, ingelesez bakarrik entrenatutako eredu bat hainbat hizkuntza ezberdin erabiliz entrenatutako bat baino zehatzagoa izango da ingelesezko HAD egitean. Beste aldetik kontuan izan behar da LMMS artikuluko jatorrizko emaitzak tamaina handiagoa duen BERT-Large erabiliz lortu direla; esku-

ragarri dauden BERT-en bertsio eleanitzak, ordea, BERT-Base formatuan bakarrik daude eskuragarri (ikusi 3.2.4), eta ondorioz emaitzetan eragin zuzena du.

BERT eredu bakoitzerako embeddingen hiru bertsio erabili dira:

- 4.1.1 - Corpusetik entrenatutako embeddingak, ereduaren izena erabiliz adierazita.
- 4.1.2 - (WordNet osora) zabaldutako embeddingak, **'Ext'** erabiliz adierazita.
- 4.1.3 - Glosak erabiliz hobetutako embeddingak, **'Glosa'** erabiliz adierazita.

Eredua	**Senseval2**	**Senseval3**	**SemEval 2007**	**SemEval 2013**	**SemEval 2015**	**ALL**
MFS	66.8	66.2	55.2	63.0	67.8	65.2
BERT-Large Ext	75.4	74.0	66.4	72.7	75.3	73.8
BERT-Large Glosa	76.3	75.6	68.1	75.1	77.0	75.4
BERT-Base	69,3	69,0	63,6	63,3	69,0	67,5
BERT-Base Ext	74,5	71,8	65,3	71,3	74,0	72,4
BERT-Base Glosa	74,7	72,9	66,8	72,7	73,5	73,1
mBERT	67,0	66,0	62,0	61,8	68,1	65,4
mBERT Ext	71,9	69,0	64,0	69,4	72,6	70,2
mBERT Glosa	70,2	66,5	59,8	69,7	68,4	68,2
mBERTeus	68,2	67,2	58,9	61,3	67,8	65,8
mBERTeus Ext	72,9	70,2	60,7	68,7	72,1	70,4
mBERTeus Glosa	72,8	70,5	62,6	68,1	72,2	70,4

6.1 Taula: Adiera-embeddingak erabiliz, ingeleseko eta BERT elebidun bertsio desberdinekin lortutako emaitzak. BERT-Large LMMS sistemaren jatorrizko emaitzak dira; gainontzekoak lan honetako emaitzak.

Tamaina txikiagoko BERT eredu bat erabiltzearen eragina ikusi ahal izateko, BERT-Base izeneko bertsio bat gehitu da 6.1 taulan. Bertsio hau, ingelesez bakarrik entrenatutako jatorrizko ereduaren bertsio txikiago bat da, BERT eleanitzen tamaina berdina duena.

Emaitzetan, glosak erabiliz hobetutako embeddingei erreparatuz, eredu txikiago bat erabiltzeak (BERT-Base vs BERT-Large) eragin aipagarria duela ikusten da 2 puntuko erorketarekin, hala ere, eragin nabarmenena eredu eleanitz bat erabiltzean gertatzen da, 3 puntu inguruko jaitsierarekin datu-multzo guztiak (ALL) erabiltzean. Glosa erabiliz hobetutako embeddingek ez dute hobekuntza nabarmenik eskaintzen eredu eleanitzetan, mBERT-en kasuan 2 puntuko jaitsiera ikusten da, emaitza onenak zabaldutako embeddingek eskainiz. Gainera, mBERTeus-en abantaila ikus daiteke mBERT-ekin konparatuz, ziurrenik, hizkuntza gutxiago erabiliz entrenatu izanagatik. Guztira 5 puntuko erorketa ikusten da LMM-ko jatorrizko emaitzekin alderatuz.

Synset-Embeddingak sortzean, adieren batezbestekoa egiteko hiru aukera ezberdin aurkeztu dira, 5.3 irudiaren jarraian. Garapenean zehar aukera ezberdinen aldea zenbaterainokoa den aztertzeko ingelesezko datu multzo guztien batezbestekoa (ALL) erabili da, 3.2.3 atalean azaldutako mBERTeus-en bi bertsio desberdin erabiliz ebaluatuz, orain arte emaitza onenak eskaini dituen eredu eleanitza delako. 6.2 taulan ikusten den bezala, batezbestekoak emaitza baxuak eskaintzen ditu haztatuarekin alderatuz. Batezbesteko haztatuaren kasuan, agerpen guztiak mantentzeak ez du abantaila handirik eskaintzen SemCor erabiliz entrenatutako embeddingetan; WordNet osoa estaltzeko zabaldutako embeddingetan ordea, adiera guztiak mantentzeak emaitza hobeak eskaintzen ditu.

Embedding	**mBERTeus**	**mBERTeus Ext**
Adiera-Embedding	65,8	70,4
Synset-Embedding - Batezbesteko sinplea	62,9	66,8
Synset-Embedding - Haztatua, 0 agerpeneko adierak ezabatu	**66,5**	68,6
Synset-Embedding - Haztatua, Agerpen kopurua +1 leunketa	66,4	**70,2**

6.2 Taula: Synset-Embeddingak sortzeko batezbestekoa egiteko aukera desberdinen konparaketa ingelesezko datu multzoen gainean.

Synset-Embeddingak lortu ondorengo emaitzak agertzen dira 6.3 taulan. Synset-Embedding hauek jatorrizko adiera-embeddingen hizkuntzaren menpe ez dauden bertsioa dira. Nola lortu diren 5.1 kapituluan azaldu da.

Eredua	**Senseval2**	**Senseval3**	**SemEval 2007**	**SemEval 2013**	**SemEval 2015**	**ALL**
mBERT	67,3	67,8	61,2	63,4	67,6	66,2
mBERT Ext	**71,5**	**69,8**	**62,6**	69,0	**72,2**	**70,0**
mBERT Glosa	70,0	66,2	58,2	**70,6**	69,0	68,3
mBERTeus	68,0	67,1	62,7	63,0	68,6	66,4
mBERTeus Ext	**72,3**	**69,5**	**63,7**	**68,4**	**72,1**	**70,2**
mBERTeus Glosa	71,8	**69,5**	62,2	67,6	**72,1**	69,7

6.3 Taula: Synset-Embeddingak erabiliz, BERT bertsio desberdinekin lortutako emaitzak

Synset-Embeddingen emaitzak adiera-embeddingenekin alderatuz, orokorrean emaitzak mantentzen direla ikusten da zabaldutako embeddingetan; nahiz eta synsetak adierak baino informazio orokorragoa kodetzen duten. Zabaldutako embeddingen abantaila mantentzen da gainontzeko bertsioekiko, argi utziz glosak erabiliz hobetutako embeddingek ez dutela balio emaitzak hobetzeko BERT eleanitzetan.

6.3.2 Euskarazko emaitzak

Aurreko kapituluan sortu dugun EuSemcor datu-multzoaren *test* partizioko ebaluazioa agertzen da 6.4 taulan. Datu-multzo hau EuSemcor-en bertsio originalaren aldakuntza da, aurreko probetan erabiltzen diren ingelesezko datu-multzoen formatura egokituta. Aldaketa hauek nola egin diren 5.3 atalean azaltzen dira sakontasunez.

Eredua	**EuSemcor**
MFS	67,1
mBERT	**59,8**
mBERT Ext	57,1
mBERT Glosa	55,4
mBERTeus	**61,8**
mBERTeus Ext	61,7
mBERTeus Glosa	61,0

6.4 Taula: EuSemcor datu-multzoaren ebaluazioa BERT bertsio desberdinekin.

Emaitzei begiratuz, mBERTeus-ek lortzen ditu emaitza onenak bertsio desberdin guztietan mBERT-ekin alderatuz. Hau gertatzeko arrazoi desberdinak daude.

Alde batetik, mBERT 104 hizkuntza erabiliz entrenatu da, mBERTeus aldiz, hiru hizkuntzatan bakarrik dago entrenatuta (euskara, gaztelera eta ingelesa). Beraz, euskarazko esaldiekin lan egiteko hobeto moldatzen da. Hau ikus daiteke, adibidez, *etxerantz* hitza bi BERT ereduek nola tokenizatzen duten alderatuz. mBERTeus-en tokenizazioan jatorrizko hitzen egitura hobeto mantentzen da, eta gramatika aldetik zentzu gehiago dauka.

```
mBERT:     Et ##xer ##ant ##z
mBERTeus:  Etxera ##ntz
```

Beste aldetik, 3.2.4 atalean ikusi den bezala, mBERTeus entrenatzeko mBERT baino euskarazko corpus handiago bat erabili da, mBERT-ek erabiltzen duen euskal Wikipediaz aparte, hainbat aldizkari eta egunkari gehitu dira. Entrenatzeko erabili den token kopurua 224 milioikoa da, euskal Wikipediak dituen 35 milioien aldean.

Arrazoi hauengatik argi geratzen da mBERTeus egokiagoa da ataza konkretu hau burutzeko, ingelesez ikasitako errepresentazioak erabili euskarazko HAD burutzeko, alegia. Hala ere, sistema guztiz eleanitz bat sortzea helburu izanez gero, mBERT-ek eskaintzen duen hizkuntza kopurua askoz handiagoa dela kontuan izan behar da.

Embeddingak sortzeko metodoari dagokionez, garapen fasean ikusi da glosak erabiliz hobetutako embeddingak ez dutela abantailarik eskaintzen BERT eredu eleanitzetan. Ondorio bera ikus daiteke euskarazko emaitzak aztertzean. Corpusetik bakarrik entrenatutako, eta zabaldutako embeddingen artean, lehenengoak eskaintzen dituzte emaitza onenak. Hala ere, kontuan izanik EuSemcor-ek euskarazko 400 izen ohikoenak bakarrik hartzen dituela, normala da zabaldutako embeddingek abantailarik ez eskaintzea, azken finean EusWordNeten dauden synsetetatik ohikoenak daudelako EuSemcor-en.

Garapen faseko emaitzak, eta WordNeten estaldura osoa duten embeddingen onurak kontuan izanda, **mBERTeus Ext** izango litzateke aukera zentzuzkoena; nahiz eta euskaraz ebaluatu ondoren ikusi den beste bertsio batek emaitza apur bat hobeak eskaini ditzakeela.

7. KAPITULUA

Ondorioak eta etorkizuneko lanak

Kapitulu honetan proiektuan zehar lortu ditugun ondorioak azaltzen dira; emaitzen inguruko hausnarketa eta ondorio pertsonalak. Azkenik, proiektuari jarraipena emateko, etorkizunean egin daitekeen lana proposatzen da.

7.1 Ondorioak

7.1.1 Lortutako emaitzak

Proiektuan zehaztutako helburuak betetzea lortu dugu, bai LMMS artikuluko emaitzak erreplikatuz eta euskarazko testuak desanbiguatzeko balio duen sistema eleanitz bat sortuz.

Sistemari dagokionez, hizkuntza-eredu eleanitz desberdinak (3.2.3) erabili dira jatorrizko adiera-embeddingak lortzeko; eta embedding hauek hizkuntza baten menpe ez dauden Synset-Embedding bihurtzeko prozesua garatu da, fase bakoitzean sistema ebaluatuz galerak identifikatu ahal izateko. BERT eredu txikiago bat erabiltzeak 2 puntuko jaitsiera, eta eredu eleanitz bat erabiltzeak 3 puntuko jaitsiera dakarrela ikusi da.

Euskarazko HAD atazarako datu-multzo bat sortu da eskuz etiketatutako EuSemcor corpusetik (5.3), bertako datuen garbiketa eginez, eta entrenamendu/garapen/test banaketa eginik. Datu-multzo hau etorkizuneko lanetarako erabilgarria da, etorkizuneko sistemen errendimendua elkarren artean konparatu ahal izateko. Gainera, datu-multzoa ingelesezko

gainontzeko datu-multzoen formatura egokitu da, ebaluazioan hizkuntzen artean desberdintasunik egon ez dadin.

Euskarazko emaitzei dagokionez, espero zen balio-tartean daude, hala ere, emaitzek ez dutenez lortu MFS-a gainditzea etorkizuneko lanetarako hobekuntzak beharrezkoak izango direla argi geratzen da. Oraingoz, ordea, bide onetik doan lehen pausu bat bezala ikus daiteke, etengabeko bilakaeran dagoen alor honetan.

7.1.2 Ondorio pertsonalak

Maila pertsonalean, proiektu honen garapena oso esperientzia aberasgarria izan da, hainbat arrazoirengatik: hizkuntzaren prozesamenduaren inguruan lortutako ezagutza, ezagutza hori aplikatzea, eta tamaina honetako proiektu bat egin den lehen aldia izanik, kudeaketa.

Ezagutzaren inguruan, ez nuen inolako ezagutzarik hizkuntzaren prozesamenduaren arloan, hori dela eta, proiektuaren hasierako faseetan, oinarrizkoak diren kontzeptu nagusiak ulertzea eta barneratzeak suposatu dit lan handiena. Pertsonalki, guztiz ezezaguna den gai bat ikasten hastea eta eskala honetako proiektu bat garatzeko ezagutza eskuratzea oso esperientzia onuragarria izan da.

Eskertzekoa izan da LMMS artikuluaren egileek eskuragarri duten GitHub-eko biltegia[1]; pausoz-pauso eginiko lana nola exekutatu, ingurunea prestatu, eta artikuluan aurkezten dituzten emaitzak nola lortu azaltzen da modu oso argi eta antolatuan; kode argi eta komentatuaren bidez. Eginiko esperimentuak ulertzeko oso lagungarria izan da, eta ingurunea prestatzeko lan asko aurreztu dit. Graduan zehar lortutako Python-eko esperientzia oso baliagarria izan da, erabilitako eta eginiko kode guztia hizkuntza honetan izan baita. Modu berean Estatistika eta Datu-Meatzaritza irakasgaietako kontzeptu asko ere baliagarriak izan dira.

Eskertzekoa izan da ere zuzendarien partetik jasotako laguntza, proiektu hau gauzatzeko ezinbestekoa izan da, eta edozein zalantza edo arazo argitzeko beti prest egon dira. Pertsonalki, oso gustura aritu naiz beraiekin lanean; eta azken finean hori da tamaina honetako proiektu bat aurrera ateratzeko gauza garrantzitsuena.

[1] https://github.com/danlou/lmms

7.2 Etorkizuneko lanak

Mota honetako proiektuek ez dute bukaera finko bat izaten, beti egin daitezke gauzak modu desberdinean, eta teknologiak aurrera egin ahala aztertu daitezkeen bide berriak agertzen dira. Proiektuan, ordea, mugak ezarri behar dira, gure kasuan ingelesez entrenatutako sistema eleanitz bihurtzea, eta euskarazko datu-multzo bat finkatzea izan da.

Eginiko lana etorkizuneko ataza desberdinetarako baliagarria izan daiteke. Hori dela eta, proiektuak izan ditzakeen hainbat hobekuntza eta jarraipen bururatzen zaizkigu.

1. **Eginiko lanaren gainean froga gehiago burutu:** Adiera-embeddingak sortzeko prozesuan embedding estatikoak gehitzeko hautazko ataza bat ikusi zen (ikusi 4.1.4). Embedding estatikoak erabiliz proba batzuk egin ziren arren, ez da embedding hauekin esperimenturik egin; honen arrazoi nagusia, hizkuntza bakoitzerako bereizita daudela da. Hala ere, posible izango litzateke hizkuntza konkretu batzuk aukeratuz, embedding estatikoak biltzea, eta *Uniformed Sense Matching* atazatan aztertzea.

   Dimentsio altuko espazioetan puntu gutxi batzuk beste puntu askoren gertukoen auzokide izateko joera izaten dute; fenomeno honi *hubness* deritzo. Hubness altuak eragin kaltegarriak izaten ditu ebaluazioan. Puntuen arteko distantzia kalkulatzeko kosinu-antzekotasuna erabili beharrean, hubnessa kontuan hartzen duen CSLS izeneko neurri bat erabili daiteke. Honen inguruan esperimentu gehiago egin litezke, emaitzak hobetzen dituen edo ez aztertzeko.

   Sistema eleanitz bat sortu denez, interesgarria izango litzateke ingelesa eta euskaraz aparte beste hizkuntza batzuk ebaluatzea; hala nola, espainiera, katalana...

2. **Adiera-embeddingak sortzeko prozesu desberdinak aztertu:** LMMS artikuluan embeddingak sortzeko prozesu konkretu bat aurkezten da, ingelesezko HAD-an emaitza onenak lortzeko doitua. Honek ez du esan nahi, ordea, embedding eleanitzak sortzeko prozesu onena denik. Entrenamenduko parametro desberdinak erabiltzea, edo azken urteotan argitaratutako metodo desberdinak probatzea aukera interesgarria litzateke.

3. **Euskarazko datu-multzoan hobekuntzak:** Nahiz eta euskarazko datu-multzo bat sortu den EuSemcor-etik abiatuz, 3.1.4 atalean azaltzen den bezala EuSemcor *lexical sample* motako corpus bat da. Azken urteetan *all-words* motako datu-multzoak

erabiltzen dira gehienbat HAD atazarako. EuSemcor-en etiketatutako hitz kopurua nahiko handia denez, esaldi berdinean hitz bat baino gehiago etiketatuta egotea maiz gertatzen da, *all-words* motako bertsio bat sortzea ahalbidetuz etorkizunean. Hala ere, kontuan izanik izenak bakarrik daudela etiketatuta, ez litzateke ingelesezko datu-multzo baliokideak bezain ona izango.

Eranskinak

A. ERANSKINA

EuSemcor. Entrenamendu/garapen/test banaketako hitzak

Hitza	Agerpen Kopurua				MFS
	Train	Dev	Test	EuSemcor	(Train)
adierazpen	53	27	27	107	06722453-n
agintari	61	31	31	123	09623038-n
agiri	56	28	28	112	06470073-n
aldaketa	87	44	44	175	00191142-n
arazo	144	72	73	289	14410605-n
aste	86	43	44	173	15169873-n
bake	64	32	32	128	13970236-n
batasun	87	44	44	175	08224274-n
batzorde	90	45	45	180	08310949-n
bide	132	67	67	266	00038262-n
borroka	52	26	27	105	00788973-n
boto	53	27	27	107	00183505-n
denboraldi	88	44	45	177	15239579-n
diru	80	40	41	161	13384557-n
egoera	144	72	72	288	00024720-n
egun	228	114	115	457	15123115-n
ekintza	60	30	30	120	00952963-n

elkarrizketa	70	35	36	141	07148192-n
elkarte	54	27	27	108	08049401-n
emaitza	62	31	31	124	07292694-n
emakume	68	35	35	138	10787470-n
enpresa	92	46	46	184	08056231-n
epe	51	26	26	103	15224156-n
erakunde	103	52	52	207	08008335-n
estatu	152	77	77	306	08178547-n
etxe	152	76	77	305	08559508-n
garai	70	35	35	140	15122231-n
gau	64	33	33	130	15167027-n
gauza	110	55	56	221	05671325-n
gizon	79	40	40	159	00007846-n
gobernu	237	119	119	475	08050678-n
harreman	52	27	27	106	07134445-n
hasiera	117	59	59	235	15265518-n
haur	72	36	37	145	09917593-n
hauteskunde	72	36	37	145	00181781-n
helburu	118	60	60	238	05980875-n
herri	148	75	75	298	08672738-n
herrialde	74	37	38	149	08544813-n
herritar	83	42	42	167	09625401-n
hilabete	81	41	41	163	15206296-n
hiri	51	26	26	103	08524735-n
hitz	84	43	43	170	06286395-n
hizkuntza	69	35	35	139	06282651-n
indar	84	42	43	169	10461424-n
indarkeria	52	27	27	106	01127245-n
iritzi	57	29	29	115	05945642-n
izen	88	44	45	177	06333653-n
jarrera	79	40	40	159	06193203-n
jende	98	50	50	198	07942152-n
jokalari	147	74	74	295	10439851-n
kale	58	30	30	118	04334599-n
lagun	144	73	73	290	00007846-n

laguntza	72	36	36	144	05154908-n
lan	208	105	105	418	00575741-n
langile	54	27	28	109	09632518-n
lege	102	51	52	205	06532330-n
leku	62	31	31	124	00027167-n
maila	134	67	67	268	05093890-n
metro	59	30	30	119	13659162-n
ministro	84	42	42	168	10320863-n
minutu	70	36	36	142	15234764-n
mundu	85	43	43	171	09270894-n
ordu	88	44	45	177	15227846-n
partida	50	25	25	100	00456199-n
partidu	227	114	114	455	00456199-n
pertsona	74	37	37	148	00007846-n
politika	62	31	32	125	06656408-n
polizia	108	54	55	217	08209687-n
postu	70	36	36	142	14429382-n
proiektu	54	28	28	110	05910453-n
proposamen	60	31	31	122	07162194-n
prozesu	65	33	33	131	01023820-n
sistema	54	27	28	109	05902872-n
udal	72	36	37	145	08225736-n
une	70	35	36	141	15180528-n
ur	58	29	29	116	07935504-n
urte	484	242	242	968	15203791-n
zati	51	26	26	103	15257829-n
zigor	60	31	31	122	01189282-n

A.1 Taula: EuSemcor-en partizio bakoitzean agertzen diren hitzak, eta hauen adibide kopurua.

Hitza	**Synset WN1.6**	**Agerpen Kop.**
herrialde	50004689-n	52
iritzi	00468967-n	1

A.2 Taula: EuSemcor-en WN3.0-ra mapatu ezin izan diren adibideak.

Bibliografia

[Agency., 1993] Agency., U. S. A. R. P. (1993). *Human language technology : proceedings of a workshop held at Plainsboro, New Jersey, March 21-24, 1993*. Morgan Kaufmann Publishers, San Francisco, CA.

[Agerri et al., 2020] Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[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.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.*

[Lopez de Lacalle and Agirre, 2010] Lopez de Lacalle, O. and Agirre, E. (2010). Hitzen adiera-desanbiguazioa.

[Loureiro and Jorge, 2019] Loureiro, D. and Jorge, A. (2019). Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation.

[Otegi et al., 2020] Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442, Marseille, France. European Language Resources Association.

[Otter et al., 2018] Otter, D. W., Medina, J. R., and Kalita, J. K. (2018). A survey of the usages of deep learning in natural language processing.

[Pociello et al., 2011] Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

[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.

[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.

[Vossen, 1998] Vossen, P. (1998). *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic, Dordrecht The Netherlands Boston.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.