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

GLEN: Adiera desanbiguazioa hizkuntza-eredu sortzaileak erabiliz

Egilea

Tasio Aguirre Blanco

Zuzendariak

Ander Barrena eta Eneko Agirre

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2022ko otsaila

Sailak: Konputazio Zientziak eta Adimen Artifiziala

This is to acknowledge that we know what is
happening and what needs to be done.
Only you know if we did it.

A letter to the future

February 2022
419ppm CO_2

Laburpena

Azken urteotako ikasketa sakonean oinarritutako sistema sortzaileek ospea lortu dute jendartean gizakien idazkera naturala imitatzeko ahalmenak bultzatuta. Honez gain, hizkuntzaren prozesamenduko hainbat atazetan emaitza aipagarriak lortzeko erabili dira, arloaren egoerako bestelako sistema tradizionalak gaindituz kasu batzuetan.

Proiektu honen helburua, ingelesezko hitzen adiera-desanbiguazioa gauzatzen duen sistema bat garatzea izan da, GLEN, hitz anbiguo bati dagokion hiztegiko glosa sortzen duen BART sistema sortzailea erabiliz. Entitate-izenen desanbiguazioa burutzeko gai den sistema batetik abiatzen da bi ataza hauen antzekotasuna baliatuz. Sistemaren errendimendua hobetzeko eginiko esperimentu guztiak aurkezten dira; hala nola, ingelesez entrenatutako GLENen bertsio eleanitz bat ebaluatzen da euskarazko testuak desanbiguatzean. Ingelesezko sistemak, arloaren egoeran emaitza lehiakorrak lortzen ditu, sistema sortzaileetan oinarritutako egungo emaitza onenak eskainiz.

Abstract

Generative models based on deep learning techniques have become recognized given their ability to mimic human language. In addition, they have been used to achieve impressive results in some tasks of natural language processing, surpassing in some cases state-of-the-art results by traditional systems.

The goal of this project is to create a system to perform English word-sense disambiguation -GLEN- using BART, a generative model that generates the correct gloss of an ambiguous word. A system that performs named-entity disambiguation is used as a starting point given the similarities of these two tasks. We present all the experiments to improve the performance, as well as the evaluation of a multilingual model trained in English for disambiguation of Basque texts. GLEN achieves competitive results, being the best state-of-the-art generative WSD system.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Eguneroko bizitzan erabiltzen dugun hizkuntzak jakintzat jotzen ditugun ezaugarri asko ditu. Horietako bat hizkuntzaren anbiguotasuna da: zentzu desberdin bat baino gehiago izan dezaketen hitz edo esaldiak. Kontzeptu desberdinei erreferentzia egiteko hitz berdinak erabiltzen ditugu maiz, ohartu gabe, hitz bati **adiera** bat baino gehiago esleituz. Hiztegietan aurki daitezke adiera desberdin hauen definizioak edo **glosak**.

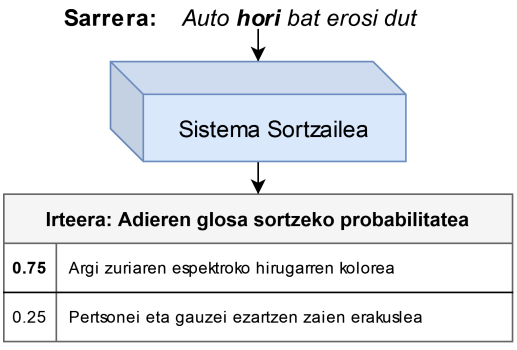
Hizkuntzaren Prozesamenduan (HP) oinarritutako ataza asko gauzatzeko beharrezkoa suertatzen da hitzen adiera egokia zein den jakitea: hizkuntza arteko itzulpen automatikoa modu zuzenean egiteko, ahotsean oinarritutako laguntzaile birtual baten erantzuna jasotzeko edo esaldi baten azterketa zuzen bat gauzatzeko. Ataza hauetarako hain baliagarria den prozesu hau Hitzen Adiera-Desanbiguazio (HAD) automatikoa deritzo. Baliagarria izaten den antzeko ataza bat Entitate-Izenen Desanbiguazioa (EID) da, non izenaipamenak (pertsona izenak, lekuak, enpresa izenak...) ezagutza-base batean bildutako entitateekin lotu behar diren. Wikipedia, edo Wikipediatik eratorritako bestelako proiektuak izanik erabiltzen diren ezagutza-base ohikoenak.

Nahiz eta hizkuntzaren prozesamenduko lehen atazetako bat izan zen HAD, azken urteetako ikasketa-sakonaren iraultzak erraztasun asko eskaini ditu tankerako ataza hauek gauzatu ahal izateko; azken urteetako argitalpenetan lortu diren emaitzetan islatuz. Ikasketa-sakonean oinarritutako metodoak modu desberdinetan erabili daitezke errendimendu handiena lortzeko: sailkatzaileetan oinarritutako sistemetatik, adieren errepresentazioetan oinarritutako sistemetara, ataza konkretu batean eginiko garapenak beste atazetarako erabilgarriak suertatzen dira. Proiektu honetan, azken urteetan emaitza aipagarriak lortzeko

erabili diren sistema sortzaileen eraginkortasuna aztertzen da HAD gauzatzeko. Entitate-izenen desanbiguazioa gauzatzeko sorturiko sistema batetik abiatuz, hitzen adiera-desanbiguazioa gauzatzeko egokituko da.

Sistema sortzaileen bereizgarri aipagarriena, gizakiek eguneroko bizitzan aurki dezaketen testu naturala sortzeko gaitasunean datza. Ezaugarri honek ikusgarritasun handia eman die ikerkuntzatik kanpo aurkitzen diren pertsonen artean, testu naturala modu automatikoan sortzeko gaitasunak izan ditzakeen ondorio kaltegarriek kezkatuta. Testu naturala sortzeko, emandako testuaren hurrengo hitz posible bakoitzari probabilitate bat esleitzen diote, testuaren koherentzia eta esanahia mantenduz. Lan honen helburua, zehazki, gaitasun honetaz baliatzea da HAD gauzatzeko. Hitz anbiguo batek izan ditzakeen adierak testu naturalean idatzitako glosa batekin lotu, eta jarraian sistema sortzaile baten bidez glosa bakoitzari probabilitate bat esleitzen zaio. Glosa probableenari dagokion adiera aukeratzen da adiera zuzen gisa. 1.1 irudian ikus daiteke lan honetan aurkezten den sistemaren ideia.

Dokumentu honetan, eginiko lanaren nondik-norakoak azaltzen dira. Lehenik, 2. kapituluan, arloaren egoera aztertzen da HAD eta EIDren definizio sakonago bat eskainiz eta erabilitako baliabideen deskribapen bat eginez. 3. kapituluan, gure sistema garatzeko oinarri bezala erabili den sistema aurkezten da, eta honen gainean eginiko aldaketak aurkezten dira. 4. kapituluan, gure sistemaren errendimendua hobetzeko eginiko garapen esperimentu eta aldaera guztiak aurkezten dira. 5. kapituluan, lortutako emaitzak aurkezten dira arloaren egoerako bestelako sistemekin alderatuz, eta euskarazko HAD egiteko ahalmena aztertzen da eredu eleanitz baten bidez. Azkenik, 6. kapituluan, proiektu honetatik atera daitezkeen ondorioak eta etorkizunean egin daitezkeen hobekuntzen gainean hausnartzen da.



1.1 Irudia: Gure sistemaren funtzionamenduaren ideia. *Hori* hitz anbiguoa testuinguruan emanik, sistema sortzaile batek adiera posibleen glosei probabilitate bat esleitzen die, probabilitate altuena duen adiera esleituz *hori* hitzari.

2. KAPITULUA

Arloaren egoera

Kapitulu honen helburua, lan honetan zehar agertzen diren kontzeptuen azalpen sakonago bat eskaintzea eta proiektua gauzatzeko erabiltzen diren tresna eta baliabideen deskribapen bat egitea da. Lehenik entitate-izenen desanbiguazioa zer den azaltzen da. Jarraian, hitzen adiera-desanbiguazioaren eta honi lotutako baliabideen azalpena aurkezten da. Ondoren, bi atazen artean aurki daitezkeen antzekotasunak azaltzen dira, izan ere, lan honen helburua entitate-izenetan oinarritutako sistema bat adiera-desanbiguaziora egokitzea da. Azkenik, hizkuntza-eredu neuronaletan oinarritutako BART sistema sortzailea aurkezten da, gure sistemaren oinarria izango dena.

2.1 Entitate-izenak

Entitateei buruz hitz egitean; izen-aipamenek erreferentzia egiten dieten pertsona-izenei, tokiei, enpresa izenei... buruz hitz egiten da. Adierekin gertatzen den antzera, aipamen eta entitateen erlazioa ez da beti zuzena izaten: izen berdina duten hainbat pertsona existitu daitezke, toki baten izen bera duen pertsona bat existitu daiteke (Paris Hilton, adibidez), eta ohikoa da izen bera duten liburu eta filmak aurkitzea. 2.1 irudian ikus daitekeen bezala, Wikipedian *London* izenak erreferentziatu dezakeen entitate zerrenda, artikuluen izenaz identifikatua, ugaria da. Jarraian, entitate-izenekin erlazionatutako bi ataza desberdin azaltzen dira, hurrengo kapituluan aurkeztuko den sistema ulertzeko.

London (disambiguation)

London is the capital city and largest metropolitan region of both England and the United Kingdom.

London may also refer to:

Places

Canada

- London, Ontario

United States

- London, Arkansas, a city

Arts, entertainment and media

Film

- *London* (1926 film), a British silent film

Literature

- *London* (Samuel Johnson poem)
- *London* (William Blake poem)

Groups and labels

- London (heavy metal band), American band
- London (punk band), British band

2.1 Irudia: Wikipedian *London* izenak errefererentziatu dezakeen entitateen zerrenda; urdinez, artikuluaren izenaz identifikatuta.

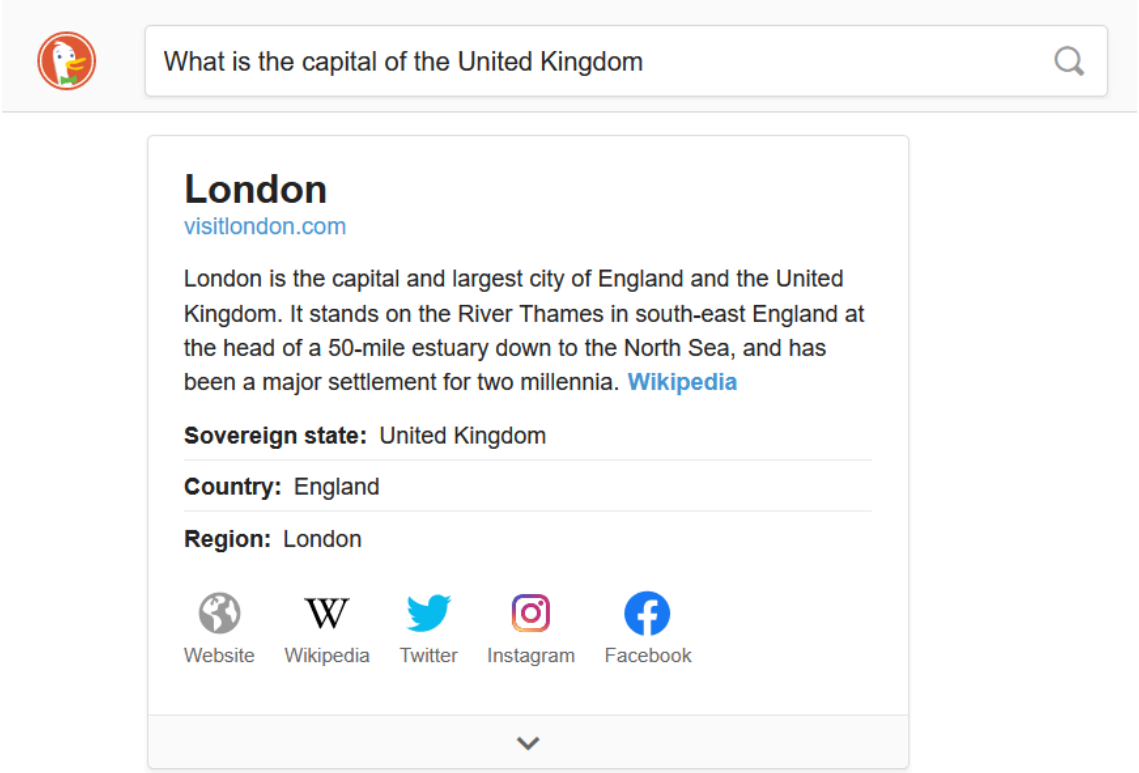
2.1.1 Entitate-Izenen Desanbiguazioa

Entitate-Izenen Desanbiguazioa (ingelesez *Named-Entity Disambiguation)* [Hoffart et al., 2011] [Lazic et al., 2015] [Barrena et al., 2018] [Cao et al., 2021b] hizkuntzaren prozesamenduko ataza bat da, non izen-aipamenak desanbiguatu behar diren dagokien entitateekin lotuz. Entitate hauek Wikipediatik eratorritako ezagutza-baseetan biltzen dira. EIDn helburua desanbiguatu beharreko izena Wikipedian dagokion artikuluaren estekarekin lotzea izan ohi da, baina bi artikuluk ezin dezaketenez izen bera izan, artikuluaren izena erabili daiteke entitatearen identifikatzaile moduan. Izen-aipamena izanda, entitate hautagai zerrenda bat izaten da atzigarri, desanbiguazioa ezagutza-baseko entitate guztien artean ez gauzatzeko. Artikulu-izenak nahiko deskribakorrak izan ohi dira erabiltzaileen erraztasunerako. 3. kapituluan ikusiko den moduan, Wikipediako artikulu-izenak entitateen identifikadore gisa erabiltzeak zentzua izan dezake sistema sortzaile baten irteera gisa, lengoaia naturalean idatzitako deskribapen laburrak izaten dituztelako, maiz parentesi artean. Adibidez, *argia*-ren entitateak Wikipedian: Argia eta Argia (aldizkaria).

2.1.2 Entitateen Berreskurapena

Entitateen Berreskurapena (ingelesez *Entity Retrieval*) [Piccinno and Ferragina, 2014] [Le and Titov, 2018] [Broscheit, 2019] [Wu et al., 2020] [Cao et al., 2021a] atazan, kontsulta bat emanik dagokion entitatea zein den asmatu behar da. Ataza konplexuagoa da, EIDn ez bezala, kontsultan ez baita izen-aipamenik agertzen. Beraz ezagutza-baseko entitate guztiak dira desanbiguaziorako hautagai. Honek mugak ezartzen ditu entitate kopuruaren arabera, bai memoria eta prozesamendu ahalmenaren aldetik.

Ataza hau oso erabilgarria da bestelako hainbat ataza gauzatu ahal izateko. Adibidez, galdera-erantzun sistemen erantzuna jasotzeko edota laguntzaile birtual edo interneteko bilatzaile bati galdera bat egiten zaionean. 2.2 irudian ikus daiteke interneteko bilatzaileak gai direla entitatea aipatu gabe, Wikipedian esaldiak erreferentzia egiten dion entitatea berreskuratzeko.



2.2 Irudia: Interneteko DuckDuckGo bilatzaileak *"Zein da Erresuma Batuko hiriburua [?]"* galderari ematen dion erantzuna. Erantzuna dagokion Wikipediako artikuluarekin estekatzen du.

2.2 Hitzen Adiera-Desanbiguazioa

Hitzen Adiera-Desanbiguazioa (ingelesez *Word-Sense Disambiguation*) [Agirre and Soroa, 2009] [Vial et al., 2019] [Bevilacqua and Navigli, 2020] [Barba et al., 2021] hizkuntzaren prozesamenduan ezarri zen lehen atazetako bat izan zen, itzulpen automatikoari lotuta 1940ko hamarkadan. Hasiera batean gizakiek bakarrik egin ahal zuten ataza bat zela pentsatu zen hizkuntzaren ezagutza izugarria kodetzeko ezinezkotasuna zela eta. 1980ko hamarkadan, ordea, aurrerapen handiak egin ziren ezagutza hori kodetzeko hiztegi elektronikoei esker. Hiztegi elektronikoak azaldu aurretik, HAD zer den azalduko da.

2.2.1 Zer da HAD?

Ataza hau zertan datzan hobeto ulertzeko har dezagun euskarazko hiztun batek aurki dezakeen adibide bat. *Hori* hitzaren glosa hiztegi batean[1] kontsultatuz gero, honako bi glosa hauek (beste batzuez gain) agertuko lirateke:

hori 1: erak. Entzuten ari denaren inguruko pertsonei eta gauzei ezartzen zaien erakuslea.

hori 2: iz. Argi zuriaren espektroko hirugarren kolorea, laranjaren eta berdearen artekoa.

Adibide gisa demagun hondoko bi esaldi hauek ditugula:

i: Horko auto *hori* erosi dut.

ii: Auto *hori* bat erosi dut.

Nahiz eta bi esaldi hauek oso antzekoak izan, ez dute inolaz esanahi berdina. Lehenengo esaldiko *hori* hitzak lehen glosari egiten dio erreferentzia, eta bigarren esaldiak bigarren glosari. Igarri daitekeen moduan, nahikoa ezagutza behar da ataza hau gauzatzeko: hizkuntzaren aldetik, erakusle eta adjektiboen arteko desberdintasuna jakiteko; hala nola, mundu errealeko ezagutza aldetik, kolore horia duten autoak existitzen direla.

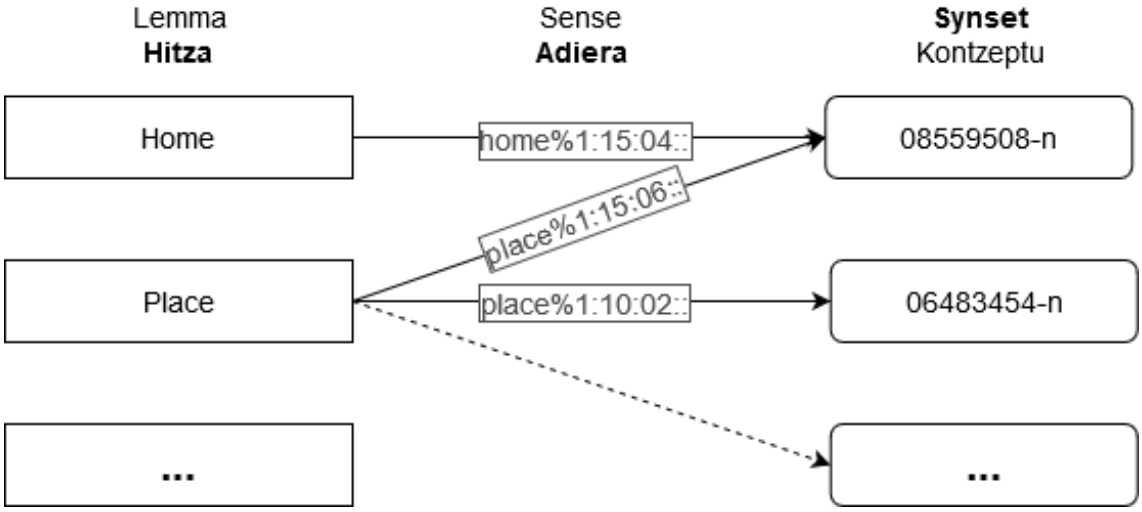
EIDn entitateei erreferentzia egiten dieten izen-aipamenak desanbiguatzen diren moduan, HADen kontzeptu bati erreferentzia egiten dieten hitzak desanbiguatzen dira. Kontzeptualki ataza berdina da, desberdintasun nagusi batekin: ez dira Wikipedia bezalako ezagutza-baseak erabiltzen; ataza hau modu automatikoan gauzatzeko gai den sistema bati ezagutza hau eramateko, WordNet bezalako hiztegi elektronikoak erabiltzen dira.

[1] http://www1.euskadi.net/harluxet/

2.2.2 Datu-base lexikalak: WordNet

WordNet [Fellbaum, 1998] da ingelesezko hizkuntzaren prozesamenduko atazetan gehien erabiltzen den datu-base lexikal edo hiztegi elektronikoa. Ingelesezko izen, aditz, adjektibo eta adberbio guztiak biltzen ditu, testuinguru desberdinetan ager daitezkeen kontzeptuekin lotuz hitz bakoitza. Kontzeptu hauetako bakoitza **synset** izeneko kode batekin identifikatzen da WordNeten.

HAD egiteko interesekoa da hitzen eta kontzeptuen arteko erlazioa. Kontzeptu bati erreferentzia egiteko hitz bat baino gehiago erabili daiteke maiz, hitzak sinonimoak diren kasuetan. Hau erraz ikus daiteke *"Deliver the package to my place"* eta *"Deliver the package to my home"* esaldietan, non *place* eta *home* hitzak trukakorrak diren esaldiaren zentzua aldatu gabe. Hitz batek dituen adierei buruz hitz egitean, hitzak eta kontzeptuak lotzen dituen erlazio honi buruz hitz egiten da, 2.3 irudian ikus daitekeen bezala. Adierak *sense_key* izeneko kode batekin identifikatzen dira, non erlazioaren ezaugarriak kodetzen diren.



2.3 Irudia: WordNeten hitzen eta *synset*en arteko erlazioaren ilustrazio bat. Hitz batek hainbat *synset* izan ditzake, eta alderantziz. Erlazio hauetako bakoitzari adiera deritzo. Iturria: [Aguirre Blanco, 2020]

Erlazio hauez aparte, erabilgarria den bestelako hainbat informazio eskaintzen da WordNeten. Adiera konkretuen agerpen kopurua corpus batean, adibide esaldiak, glosak, fitxategi lexikografikoak...

Glosak

WordNetek, hiztegi baten antzera glosak eskaintzen ditu, hau da, lengoaia naturalean idatziko azalpen laburrak. Hiztegietan ez bezala, glosak ez dira adieren arabera taldekatuta agertzen, kontzeptuen arabera baizik, *synset* bakoitzari lotuta. WordNeteko *synset* guztiek dute glosa bat lotuta. Honek esan nahi du glosek ez dutela balio adiera konkretu bat identifikatzeko, kontzeptu baino adiera gehiago baitaude. Hala ere, hitz bat desanbiguatzeko unean lema erabiliz gero ez dira bi glosa berdin inoiz topatuko, desanbiguaziorako guztiz baliagarriak izanik.

Fitxategi lexikografikoak edo *lexname*ak

WordNeteko adiera guztiak fitxategi lexikografikoen arabera taldekatuta daude. Guztira 45 fitxategi desberdin daude, ia guztiak izenei eta aditzei eskainiak, domeinu desberdinetako adierak taldekatzeko. *Sense_key* kodean adierazten da zein fitxategiri dagokion adiera.

Informazio hau erabilgarria izan daiteke adieren arteko erlazio oinarrizkoena bezala antzeko adierak identifikatzea baliagarria den kasuetan. Lan honetarako interesgarria zaigu fitxategien informazioa eskaintzen duen taulako azalpena[2], lengoaia naturalean idatzitako esaldi bat baita, adieren nolabaiteko informazio gehigarria eskainiz. Taularen lagin bat ikus daiteke 2.4 irudian.

File Number	Name	Contents
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts

2.4 Irudia: WordNeteko *lexname*en taularen lagina. Irudian izenen 5 fitxategiren identifikatzailea eta azalpen testua.

[2]Taula hemen topatu daiteke: https://wordnet.princeton.edu/documentation/lexnames5wn

Lan honetan WordNet 3.0 bertsioa erabiltzen da, gaur egun oraindik ataza hauetarako bertsio ohikoena izanik nahiz eta 3.1 bertsio berriagoa eskuragarri egon. Bertsio honetan 200.941 adiera desberdin biltzen dira, eta hizkuntza desberdinetako beste WordNetekin bateragarria da *synset* eta adiera-kode berdinak erabiliz hizkuntza gehienetan.

2.2.3 Anotatutako Corpusak: Semcor eta EuSemcor

Semcor [Agency., 1993] ingelesez eskuz semantikoki etiketatutako corpus bat da. WordNet-en egile berak egina, esparru desberdinetako esaldiak biltzen ditu, esaldi hauetan agertzen diren hitz anbiguo guztiei dagokien lema, kategoria gramatikal eta adiera eskainiz. 200.000 agerpen baino gehiago biltzen ditu WordNet 3.0 bertsioa erabiliz.

Semcorren erabilera zabalduta dago HAD sistemen entrenamendu datu bezala; datu kopuru handiak eta erabiltzeko erraztasunak oso ohikoa egin du lan gehienetan gaur arte. Hala ere, datu kopurua zabaltzeko asmoarekin, azken urteetako argitalpen anitz WordNetek eskaintzen dituen etiketatutako glosak eta adibideak[3] erabiltzen hasi dira. Entrenamendu datu estra honi **WNGE** (*WordNet Glosses and Examples*) izena ematen zaio. Alderagarritasuna mantentzen aldera, Semcor eta Semcor+WNGE erabili duten sistemak bereizten dira. Lan honetan bi entrenamendu datu hauen eragina aztertuko da garapen fasean, emaitzak hobetzeko baliagarriak diren aztertzeko.

EuSemcor euskarazko corpus etiketatua da [Pociello et al., 2011]. Semcorren antzera, hitz anbiguoak testuinguruan biltzen ditu dagokion WordNeteko adierekin lotuz. Ingelesezko Semcor ez bezala, euskarazko 407 izen ohikoenak bakarrik aukeratu ziren, hauen 40 mila agerpen baino gehiago bilduz[4]. Esaldi bakoitzean hitz anbiguo bakarra dago ingelesezko Semcorren ez bezala. Aditzak, adjektibo eta adberbioak ez biltzeaz gain, izen askok adiera bakarra dutenez etiketatuta (edo ez daudenez adiera batzuentzat nahikoa adibide) ez da oso aproposa HAD atazarako. Hala ere, euskarazko baliabide garrantzitsu bat izanik, [Aguirre Blanco, 2020] lanean corpus hau prozesatu zen Semcorren formatu berdinera doituz, testuaren kodeketa arazoak konponduz, eta jatorrizko adiera-kodeak WordNet 1.6 bertsiotik 3.0 bertsiora lotuz. Gainera, entrenamendu/garapen/test partizioak sortu ziren, HADerako garrantzitsuak diren honako ezaugarriak kontuan izanik:

- Izenak 100 agerpen edo gehiago izatea.
- Izenak gutxienez 2 adiera izatea.

[3] https://wordnetcode.princeton.edu/glosstag.shtml
[4] Estatistikak hemen: https://ixa2.si.ehu.eus/mcr/estatistikak.html

- 50, 25, 25 adibide proportzioak entrenamendu, garapen eta test partizioetan hurrenez-hurren.
- Adiera guztiak hiru partizioetan.

2.2.4 Ebaluazio datu-multzoak

Ingelesezko HADen jada ezarrita dagoen moduan, [Raganato et al., 2017]-ren ebaluazio datu-multzoak erabiltzen dira lan honetan. Artikuluan bost datu-multzo desberdin aurkezten dira, *Senseval* [Edmonds and Cotton, 2001] [Snyder and Palmer, 2004] eta *SemEval* [Pradhan et al., 2007] [Navigli et al., 2013] [Moro and Navigli, 2015] lehiaketako ataza desberdinetatik aterata. Datu-multzo hauek sistemaren errendimendua ebaluatzeko erabiltzen dira, eta beste egungo sistemekin alderatzeko aukera eskaintzen dute. Orokorrean bost datu-multzoen kateamendua erabiltzen da *ALL* izenarekin.

Esperimentu eleanitzetarako euskarazko eta gaztelerazko datu-multzo bana erabiltzen da. Euskarazko ebaluazioa egiteko EuSemcorren test partizioa erabiliko da. Datu-multzo nahiko berria izanik, ezin daiteke beste sistemekin alderatzeko erabili, hala ere, hau izanik euskarazko HAD sistema bat ebaluatzeko aukera bakarra, sistemaren errendimendu orokorra ikusteko baliagarria da. Gaztelerazko kasuan, *SemEval2013*ko ataza bateko datu-multzoa erabiltzen da. Datu-multzo hauek izenak bakarrik biltzen dituzte.

Datu-multzo guztien informazioa agertzen da 2.1 taulan.

Hitzak	**EuSemcor**	**SE13-ES**	**S2**	**S3**	**SE07**	**SE13**	**SE15**	**ALL**
Agerpenak	3743	1218	2282	1850	455	1644	1022	7253
Izenak	+	+	+	+	+	+	+	4300
Aditzak			+	+	+		+	1652
Adj.			+	+			+	955
Adb.			+	+			+	346

2.1 Taula: Euskarazko, gaztelerazko eta ingelesezko HAD ebaluazio datu-multzoen agerpenen kontaketak. + ikurraz kategoria gramatikal horretako hitzak agertzen direla adierazten da. *ALL* zutabeak ingelesezko datu-multzoen bildura adierazten du.

2.3 Hizkuntza-eredu sortzaileak: BART

Dokumentu honetan eredu sortzaileei buruz hitz egitean, hizkuntzaren prozesamenduan erabiltzen diren sekuentziatik-sekuentziarako (ingelesez *seq2seq*) ereduei buruz hitz egi-

ten da. Eredu hauek, esaldi bat emanik, beste bat sortzeko gaitasuna dute, ataza konplexuak gauzatzeko oso erabilgarriak izanik: itzulpen automatikoa, galdera-erantzunak, testu laburketa... Itzulpen automatikoan erabilgarri izateak agerian uzten du sarrera esaldi bat emanik, esaldi berriak sortzeko ahalmena. Hizkuntza naturalean idatzitako esaldiak sortzeko maiz ustiatu da ahalmen hau, noranzko bakarreko deskodetzaile bat erabiliz sententzia baten hurrengo hitz probableena zein den itzultzeko erabili daiteke eta.

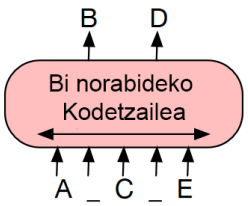
Atal honetan BART [Lewis et al., 2019] *seq2seq* eredua aurkezten da, garatuko den lanaren oinarri izango dena. Lehenik BARTen arkitektura azalduko da, honen funtzionamendua ulertzeko. Jarraian, BART nola entrenatzen den azalduko da; eta eredua birdoitzeko aukera desberdinak aipatuko dira, lan honetarako maiz egingo baita. Azkenik, eskaintzen diren BART eredu desberdinak aurkeztuko dira, hala nola, BART eleanitzak nola garatzen diren azalduko da.

2.3.1 BARTen arkitektura

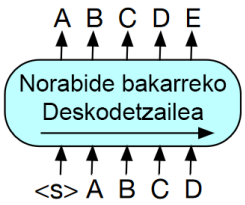
Sekuentziatik-sekuentziarako ereduak kodetzaile-deskodetzaile arkitekturan oinarritzen dira. Kodetzaileak sarrerako testua jasotzen du, tokenizatu, eta hitz-bektoreak erabiliz, esaldiak kodetzen ditu hauen esanahia irudikatu ahal izateko. Deskodetzaileak, aldiz, kodetzaileak kodetutako informazioa kontuan izanik, token baten hurrengoa iragartzen du, hurrengo token posibleei probabilitate bat esleituz; irteerako testua eraikiz. BARTen arkitektura ulertzeko BERT eta GPT azalduko dira lehendabizi, antzeko egitura erabiltzen duelako.

BERT [Devlin et al., 2019], transformerren arkitekturan [Vaswani et al., 2017] oinarritutako hizkuntza-eredua da, bi noranzkoko atentzio mekanismoetan oinarrituta. Sarrerako token batzuk ausaz ezkutatzen dira (maskaratuz, edo beste token batekin ordeztuz), eta BERTen eginkizuna, ezkutatutako token hauek zein ziren iragartzea da. Horretarako, aurreko zein hurrengo tokenen informazioa erabiltzen du (bi noranzkoak). Ondorioz, BERT ezin daiteke testu sorkuntzarako zuzenean erabili. BARTen kodetzailea BERTen egituran oinarrituta dago.

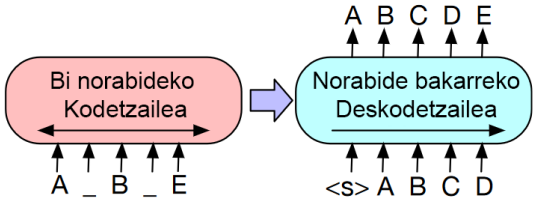
GPT [Radford and Narasimhan, 2018], norabide bakarreko eredu sortzailea da. Hasierako token bat emanik, GPTk hurrengo token posibleen probabilitatea kalkulatzen du bere ezagutza propioa erabiliz, eta token probableena itzultzen du soluzio moduan. Iragarritako token hau esaldi bukaeran kateatuz, hurrengoa kalkulatzeko erabiltzen du iterazio bakoitzean. Testu sorkuntzarako hain erabilgarria izanik, BARTen deskodetzailea egitura



(a) BERT: Tokenak ausaz maskaratzen dira eta dokumentua bi norabidetan kodetzen da. Ezkuntatutako tokenak iragartzen dira, beraz BERT ezin da sorkuntzarako erabili



(b) GPT: Tokenak ezkerretik eskuinera iragartzen dira, beraz GPT sorkuntzarako erabili daiteke. Ezkerreko informazioa bakarrik erabiliz, ezin ditu ikasi bi norabideko interakzioak.



(c) BART: Kodetzailearen sarrera eta deskodetzailearen irteera ez dira parekatuta egon behar, sarrerako testuari hautazko transformazioak onartzeko (luzera aldatu, hauen artean). Adibide honetan, dokumentu bat hondatu da testu zatiak maskaratuz. Hondatutako testua (ezkerra) bi norabideko eredu baten bidez kodetzen da, eta jatorrizko testuaren probabilitatea (eskuina) norabide bakarreko deskodetzaile baten bidez kalkulatzen da. Sistema birdoitzeko, transformaziorik gabeko testua ematen zaio kodetzaileari eta deskodetzaileari. Deskodetzailearen geruza bakoitzak kodetzailearen azken geruzako informazioa jasotzen du.

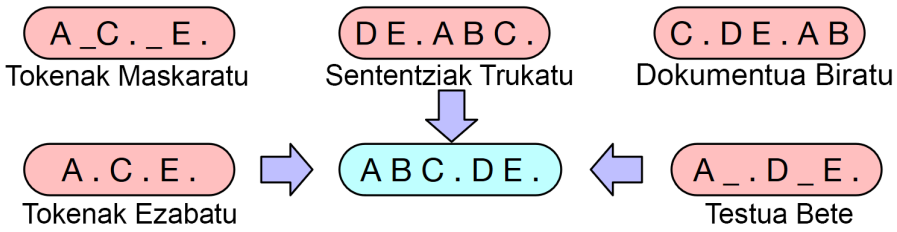
2.5 Irudia: BARTen kodetzailea eta deskodetzailearen arkitektura BERT eta GPTrekin alderatuz. Iturria: [Lewis et al., 2019].

honetan oinarritzen da. Kodetzailea eta deskodetzailea lotzeko, deskodetzailearen geruza bakoitzak kodetzailearen azken geruzako informazioa jasotzen du. Arkitekturaren ezaugarri gehiago emango dira BARTen bertsio desberdinak azaltzean, bertsio desberdinek parametro desberdinak dituzte eta.

BERT eta GPTren ezaugarriak 2.5 irudian azaltzen dira BARTen arkitekturarekin alderatuz. Kodetzaile eta deskodetzailearen lana desberdina da sistema entrenatzeko moduaren arabera. Jarraian, BART aurre-entrenatzeko, eta ataza desberdinetan birdoitzeko prozesua azalduko da.

2.3.2 Aurre-entrenamendua

BART aurre-entrenatzeko, sarrerako testuetan zarata gehitzen da eta jatorrizko testua berregiteko optimizatzen da, sistemaren irteera eta sarrerako testuaren arteko desberdintasuna ahalik eta txikiena izateko entrenatuz. Testuari zarata gehitzeak honen ulergarritasuna hondatzea du helburu. 2.6 irudian, eta jarraian, ikus daitezke testua hondatzeko erabiltzen diren transformazioak. Transformazio hauek bakarka, edo transformazio anitzen konbinaketa moduan erabili daitezke.

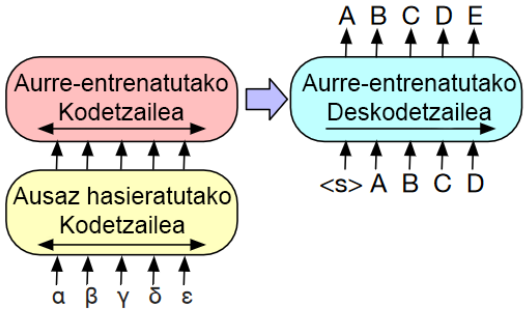


2.6 Irudia: BARTen sarreran zarata gehitzeko erabiltzen diren transformazioak. Iturria: [Lewis et al., 2019].

- **Tokenak maskaratu:** BERT azaltzean aipatu den moduan, ausaz aukeratutako token batzuk ezkutatzen dira token berezi bat erabiliz.
- **Tokenak ezabatu:** Ausaz aukeratutako token batzuk ezabatzen dira testutik. Maskaratzean ez bezala ez da token hauen existentziaren arrastorik uzten.
- **Testua bete:** Elkarren jarraian aurkitzen diren token sekuentziak maskaratzen dira token berezi bakarra erabiliz. Sekuentziaren luzera probabilitate-banaketa bat erabiliz kalkulatzen da, eta zero denean, maskaratzeko erabiltzen den tokena txertatzen da testuan ezer ezabatu gabe. Ataza honek, jatorrizko testuan falta diren token kopurua identifikatzen erakusten dio ereduari.
- **Sententziak trukatu:** Testuan aurkitzen diren puntuz banatutako esaldiak ausazko ordenan nahasten dira.
- **Dokumentua biratu:** Dokumentuko token bat ausaz aukeratzen da, eta dokumentuaren lehen token bezala zehazten da. Ataza honek dokumentuaren hasiera zein den identifikatzen irakasten dio ereduari.

2.3.3 Birdoitze edo *fine-tuninga*

Behin BART eredu bat aurre-entrenatuta izanik, ataza konkretuak gauzatzeko birdoitu daiteke. BARTen errendimendua hainbat ataza desberdin gauzatzeko aztertu da, modu desberdinetan birdoituz esaldiak sortzeko, tokenak edo esaldiak sailkatzeko, eta lan honetan interesgarriena, itzulpen automatikoa gauzatzeko. Itzulpen automatikoan birdoitzeko, BART (kodetzaile eta deskodetzailea, biak) aurre-entrenatutako deskodetzaile bakar baten moduan erabiltzen dira. Ingelesez aurre-entrenatutako BART eredua ez da gai beste hizkuntza bateko testua ulertzeko: tokenizazioa desberdina izan daiteke, testu-karaktere desberdinak erabili... Horretarako, sarreran kodetzaile txiki bat gehitzen da, eta honen



Itzulpen automatikorako, BARTen hitz-bektoreak ordezkatzen dituen kodetzaile gehigarri txiki bat ikasten da. Kodetzaile berriak lexiko disjuntu bat erabili dezake.

2.7 Irudia: BARTen birdoikuntza itzulpen automatikoa egiteko. Iturria: [Lewis et al., 2019].

irteera, BARTen hitz-bektoreen sarrerara lotzen da. Kodetzaile berri honen helburua, hizkuntza desberdin bateko testua BARTek uler dezakeen zerbaitetan bihurtzea da. Jarraian, BARTek ingelesezko testua berregingo du atzerriko hizkuntza bateko testua ingelesera itzuliz. Egitura hau 2.7 irudian ikus daiteke.

Lan honetan BART zabalki erabiltzen da hitz anbiguoen (adieren) glosak sortzeko, eta sortutako glosa bakoitzari probabilitate bat esleitzeko. Ataza honetarako birdoitzeko, itzulpen automatikoa egiteko erabiltzen den egituraz baliatzen da.

2.3.4 BART eredu desberdinak

BART eredu aurre-entrenatu desberdinak publikoki eskuragarri aurkitu daitezke; hala nola, ataza konkretuak gauzatzeko jada birdoitutako ereduak ere. Eredu aurre-entrenatuei dagokionez, erabiltzailearen esku uzten da ataza konkretuetarako birdoitze prozesua nola egin. Hauen bi bertsio argitaratzen dira:

- **BART-Base:** 6 kodetzaile eta deskodetzaile transformer geruza, 768 dimentsioko hitzen barne errepresentazioak. Eredu honen parametro kopurua guztira 140M da.
- **BART-Large:** 12 kodetzaile eta deskodetzaile transformer geruza, 1024 dimentsioko hitzen barne errepresentazioak. Ereduaren parametro kopurua guztira 400M da.

Tamaina handiagoko bertsioak emaitza hobeak eskaintzen ditu, konputazio ahalmen handiagoa eskatuz. Lan honetan BART-Base erabiliko da garapen esperimentuak gauzatzeko, eta BART-Large azken emaitzak lortzeko. Modu honetan garapen esperimentuak arindu daitezke konputazio baliabide gutxiago erabiliz.

BART eleanitzak

mBART [Liu et al., 2020] BARTen bertsio eleanitza da. Ingelesaz gain, aldi berean beste hainbat hizkuntzatan aurre-entrenatutako eredua da. Lehenik, hainbat hizkuntzetako testu elebakarra erabiliz aurre-entrenatzen da, 2.3.2 atalean ikusi den moduan. Honela, hizkuntza desberdinetako testua kodetu eta ulertzeko gai den sistema bat lortzen da. Ondoren, eredu hau birdoitu daiteke nahi den hizkuntza-parea erabiliz itzulpen automatikoa gauzatzeko, 2.3.3 atalean ikusi den moduan, jatorrizko hizkuntzako testua kodetzaileari, eta helburu testua deskodetzaileari emanik. Itzulpen automatikoaz gain, beste edozein ataza gauzatzeko birdoitu daiteke eleaniztasun honetaz baliatuz.

Aurre-entrenatzeko, Common Crawl (CC) [Wenzek et al., 2020] [Conneau et al., 2020] baliabidea erabiltzen da, hortik lortutako 25 hizkuntza desberdinetako testua erabiliz. Testu hau internetik erauzia izan da, eta modu librean banatzen da. mBART ofizialaren 25 hizkuntzen artean ez da euskara gehitzen, baina ikerketa talde desberdinek mBART sortzeko prozesu berdina jarraitu dute CCko hizkuntza desberdinak erabiliz. Honela aurki daitezke *mBART.cc25* (25 hizkuntza), *mBART.cc50* (50 hizkuntza), eta gure lanerako erabilgarria den *mBART.cc100* (100+ hizkuntza).

mBART.cc100 [Cao et al., 2021b] 100 hizkuntza baino gehiago[5] erabiliz aurre-entrenatutako BART sistema eleanitza da. Hizkuntza hauetan aurre-entrenatu ondoren, hizkuntza kopurua 125-era igotzen da Entitate Berreskurapenerako birdoituz. Birdoiketa hau ez da gure lanerako baliagarria, baina euskarazko eskuragarri dagoen mBART bakarra izanik, esperimentu eleanitzak gauzatzeko erabiliko da HAD atazarako birdoitu ondoren. Eredu hau BART-Large ereduan oinarritzen da, geruza eta dimentsio kopuru berdina izanik.

[5] Hizkuntza zerrenda hemen: https://data.statmt.org/cc-100/

3. KAPITULUA

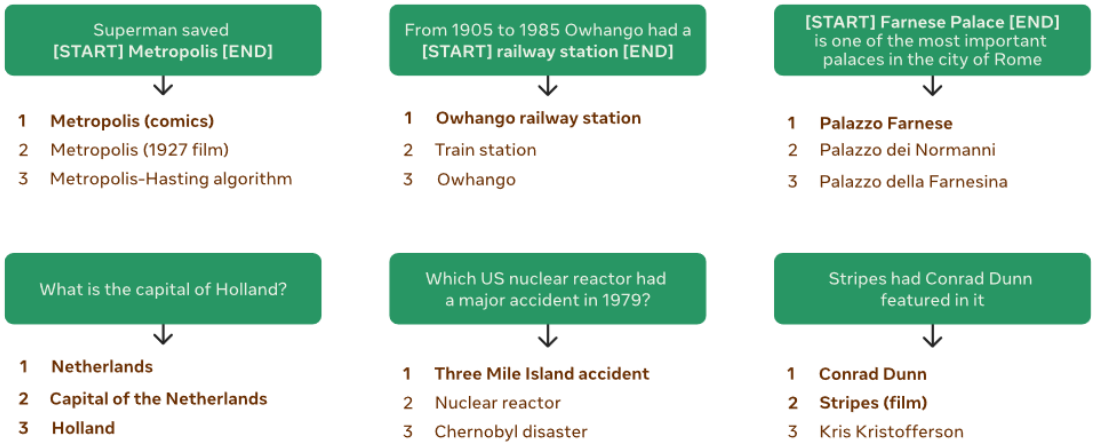
Gure sistema: GLEN

Kapitulu honetan, garatutako sistema eta honen ezaugarriak azaltzen dira: GLEN (*Gloss & Lexname gENerator*). Lehenik, sistema garatzeko oinarri bezala erabili den GENRE sistema aurkezten da, eta gure azken helburua betetzen duen antzeko sistema bat aurkezten da, Generationary. Azkenik, GENRE HAD ataza gauzatzeko sistemaren egokitze-prozesua azaltzen da.

3.1 Entitateen Berreskurapenerako sistema: GENRE

GENRE (*Generative ENtity REtrieval*) [Cao et al., 2021a] da lan hau garatzeko oinarri bezala erabili den sistema. 2021eko martxoan Facebook ikerkuntzako talde batek argitaratuta, entitateen berreskurapena (ikusi 2.1.2 atala) gauzatzeko sistema sortzaile bat erabiltzen du egungo sistemen errendimendua gainditu, eta entitateen berreskurapena ikuspegi desberdin batetik ebazteko.

Sailkatzaileetan oinarritutako sistema tradizionalek gabezia aipagarriak dituzte: konputazio kostu handia du milioika entitateen artean sailkatzaile bat erabiltzeak, eta entitate guztientzako errepresentazioak gordetzeak memoria kostu oso handia du entitate kopuru handia erabiltzen den errealitateko kasuetan. Adibide bat jartzearren, Wikipediako gutxi gorabeherako 6 Milioi entitateen errepresentazioak gordetzeko 24GB inguruko memoria beharrezkoa da. Sistema sortzaile bat erabiltzeak arazo hauek konpon ditzake, ataza ebazteko beste ikuspegi bat eskainiz.



3.1 Irudia: GENRE sistemak jasotzen duen sarrera eta itzultzen duen irteeraren adibideak. Goiko adibideek entitate-izenen desanbiguazioa, eta behekoek entitate berreskurapena erakusten dute. Iturria: [Cao et al., 2021a].

GENREk, BART-Large eredu bat birdoituz arloaren egoera gainditzen du ia ataza guztietan baliabide askoz gutxiago erabiliz, RAM eta disko-memoriari dagokionez. Adibidez, Wikipediako artikulu-izen guztien informazioak 600MB inguru hartzen ditu, lehen aipatutako sistema tradizional batek baina 40 aldiz gutxiago izanik.

Entitateen berreskurapena nola gauzatzen den 3.1 irudian ikus daiteke. Goiko hiru adibideek EIDren adibideak erakusten dituzte, izen-aipamena testuinguruan emanik sistemak iragarritako hiru entitate probableenak aurkezten dira. Entitateen berreskurapenaren kasuan, izen-aipamena agertzen ez den esaldi batek erreferentziatzen duen entitatea identifikatu behar da (*Zein da Holandako hiriburua?*). Zoritxarrez ez dago datu-multzo nahikoa handirik ataza hau irakatsi ahal izateko. Hori dela eta, sistema ataza honen barne-atazak gauzatzeko garatzen da, aurre entrenatutako BART eredu bat barne ataza hauek egiteko birdoituz. Ataza hauekin, entitateen berreskurapena egiteko gai den sistema egoki bat lortzen da. GENREk erabiltzen dituen barne atazen artean, honako hauek dira interesgarrienak:

1. **Entitate-Izenen Desanbiguazioa**: Publikoki eskuragarri dagoen aurre-entrenatutako BART eredu batetik abiatuz, Wikipediako 9 milioi izen-aipamen biltzen dituen datu-multzo bat erabiltzen da eredua birdoitzeko. Ataza honetan, izen-aipamenak esaldi batean markatuta ageri dira bi token berezien artean, eta dagokien entitatearen artikulu-izena eskaintzen da aipamenaren soluzio gisa sistema entrenatu ahal izateko.

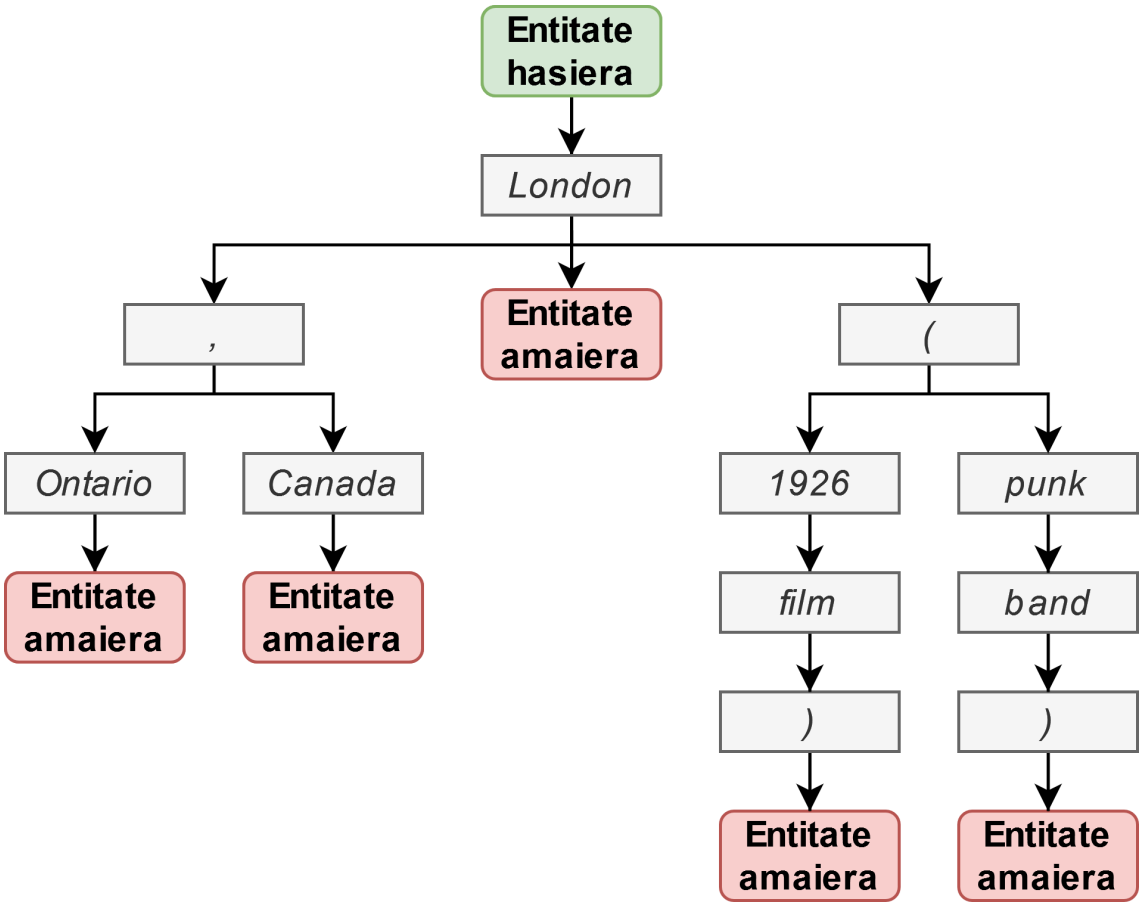
2. **Muturretik-Muturrerako Entitateen Ezagutze eta Desanbiguazioa:** Ataza honetan Wikipediatik eratorritako testua eskaintzen zaio ereduari entrenamendu moduan. Ereduak esaldi baten agertzen diren izen-aipamenak identifikatu eta dagokien entitate-izena txertatu behar du testuan. Aurreko atazan bezala, desanbiguazio osagai garrantzitsu bat dago entitateak identifikatu ostean. Hau hobeto ulertzeko adibide hau eskaintzen da: esaldi hau emanik: *"In 1503, Leonardo began painting the Mona Lisa."*, sistemak honako esaldia itzuli beharko luke: *"In 1503, [**Leonardo**](Leonardo da Vinci) began painting the [**Mona Lisa**](Mona Lisa)"*

3. **Dokumentu Berreskurapena:** Berreskurapenarekin lotutako bestelako atazak gauzatzeko birdoitzen da sistema: galdera-erantzunak, entitate lotura, elkarrizketa sorkuntza...

BART bezalako sistema sortzaileen funtzionamendua, testu bat emanik hurrengo hitz posibleen probabilitatea itzultzean datza (hau da, esaldiak hitzez-hitz sortzen dituzte). Hori dela eta, ezin daiteke esaldi osoa sortu arte honen probabilitatea zein izango den jakin. Izen-aipamen bat desanbiguatzeko honen Wikipediako artikulu-izena sortu behar denez, entitate posible guztien izena sortu beharko litzateke gertagarriena zein den jakiteko. Konputazio kostu hau ekidin daiteke *Beam Search* bezalako algoritmo bat erabiliz, dagoeneko sortu den testuaren probabilitatea kalkulatuz hitz bat gehitzen den bakoitzean, entitate asko baztertzeko aukera eskainiz. Ez du ziurtatzen probabilitate handien duen emaitza sortzea, baina bai denbora askoz laburragoan emaitza optimoa edo optimora asko hurbiltzen den soluzio bat ematea. Honez gain, bilaketa-zuhaitzak erabili daitezke bilaketa oraindik eta gehiago murrizteko, hautagai posibleak aurreprozesatuz.

3.1.1 Bilaketa-zuhaitzak

Entitatearen bilaketa errazteko, garrantzitsua da bilaketa existitzen diren entitateetara bakarrik mugatzea, denborarik ez galtzeko existitzen ez diren artikulu-izenak sortzen. Exekuzioan mugaketa hau erraz egiteko, existitzen diren artikulu-izenen zuhaitzak aurreprozesatu daitezke, hitz bat izanik hurrengoak zein izan daitezkeen jakiteko. Hau *trie* izeneko bilaketa zuhaitzen bidez lortzen da, 3.2 irudian ikus daitekeen bezala.



3.2 Irudia: Bilaketa-zuhaitz baten adibidea *London*-en entitate desberdinak sortzeko.

Lan honen helburua sistema honetatik abiatzea da HAD gauzatzeko, beharrezkoak diren aldaketak eginez; entitateen identifikatzaileak sortu beharrean adieren glosak sortu, eta ondoren, glosa probableenari dagokion adiera esleitzeko. Hurrengo ataletan, eginiko aldaketak azalduko dira, eta lortutako emaitzak arloaren egoerako beste sistemekin alderatuko dira. Hori egin aurretik, ordea, beharrezkoa da HAD egiteko sistema sortzaileak erabili dituzten argitalpenak kontuan izatea; konkretuki Generationary sistema, gure lanaren oso antzeko ikuspuntu batez baliatzen baita.

3.2 Glosa-Sorkuntza sistema: Generationary

Generationary [Bevilacqua et al., 2020], 2020ko azaroan argitaratutako sistema da, BART erabiliz HAD egiteko. Nahiz eta gure lanaren helburu antzekoa izan eta baliabide berdinak erabili, Generationaryren ideia desberdina da. Sistema glosa-sorkuntza (*Definition Modeling*) ataza gauzatzeko diseinatuta dago, non hitz bat testuinguruan emanez, kasu

horretarako onargarria den glosa sortu behar den. Hau da, sistema sortzailearen irteera ez da mugatzen jadanik gizakiek sortutako hiztegi bateko glosa posibleetara, baizik eta glosa berriak sortzea du helburu. Kontzeptu batentzat glosa egokiak sortzea ez da lan erraza, hitz gutxitan kontzeptu bat ulertzeko erraza den esaldi aproposa aurkitu behar delako.

Glosa-sorkuntzako sistema baten errendimendua ebaluatzea ez da ataza erraza izaten: ebaluazio datu-multzo irekiak oso urriak dira erabiltzen diren glosen egile-eskubideak direla eta, eta sortutako glosa baten kalitatea modu automatikoan neurtzea ez delako ataza erraza. Hiztegiko glosetara antzekotasuna kalkulatu daitekeen arren itzulpen automatikoko sistemak ebaluatzeko erabiltzen diren neurriak erabiliz (BLEU [Papineni et al., 2002], Rouge [Lin, 2004], METEOR [Banerjee and Lavie, 2005]...), ez dira neurri guztiz fidagarriak glosa baten ulergarritasuna ebaluatzeko, nahiz eta orokorrean oso neurri baliagarriak izan edozein testuren kalitatea neurtzeko. Hori dela eta, Generationary garatutako taldeak gizakiak erabiliz neurtzen du kalitatea zenbakizko eskala bat erabiliz adibide bakoitzean. Irtenbide hau egokia da sistemaren errendimendua modu isolatuan ebaluatzeko, baina ez beste sistemekin alderatzeko. GENRErekin gertatzen den antzera, sistema ataza ezagunagoetan ebaluatzeko doitzea izan ohi da irtenbide errazena.

Generationaryren sortzaileek, glosen sorkuntza adiera desanbiguaziora egokitzea proposatzen dute garatutako sistemaren errendimendua ebaluatu, eta beste sistemekin alderatzeko. Hau gauzatzeko sortutako glosak hiztegi batekoekin alderatzen ditu, eta antzekotasunaren arabera adiera bat aukeratzen du. Antzekotasuna kalkulatzeko Sentence-BERT (SBERT) [Reimers and Gurevych, 2019] hizkuntza-eredua erabiltzen da, sortutako glosak WordNeten aurki daitezkeenekin alderatuz. Glosak bektore-espazio berdinean daudenez, kosinu-antzekotasunaren formula erabiltzen da antzekotasuna neurtzeko, eta antzekotasun handiena duen WordNeteko glosa aukeratzen da adiera zuzen bezala. Nahiz eta arloaren egoerako beste sistemen emaitzak ez gainditu HADen, emaitza aipagarriak lortu zituen uneko sistemetara nahikoa hurbilduz.

GENRE eta Generationary ikusi ondoren, argi dago sistema sortzaileen gaitasuna ustiatu daitekeela HAD gauzatzeko. Generationaryren kodea ez dagoenez publikoki atzigarri (eskaera bidez bakarrik lor daiteke), gure lanaren lan-lerroa GENRE sistema egokitzea izango da ataza hau gauzatzeko eta ahal den neurrian Generationaryk lortutako emaitzak hobetzeko.

3.3 GENRE HAD egiteko egokitu

Aurreko ataletan sistema sortzaileen funtzio-aniztasuna agerian geratu da, hala nola, ataza batetik bestera sistema bat egokitzeko aukera maiz erabiltzen dela ikusi da hizkuntzaren prozesamenduan. Lan honen helburua GENRE HAD egiteko egokitzea da, Generationaryk eskaintzen duen antzeko lan bat eginez, baina BARTen irteera WordNeteko glosetara mugatuz bilaketa zuhaitzak erabiliz. Sistema honi GLEN deitu zaio, *Gloss*, *Lexname* eta *gENerator* hitzen nahasketa bezala. Hurrengo ataletan GLENen oinarri-lerroa definituko da, sarrera eta irteerako datuen formatua definituz HAD egin ahal izateko.

3.3.1 Irteeraren formatua: Glosak

Hitz anbiguo batek dituen adierak identifikatzeko, glosak erabiltzea proposatzen da, aldaketa pare bat eginez:

- WordNeten glosak kontzeptuetara daude lotuta. Arrazoi hau dela eta glosak erabiliz ezin daiteke adiera konkretu bat identifikatu guztien artean. Hau ez da arazo bat HAD egitean adiera-hautagaiak kopurua lema bidez filtratzen direnean, baina bai adiera guztien artean desanbiguatzen denean. Glosak desberdintzeko, dagokion adieraren lema gehitzen da glosaren aurretik.
- BARTen lengoaia naturala idazteko ahalmenaz baliatzeko, ez da glosa zuzenenean sortuko baizik eta lengoaia naturalean idatzitako esaldi batean gehituko da.

3.3 irudian ikus daitezke proposamen hauek. Formatu honetako hiztegi bat sortuko da, adiera-kode guztiak dagokien testu identifikatzailearekin lotuz.

home%1:15:04::	The definition of **home** is: where you live at a particular time
place%1:15:06::	The definition of **place** is: where you live at a particular time
place%1:10:02::	The definition of **place** is: an item on a list or in a secuence

3.3 Irudia: GLENek adieren identifikatzaile gisa erabiltzen duen testuaren formatua. Horiz markatutako adierek kontzeptu berari egiten diote erreferentzia, eta beraz, testuan txertatutako glosak berdinak dira. Lema erabiltzeak testu bera duten bi adiera ez egotea ziurtatzen du.

Sarreran laburki azaldu den moduan, sistema sortzaileek testua sortzen dute sarreran jasotako testuaren hurrengo hitz posibleei probabilitate bat esleituz. Bilaketa zuhaitzen bidez, glosa guztiak sortzeko hitzak erraz topatu daitezke. Hitz anbiguo baten l lema eta bere t testuingurua izanik, g glosaren probabilitatea estimatzen da glosa osatzen duten h hitzen probabilitateen biderkadura kalkulatuz.

$$p(g|t,l) = \prod_{n=1}^{N} p(h_n)$$

Testuingurua eta adiera-hautagaien glosak sortzen dituen bilaketa zuhaitza emanik BARTi, glosa bakoitzari esleitzen dion probabilitatea itzuliko du, probabilitatearen arabera ordenatuta. Zerrenda honetako lehenengo glosari dagokion adiera aukeratuko da, eta hau izango da egingo den iragarpena.

3.3.2 Sarreraren formatua: KILT datu-multzoak

GENRE entrenatu eta ebaluatzeko datu-multzoak KILT [Petroni et al., 2021] atazen formatu berdinean ematen dira. 2.2.3 atalean ikusi den bezala, Semcor eta ebaluazio datu-multzoak XML formatuan datoz, non esaldi bateko hitz anbiguo guztiak etiketatuta agertzen diren. 3.4 eta 3.5 irudian ikus daiteke esaldi berdina nola kodetzen den bi formatuetan.

Erabiliko diren datu-multzo guztiak KILT formatura egokituko dira. Esaldi bakoitzean hitz anbiguo bakarra etiketatu daitekeenez, Semcor eta ebaluazio datu-multzoetako esaldiak hitz anbiguo adina aldiz errepikatuko dira, bakoitzean hitz desberdin bat etiketatuz.

```
<sentence id="senseval2.d000.s000">
<wf lemma="the" pos="DET">The</wf>
<instance id="senseval2.d000.s000.t000" lemma="art" pos="NOUN">art</instance>
<wf lemma="of" pos="ADP">of</wf>
<instance id="senseval2.d000.s000.t001" lemma="change" pos="NOUN">change</instance>
<wf lemma="." pos=".">.</wf>
</sentence>

senseval2.d000.s000.t000 art%1:09:00::
senseval2.d000.s000.t001 change%1:04:00::
```

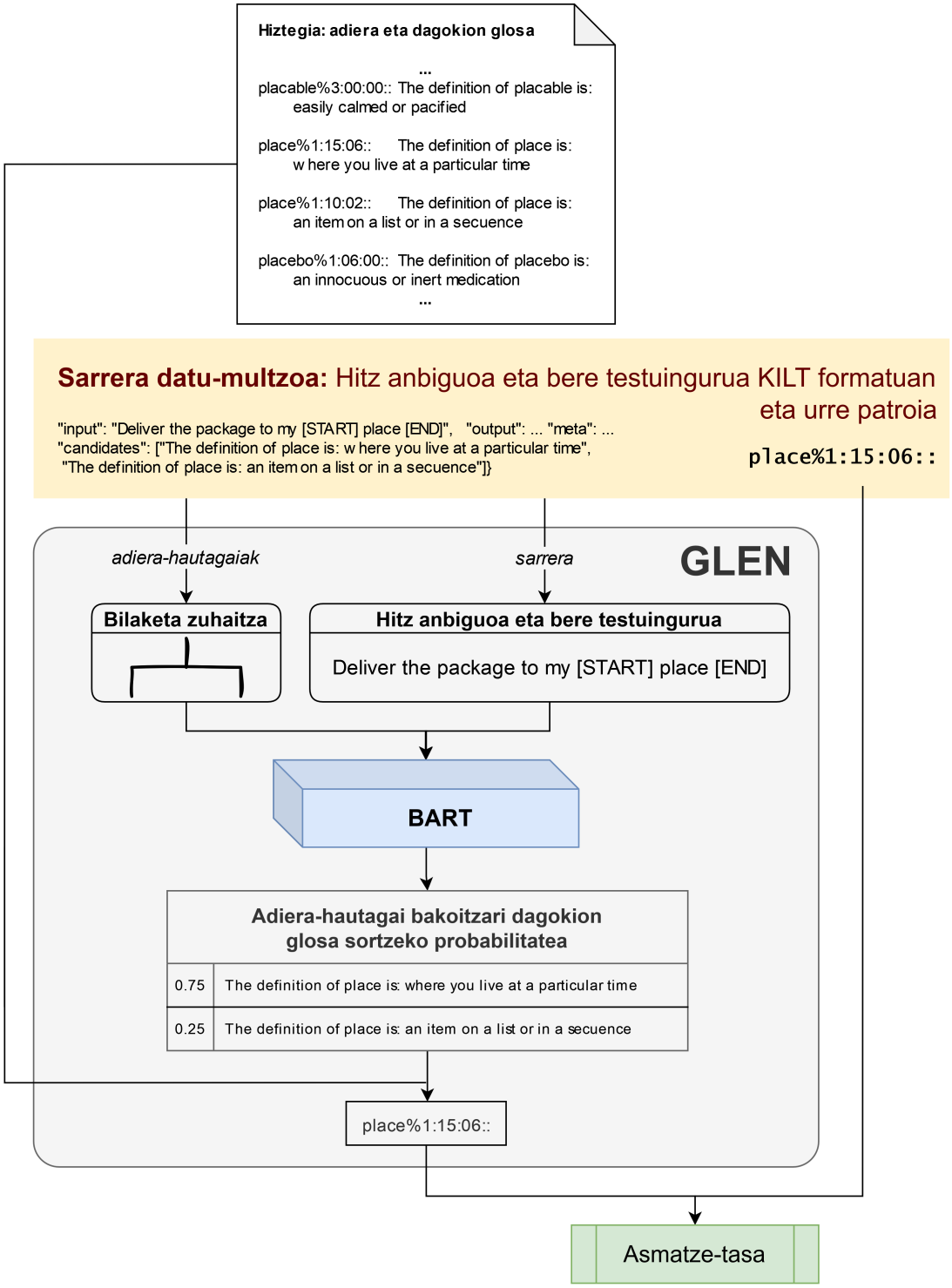
3.4 Irudia: Semcorrek erabiltzen duen formatuaren adibidea. "The art of change″ esaldia, non bi hitz anbiguoen adiera kodea agertzen den.

```
{"id": 0 , "input": "The [START_ENT] art [END_ENT] of change",
"output": [{"answer": "The definition of art is: a superior skill...
"meta": {"left_context": "The", "right_context": "of change", "mention": "art"},
"candidates": [ "The definition of art is: the creation of beautiful or...",
                "The definition of art is: the products of human creativity"]}

{"id": 1 , "input": "The art of [START_ENT] change [END_ENT]",
"output": [{"answer": "The definition of change is...
"meta": {"left_context": "The art of", "right_context": "", "mention": "change"},
"candidates": [..., ..., ..., ...]}
```

3.5 Irudia: KILT formatuaren adibidea. "The art of change″ esaldia, non bi hitz anbiguo etiketatzen dira esaldia errepikatuz.

Behin sarrera eta irteerako datuen formatua definituta 3.6 irudian ikus daiteke zein den GLENen funtzionamendua. Sarrerako datu-multzoak KILT formatura bihurtzen dira eta adiera-hautagai posibleak kalkulatzen dira 3.3 irudiko irudian azaltzen den hiztegiaz baliatuz. Jarraian, adiera-hautagai hauekin bilaketa zuhaitz bat kalkulatzen da. BARTek bilaketa zuhaitza eta KILT formatuan esaldi bat emanik, glosa posibleen probabilitatea itzultzen du. Probabilitate handiena duen irteera dagokion WordNeteko adiera-kodearekin lotzen da, eta soluzio bezala eskaintzen da. Soluzio hau erabili daiteke asmatze-tasa kalkulatzeko, datu-multzoaren urre patroiarekin alderatuz. Hurrengo kapituluan oinarri-lerro hau erabiliko da hobekuntza posibleak aztertzeko.



3.6 Irudia: GLEN sistemaren datuen erabileraren adibidea. Sarrera moduan glosa hiztegi bat, eta datu-multzoko adibide bat KILT formatuan ematen dira. Adibide bakoitzaren adiera-hautagaientzat bilaketa zuhaitz bat kalkulatzen da. BART ereduak adiera-hautagai bakoitzari dagokion glosa sortzeko probabilitate bat esleitzen dio, eta probabilitate altuena duen glosa aukeratzen da. Hiztegia erabiliz dagokion adierarekin lotzen da. Iragarpen honekin, erabiltzaileak asmatze-tasa kalkulatu dezake urre patroia erabiliz.

4. KAPITULUA

Garapen faseko emaitzak

Orain arteko kapituluetan, GLENen funtzionamendua ulertzeko beharrezkoa den informazioa aurkeztu da. Entrenamendu eta ebaluazio datu-multzoen formatu aldaketa egin ondoren, sistema birdoituko da HAD atazarako eta lehen emaitzak lortuko dira ingelesezko ebaluazio datu-multzoetan. Emaitza hauek lortu aurretik, garapen esperimentu batzuk egingo dira, emaitzak alderatzeko entrenamenduko datuen lagin txiki bat erabiliz. Semcorreko bi mila eta bat adibidez osatuta, garapen datu-multzo honi **Dev2001** izena jarri zaio, eta garapen esperimentuen arteko hobekuntzak identifikatzeko lagungarria izango da. Jarraian, eginiko esperimentu guztiak aurkeztuko dira modu ordenatuan. Kapitulu honen amaieran Dev2001 datu-multzoaren gainean lortutako emaitza guztiak aurkeztuko dira taula batean, eta emaitza onenak eskaintzen dituen eredua erabiliko da ebaluazio datu-multzoen gainean GLEN ebaluatzeko.

4.1 Diseinu esperimentala

Lehen esperimentuak egiteko, GLENen oinarri den ingelesezko BART hizkuntza-ereduaren *base* bertsioa erabiliko da. Entrenamenduan Semcor erabiliz, datu-multzoaren % 80a erabiltzen da entrenamendurako eta % 20 garapenerako[1]. Honez gain, Dev2001 datu-multzoa erabiltzen da eginiko esperimentuen emaitzak lortzeko.

[1] Birdoitzean eredu onena aukeratzeko.

Ebaluatzeko neurriak

Antzeko lanetan ohikoa den bezala, F-neurria (*F1*) erabiltzen da GLENen asmatze-tasa kalkulatzeko. Neurri hau doitasunaren eta estalduraren batezbesteko harmonikoa kalkulatuz lortzen da. Doitasunak sistemak iragarri ditzakeen artean asmatutako elementuen proportzioa adierazten du.

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

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

$$estaldura = \frac{asmatutakoak}{urre\ patroia}$$

F1 neurria, beraz, honela kalkulatzen da.

$$F1 = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura}$$

4.2 Entrenamendu datu kopurua handitu: WNGE

2.2.3 atalean aipatu den bezala, Semcor corpusaz gain WordNeteko eskuz etiketatutako glosak eta adibideak entrenamendu moduan erabiltzea ohiko praktika bihurtu da azken urteetan garatutako sistemen errendimendua modu erraz batean hobetzeko. Entrenamenduan erabiltzen diren adibide kopurua 226.037-tik 840.474-ra handitzen da, entrenamendu kopuru handia izateaz baliatzen diren sistemetan oso erabilgarria izanik.

4.3 Glosak sortzen: GLEN(Gloss)

GLEN sistemaren oinarri-lerro moduan 3.3 atalean definitutako sistema erabiltzen da. Honen azalpena 3.6 irudian ere aurkezten da irudi baten bidez. Oinarri-lerroa honako formula erabiliz definitu daiteke. Notazio matematikoaz gain, **GLEN(Gloss)** izenaz ere identifikatuko da testuan erreferentziatzean ulergarriagoa izan dadin.

$$Glen(Gloss) = p(g|t,l) = \prod_{n=1}^{N} p(h_n)$$

GLEN(Gloss)	The definition of **home** is: where you live at a particular time
GLEN(Lex+Gloss)	**This is a location** with definition: where you live at a particular time

4.1 Taula: Gloss eta Lex+Gloss arteko desberdintasuna. GLEN(Lex+Gloss) ereduak hizkuntza naturalean txertatutako esaldi desberdina du, *lexname*en informazioa erabiliz lemaren ordez (beltzez markatuta kasu bakoitzean).

4.4 Glosak eta *lexname*ak sortzen: GLEN(Lex+Gloss)

WordNet azaltzean *lexname*ak zer diren ikusi da 2.2.2 atalean. Hauek, adierei buruzko informazio gehigarria ematen dute, eta informazio hau lengoaia naturalean kodetu daitekeenez abantaila aproposa eskaintzen dute lan honetan garatzen den sisteman erabiltzeko. 45 *lexname* bakarrik daudenez, ez dute nahikoa informazio eskaintzen hitz baten adiera guztien artean bat identifikatu ahal izateko, baina glosetara gehitu daitezke informazio gehigarri gisa. Eranskineko A.1 taulan daude erabiltzen diren taldekatze-izenak. Izenak eta aditzak desanbiguatzeko dira erabilgarriak gehienbat, adjektibo eta adberbioen kasuetan ez dutelako informazio gehiegi eskaintzen. Kontuan izan behar da ere, batzuetan hitz baten adiera asko *lexname* berdinekoak direla; kasu hauetan desanbiguaziorako eskaintzen duten informazioa mugatua izango da.

Hitz bat desanbiguatzeko lema horrek izan ditzakeen adieren artean aukeratzea izaten da ohikoena HADen (lema bidezko filtrazioa). Hori dela eta *lexname*ekin eginiko lehen esperimentuan, oinarri-lerroan adieren identifikatzaile gisa erabiltzen den testuan *lexname*a gehituko da lemaren partez. 4.1 taulan irudikatzen da aldaketa hau nola egin den. Esperimentu hau hurrengo formula erabiliz definitu daiteke; *lex lexname*aren probabilitatea izanik:

$$GLEN(Lex+Gloss) = p(lex, g|t, l) = \prod_{n=1}^{N} p(h_n)$$

4.5 Bi eredu desberdin: GLEN(Lex) eta GLEN(Gloss)

Eredu bakarrean *lexname*a eta glosa iragarri beharrean, aukera berri bat aztertuko da *lexnamea* bakarrik iragartzeko aparteko eredu bat birdoituz (**GLEN(Lex)**), eta honek itzultzen dituen probabilitateak jatorrizko (**GLEN(Gloss)**) ereduarekin konbinatuz. 4.2 taulan ikus daiteke eredu bakoitzak iragarri beharko lukeen testuaren formatua.

Adiera	**GLEN(Lex)**	**GLEN(Gloss)**	**Ager. Kop.**
home%1:15:04::	This is a location	The definition of home is: where you live at a particular time	55
home%1:06:00::	This is a man made object	The definition of home is: housing that someone is living in	48
home%1:15:02::	This is a location	The definition of home is: the country or state or city where you live	10

4.2 Taula: *Home* izenaren hiru adiera ohikoenak. GLEN(Lex)ek *lexname*a eta GLEN (Gloss)ek dagokion glosa iragartzen du. Adiera bakoitzaren agerpen kopurua Semcor corpusean kalkulatu da.

Lehen aipatu den moduan, **GLEN(Lex)**ek soilik ezin du adiera konkretu bat hautagai guztien artean identifikatu, baina bai adiera-hautagai kopurua murrizten lagundu. Eredu honen errendimendua ebaluatu ahal izateko bi irtenbide proposatzen dira: ereduak onartzen dituen adiera-hautagaien artean ausaz bat aukeratzea eta adiera-hautagaien artean ohikoena aukeratzea. *Lexname*ek eskaintzen duten informazioa baliagarria izanik, bi ereduen irteera konbinatzea da hurrengo pausua.

4.5.1 Ereduen konbinazioa: GLEN(Lex)+GLEN(Gloss)

Eredu bakoitzak, sarreran emandako lemaren adiera-hautagai bakoitzari probabilitate bat esleitzen dio. Orain arteko esperimentuetan, ereduak itzulitako irteera zuzenean erabiltzen zen adiera identifikatzeko. Atal honetan bi ereduen irteera, eta hauen probabilitateak konbinatzeko aukera desberdinak aztertuko dira.

Probabilitateak konbinatu aurretik, bi ereduen irteera-testua konbinatzea beharrezkoa da. Horretarako ereduek erabiltzen duten testua konbinatuko da **GLEN(Lex+Gloss)** ereduak erabiltzen duen testua sortzeko (ikusi 4.1 taula). Bi ereduen testuak konbinatuz lor daitezkeen konbinazio posible guztiak kalkulatuko dira, kontuan izanik konbinazio hauetako batzuk ez direla existituko. Kasu hauek iragazteko aurreko esperimentua gauzatzeko sortu den hiztegia erabiliko da. 4.3 taulan agertzen dira *Home* izenaren hiru adiera ohikoenak lortzeko adibide moduan jarri diren konbinazio posibleak (4.2 taula).

Jarraian, ereduen zenbakizko probabilitateak konbinatzeko hiru metodo desberdin alderatuko dira. Zenbaki hauek logaritmo negatiboak erabiliz ematen dira; [-inf, 0] eskalan, zenbat eta handiagoak izan (zerotik gertuago) orduan eta probabilitate handiagoa adierazten dute. Probabilitateak konbinatzeko α *alpha* balio bat erabiliko da eredu bakoitzari

Konbinazioa	**Adiera**
This is a location with definition: where you live at a particular time	home%1:15:04::
This is a location with definition: housing that someone is living in	—
This is a location with definition: the country or state or city where you live	home%1:15:02::
This is a man made object with definition: where you live at a particular time	—
This is a man made object with definition: housing that someone is living in	home%1:06:00::
This is a man made object with definition: the country or state or city where you live	—

4.3 Taula: 4.2 taulako **GLEN(Lex)** eta **GLEN(Gloss)** ereduen konbinazio posibleak. Existitzen direnei dagokien adiera kodea esleitzen zaie. Gainerakoak baztertu egiten dira.

pisu bat esleitzeko. Hitz anbiguo baten l lema eta bere t testuingurua izanik, g glosa eta *lex lexname*aren probabilitatea estimatzen da.

- **Probabilitateen biderkadura:** Logaritmo negatiboekin adierazitako probabilitateak biderkatu daitezke biak batuz. Hurrengo formuletan bezala, balio handienak adierazten du probabilitate handiena.

$$p(lex,g|t,l) = pLog(lex|t,l) * (1-\alpha) + pLog(g|t,l) * \alpha$$

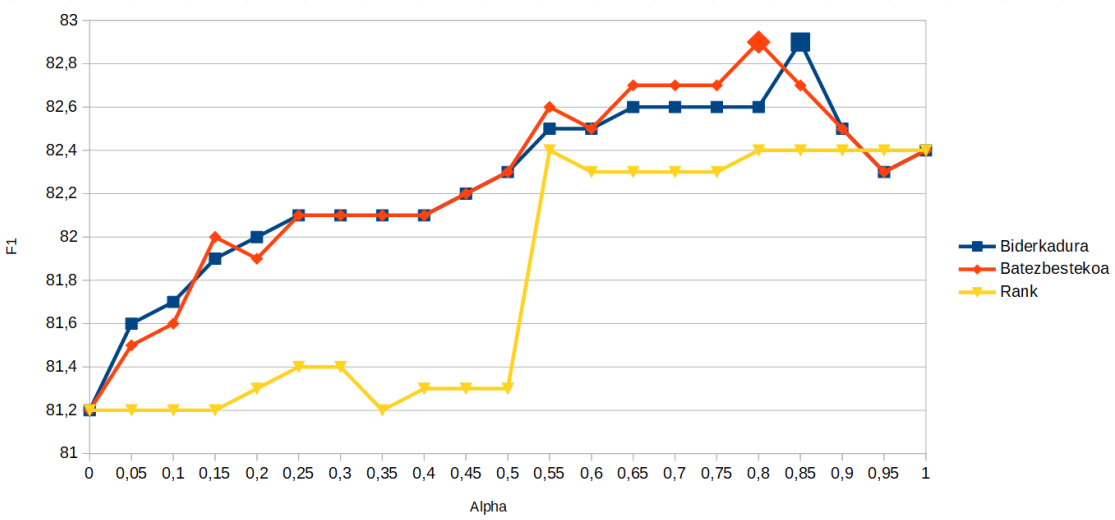
- **Probabilitateen batezbestekoa:** Probabilitate negatiboak positibo bihurtuko dira funtzio esponentziala erabiliz, [0, 1] tartean utziz, eta batezbesteko haztatua egingo da *alpha* parametroaren bidez.

$$p(lex,g|t,l) = p(lex|t,l) * (1-\alpha) + p(g|t,l) * \alpha$$

- **Rank:** Zenbakizko probabilitateak zuzenean erabili beharrean, handienetik txikienera ordenatu, eta adiera-hautagai bakoitzari zenbaki bat esleitzen zaio posizioaren arabera (probabilitate handiena duenari, 1; bigarrenari, 2...). Posizioaren zenbakia batu eta negatibo bihurtuko da, zenbaki positibo txikienak probabilitate handiena adieraz dezan.

$$p(lex,g|t,l) = -(pos(lex|t,l) * (1-\alpha) + pos(lex|t,l) * \alpha)$$

4.1 irudian ikus daiteke hiru metodo hauen errendimendua *alpha* balioaren arabera. $\alpha = 1$ denean $p(g|t,l)$ bakarrik estimatzen da, eta $\alpha = 0$ denean $p(lex|t,l)$ bakarrik. $\alpha = 0$ denean emaitzak beste ereduaren irteerari baldintzatuta daude inplementazio arrazoiak direla eta, baina ez dute beste emaitzetan eraginik. Tarteko balio guztiek bi ereduen konbinazio bat erakusten dute. Ikus daitekeen moduan, biderketa eta batezbestekoa eginez banakako ereduen errendimendua hobetu daiteke +0,5 puntutan, informazio osagarria dutela erakutsiz. Batezbestekoa $\alpha = 0.8$ ezarriz erabiliko da GLENen oinarri bezala hurrengo esperimentuetan, orokorrean batezbestekoak eskaintzen dituelako emaitza onenak. Eredu hau **GLEN(Gloss)+GLEN(Lex)** deituko da aurrerantzean.



4.1 Irudia: GLEN(Lex)+GLEN(Gloss) konbinazioa hiru metodo desberdin erabiliz. Dev2001en emaitza onenak batezbestekoa $\alpha = 0.8$ finkatuz lortzen dira.

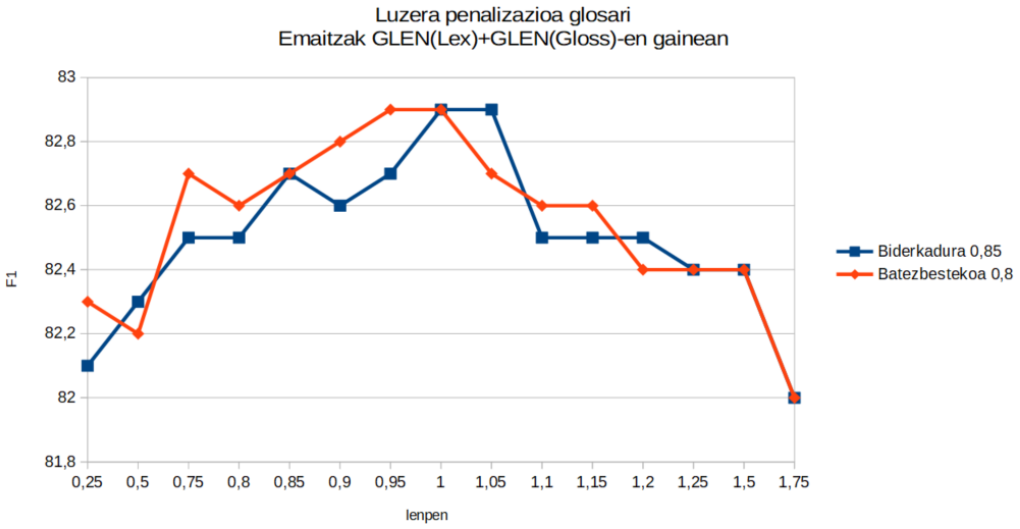
4.6 Esaldi-luzeraren penalizazioa

GENRE-ren egileek proposatzen duten beste hobekuntza bat aztertuko da atal honetan: sortzen den irteeraren luzeraren arabera probabilitatea baldintzatzea. Ez duenez zentzu gehiegirik *lexname*ak sortzen dituen ereduan hau egitea, glosak sortzen dituen ereduaren irteera bakarrik erabiliko da glosen luzera baldintzatzea delako interesgarriena. Honako formula erabiltzen da, non h glosa horrek duen hitz kopurua den. Aurreko atalean bezala, $lenpen$ parametroa luzeraren eragina doitzeko erabiltzen den parametroa da. $lenpen < 1$ denean glosa luzeagoak hobesten dira, eta $lenpen > 1$ glosa laburragoak.

$$p(lex,g|t,l,h) = p(lex|t,l) * (1-\alpha) + (h * p(g|t,l)/h^{lenpen}) * \alpha$$

Hipotesi bat, hitz baten adieren artean aukeratzean luzera handiagoa duten glosak garrantzia handiagoa izan dezaketela da. Beste hipotesi bat, glosa laburragoak ohikoagoak diren adieretara mugatuta egon daitezkeela da, informazio gutxiagoa beharrezkoa baita kontzeptu ezagun hauek deskribatzeko. Hau aztertzeko *lenpen* parametroaren eragina aztertuko da aurreko ataleko emaitza onenen gainean.

4.2 irudian ikus daitekeen bezala, ez du hobekuntzarik eskaintzen luzera handiagoko edo laburragoko glosak hobestea. Bi konbinazio modu hauetaz aparte, desberdinak egiaztatu dira, inolako hobekuntzarik lortu gabe. Hau ikusita, esperimentu hau alde batera utzi, eta hurrengoarekin jarraituko da.



4.2 Irudia: GLEN(Lex)+GLEN(Gloss) konbinazioa. Dev2001en emaitza onenak $\alpha = 0.8$ eta *lenpen* $= 1$ finkatuz lortzen dira. Aurreko esperimentuen emaitzak ez dira hobetzen

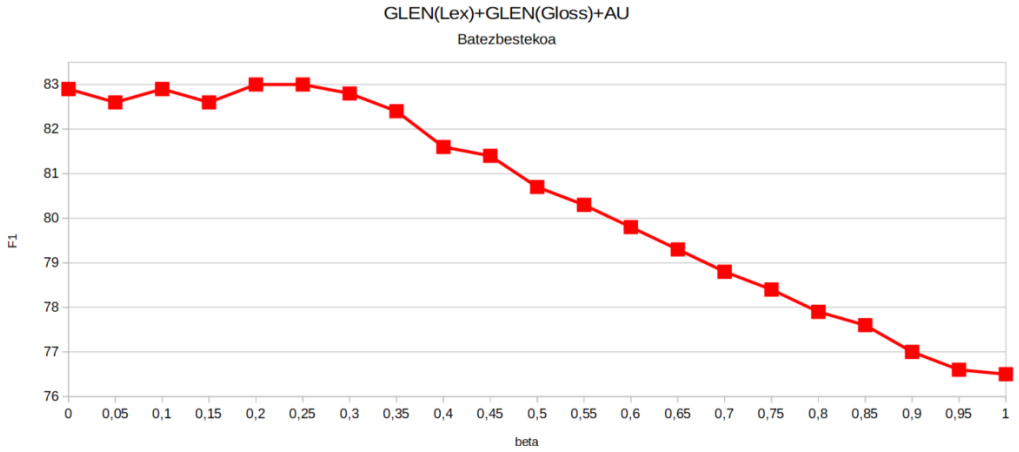
4.7 Adiera usuena konbinatu: GLEN(Lex)+GLEN(Gloss)+AU

Emaitzak hobetze aldera proposatzen den beste hobekuntza bat, adiera usuenaren probabilitatea konbinatzea da azken iragarpena egin aurretik. WordNetek eskaintzen duen informazioaren artean, adiera bakoitza Semcorren eskuz etiketatu den aldi kopurua eskaintzen da. Horretarako, adiera-hautagai guztien agerpen kopurua batu, eta *a* adiera bakoitzari probabilitate bat esleituko zaio hurrengo formula erabiliz. *Agerpenak Adiera* adiera konkretu baten agerpen kopurua Semcorren adierazten du; *Agerpenak Hautagaiak*, aldiz, hitz bat desanbiguatzeko hautagai diren adiera guztien agerpenen batura.

$$p(a|l) = \frac{Agerpenak\ Adiera}{Agerpenak\ Hautagaiak}$$

Semcorren agertzen ez diren adierak ez ezabatzeko (hauen probabilitatea zero izanik), lehenik, adiera guztien agerpen kopuruari $+1$ egingo zaio. Probabilitate hauek konbinatzeko, aurreko esperimentuetan bezala, *beta* parametro bat erabiliko da aurreko esperimentuetako emaitza onenaren gainean, non probabilitateak jada positibo bihurtu diren funtzio esponentzialaren bidez.

$$p(lex,g,a|t,l) = p(lex,g|t,l) * (1-\beta) + p(a|l) * \beta$$



4.3 Irudia: GLEN(Lex)+GLEN(Gloss)+AU konbinazioa. Dev2001en emaitza onenak $\alpha = 0.8$ eta $\beta = 0.25$ finkatuz lortzen dira.

4.3 irudian ikus daitekeen moduan, aurreko esperimentuetako emaitza onenaren (82.9 puntu) eta adieren agerpenen probabilitatearen arteko batezbesteko haztatuak emaitzak oso gutxi hobetzen ditu: 0.1 puntu $\beta = 0.2$ eta $\beta = 0.25$ denean. Esperimentu honekin % 83ko F1era iritsi da, sistemaren oinarri-lerroa 2.4 puntutan hobetuz. Sistema hau **GLEN(Lex)+GLEN(Gloss)+AU** deituko da aurrerantzean.

4.8 Garapen emaitzak

BART azaltzean ikusi den moduan (2.3 atala), ingelesezko BARTen bi bertsio desberdin daude eskuragarri ereduaren geruza kopuruaren arabera. Eskuarki, eredu handiago bat erabiltzeak emaitza hobeak eskaintzen ditu, baina beharrezkoak diren baliabideak nabarmen handitzen dira: txartel grafikoen konputazio ahalmenaren eta exekuzio denboraren

Dev2001 (BART-Base)	**F1 Score**
Ausaz	39.6
Adiera Usuena	76.5
GLEN(Gloss) *	80.6
GLEN(Gloss)	82.4
GLEN(Lex+Gloss) *	80.2
GLEN(Lex+Gloss)	82.5
GLEN(Lex) (Ausaz)	56.0
GLEN(Lex) (Adiera usuena)	78.8
GLEN(Lex)+GLEN(Gloss) ($\alpha = 0.8$)	82.9
GLEN(Lex)+GLEN(Gloss) ($\alpha = 0.8$, $lenpen = 1$)	82.9
GLEN(Lex)+GLEN(Gloss)+AU ($\alpha = 0.8$, $\beta = 0.25$)	**83.0**

4.4 Taula: BART-Base erabiliz garatutako esperimentuen emaitzak. * Semcor bakarrik erabiliz lortutako emaitzak.

aldetik. Ohiko praktika da eredu txikiagoa erabiltzea esperimentuetarako eta azken pausuan eredu handia exekutatzea. Orain arteko esperimentu guztiak BART-Base bertsioa erabiliz gauzatu dira. Esperimentu hauen emaitzak 4.4 taulan aurkezten dira. Nahiz eta kapitulu amaieran emaitzak sakonago aztertuko diren GLENen behin-betiko egitura definitzeko (4.10 atalean), orain arteko esperimentuak komentatzea interesgarria da.

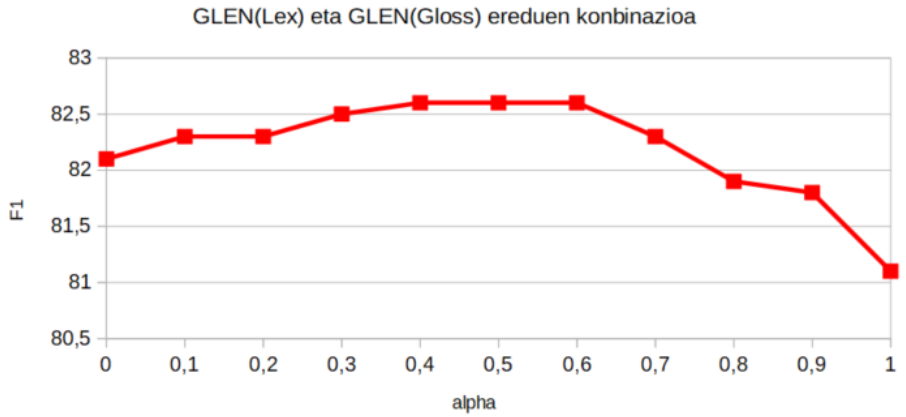
Entrenamendu corpusean Semcorri WNGE gehituz emaitzak asko hobetzen dira, 2 puntu inguruko hobekuntza lortuz. *Lexname*ak eta glosak eredu bakarrean iragartzeak ez ditu emaitzak hobetzen, baina *lexname*ek hautagai kopurua murrizteko baliagarriak direla ikusten da; adiera usuena bi puntutan hobetzeko gai dira, eta ausaz aukeratuz emaitzak nabarmenki hobetzen dira. Bi eredu desberdin erabiliz *lexname*ak eta glosak iragartzeko, eta hauen probabilitateak konbinatuz, emaitzak 0.4 puntu hobetzen dira. Luzera penalizazioak ez du inolako hobekuntzarik ekartzen, eta adiera usuenaren konbinazioak nahiz eta emaitzetan apur bat hobea izan, 0.1 puntuko hobekuntza ez da orokorrean aintzat hartzen.

BART-Base erabiltzen duen GLENen bertsio bat egiteko **GLEN(Lex)+GLEN(Gloss)** ($\alpha = 0.8$) proposatzen da. Jarraian esperimentu hauetako batzuk errepikatuko dira BART-Large erabiliz emaitzak oraindik eta gehiago hobetzeko.

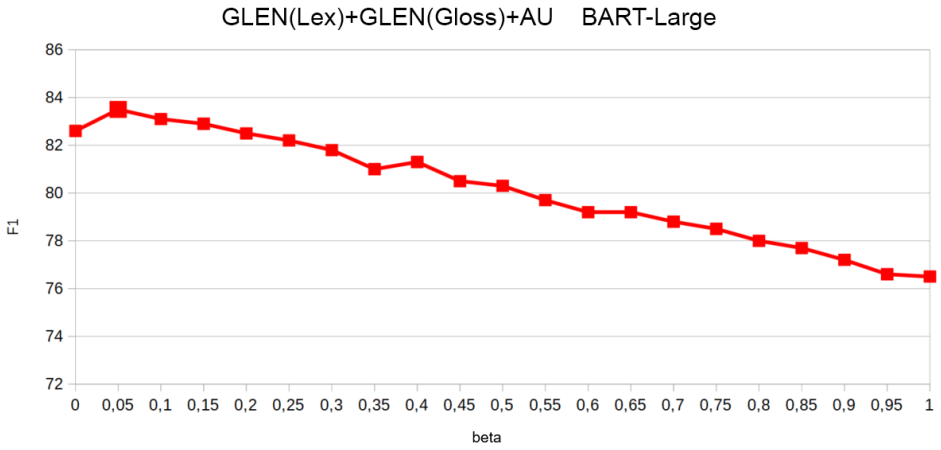
4.9 BART eredu handiagoa erabili

Atal honetan, BART-Large eredua erabiliz esperimentu batzuk errepikatuko dira, aurreko esperimentuetako parametroak aldatu daitezkeelako eredu berriak entrenatzean. BART-

Base erabiliz emaitzak hobetu ez dituzten esperimentuak baztertuko dira (GLEN(Lex+Gloss) eta luzera penalizazioa). Lehenik, **GLEN(Lex)** eta **GLEN(Gloss)** ereduen konbinaziorako *alpha* parametroa finkatuko da. 4.4 irudian ikus daitekeen moduan batezbesteko sinpleak ematen ditu emaitza onenak ($\alpha = 0.5$). BART-Baserekin lortutako emaitzekin alderatuz apur bat okerragoak dira. Hau arraroa da, baina emaitza hauei adiera usuena konbinatuz emaitzak nabarmenki hobetzen dira 4.5 irudian ikus daitekeen moduan, $\beta = 0.05$ finkatu ondoren emaitzak ia puntu bat igotzen dira.



4.4 Irudia: GLEN(Lex)+GLEN(Gloss) konbinazioa BART-Largen oinarritutako ereduak erabiliz. Dev2001en emaitza onenak batezbestekoa $\alpha = 0.5$ finkatuz.



4.5 Irudia: GLEN(Lex)+GLEN(Gloss)+AU konbinazioa BART-Largen oinarritutako ereduak erabiliz. Dev2001en emaitza onenak $\alpha = 0.5$ eta $\beta = 0.05$ finkatuz lortzen dira.

BART-Large erabiliz lortutako emaitzak 4.5 taulan aurkitu daitezke. BART-Baseren emaitza onenak 0.5 puntutan hobetu dira, oinarri-lerroa hiru puntutan hobetuz; hori dela eta, hau izango da hurrengo atalean GLENen egitura definitzeko erabiliko den egitura.

Dev2001 (BART-Large)	**F1 Score**
Adiera Usuena	76.5
GLEN(Gloss)	81.1
GLEN(Lex) (Adiera usuena)	79.6
GLEN(Lex)+GLEN(Gloss) ($\alpha = 0.5$)	82.6
GLEN(Lex)+GLEN(Gloss)+AU ($\alpha = 0.5$, $\beta = 0.05$)	**83.5**

4.5 Taula: Garapen faseko emaitzak BART-Large erabiliz. Atal honetako esperimentuak. Bi eredu konbinatuz emaitzak hobetzen dira, eta are gehiago hobetu daitezke adiera usuenaren informazioa konbinatuz.

4.10 Esperimentuen ondorioak

Kapitulu honetan GLENen oinarri-lerroa hobetzeko asmoarekin egin diren esperimentu guztiak azaldu dira, Dev2001 datu-multzoa erabiliz emaitzak alderatu ahal izateko. Nahiz eta emaitza hauek oso fidagarriak ez izan ebaluazio datu-multzoei begira, ondorio interesgarriak atera daitezke. Ondorio hauek kontuan izanik, GLENen egitura definituko da honako erabakiak hartuz:

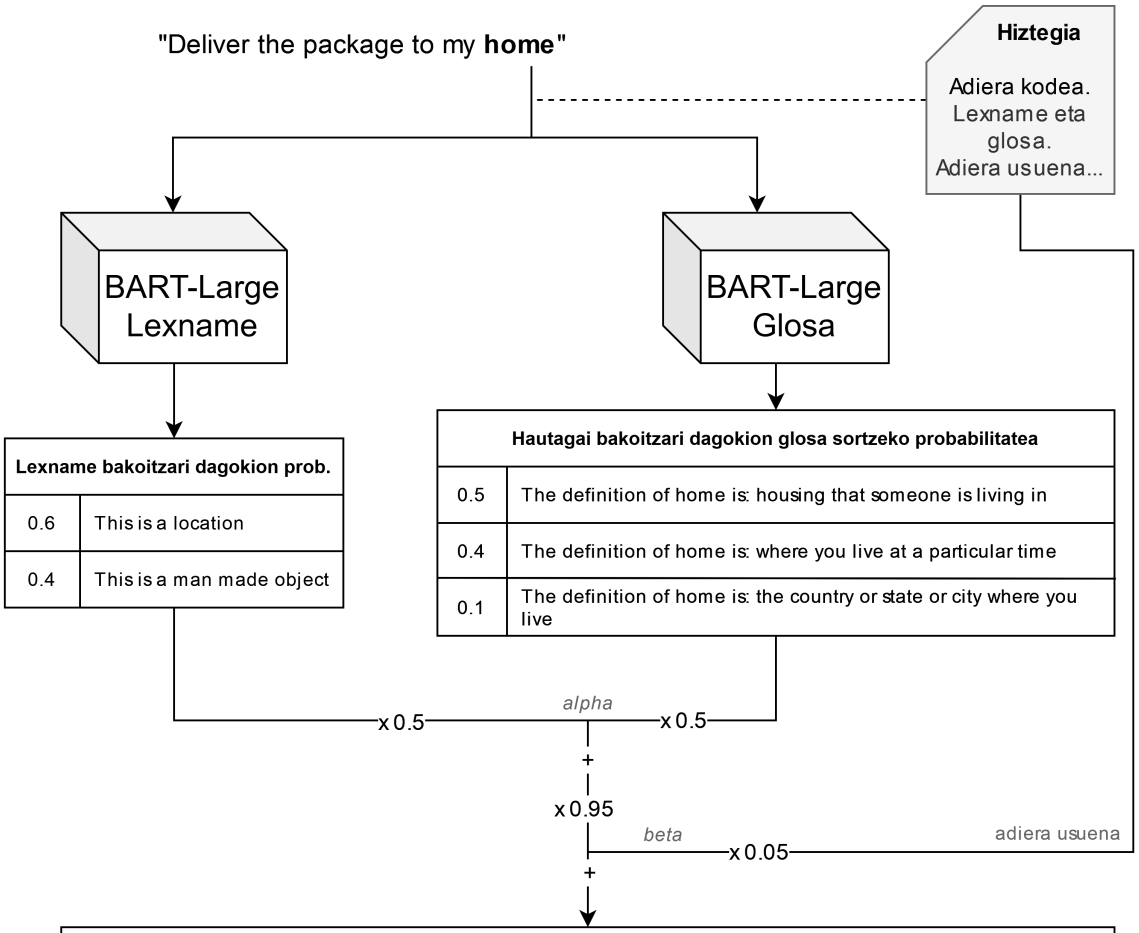
1. Entrenamendu kopurua handitzeak hobekuntza nabaria dakar. **Semcor**rez aparte **WNGE** ere erabili da sistema entrenatzeko.

2. Emaitzak oraindik eta gehiago hobetu daitezke *lexname*en informazioa erabiliz. Horretarako bi eredu desberdin entrenatu dira: **GLEN(Lex)** eta **GLEN(Gloss)**. Bi ereduek itzultzen dituzten probabilitateak **batezbesteko sinplea** ($\alpha = 0.5$) eginez konbinatzen dira.

3. Ereduak lortzeko, aurre-entrenatutako **BART-Large** eredu bat birdoitu da.

4. Adieren agerpen kopuruaren probabilitatearen informazioa (**adiera usuena**) ere **konbinatu da** $\beta = 0.05$ balioa erabiliz.

Hau kontuan izanda, 4.6 irudian erakusten den egitura proposatzen da HAD gauzatzeko. Aukera honetan bi BART-Large eredu desberdin erabiltzen dira adiera iragartzeko (**GLEN(Lex)** eta **GLEN(Gloss)**), eta hauen probabilitateak Semcorretik lortutako adiera usuenarekin konbinatzen dira. Hurrengo kapituluan konfigurazio hau erabiliko da GLENen behin-betiko emaitzak lortzeko.



4.6 Irudia: GLENen probabilitateen kalkuluaren adibidea. *Home* hitza desanbiguatzen da glosa eta *lexname*en informazioa konbinatuz. Tauletan sistemen irteera eta bakoitzari dagokion probabilitatea (0-1 eskalan) adierazten da. Adiera zuzena aukeratzen da.

5. KAPITULUA

Emaitzak

Kapitulu honetan, GLEN 2.2.4 ataleko datu-multzoetan ebaluatuko da, eta arloaren egoerako beste artikuluekin alderatuko da (Generationary hauen artean). Ingelesezko emaitzez gain, ingelesez entrenatutako sistema eleanitz bat euskarazko HADen ebaluatuko da mBART eredu bat eta EuSemcor datu-multzoa erabiliz.

5.1 Arloaren egoerako HAD sistemak

GLENek lortzen dituen emaitzak komentatu ahal izateko, azken urteetako arloaren egoera aztertzeko hiru sistema erakusten dira: SVC, EWISER eta ConSeC. Sistema hauek Semcor + WNGE erabiliz entrenatu direlako, eta arloaren egoeraren ikuspegi zabala erakusten dutelako aukeratu dira.

SVC [Vial et al., 2019] WordNeteko erlazioez (sinonimia, hiperonimia eta hiponimia) baliatzen da aurre-entrenatutako BERT hizkuntza-eredu baten hitz-bektoreak konprimitu, eta kontzeptu berari lotuta dauden adieren errepresentazioak talde batean biltzeko. Ebaluazio garaian, *softmax* sailkatzaile baten bidez hitz anbiguoaren errepresentazioa talde hauen artean sailkatzen da. Modu honetan entrenamenduan erabilitako adibideen informazioa beste adieretara estrapolatzen da, gorde beharreko informazioa nabari murriztuz.

EWISER [Bevilacqua and Navigli, 2020], ingelesezko datu-multzoetan % 80ko F1 lortu zuen lehen sistema izan zen. SVC bezala, WordNeteko erlazioez baliatzen da, hala nola,

beste sistemen ezagutzaz, hauen hitz-bektoreak erabiliz EWISER hasieratzeko. Azkenik, adiera guztien artean sailkapena gauzatzen du.

ConSeC [Barba et al., 2021] da interesgarriena HAD atazarako. Esaldi bateko hitz anbiguo bakarra desanbiguatu beharrean, esaldian agertzen diren bestelako hitz anbiguo guztiak identifikatzen ditu. Jarraian, algoritmoaren iterazio bakoitzean, adiera-hautagai gutxien dituen hitza desanbiguatzen du (hau da, desanbiguatzeko hitz errazena) eta honen glosaren informazioa erabiltzen du hurrengo hitzak desanbiguatzeko. Ikuspegi honek abantaila handia ematen dio sistema honi, eta etorkizuneko lanetarako interesgarria izan daiteke. Arkitektura, transformerretan oinarritutako DeBERTa [He et al., 2021] ereduan oinarritzen da, honen gainean sailkatzaile lineal bat gehituz.

5.2 Ingelesezko emaitzak

Atal honetan, GLENen ebaluazioa aztertzen da 5.1 taulako emaitzak aurkeztuz. *Senseval* eta *SemEval* lehiaketetatik ateratako bost datu-multzo erabiltzen dira, arloaren egoerako HAD sistemekin alderatzeko. GLENen bi bertsio aurkezten dira, BART bertsioen arabera.

- **GLEN-Base:** BART-Base ereduan oinarrituta, **GLEN(Lex)+GLEN(Gloss)** eredua erabiltzen da, hauen konbinaketarako $(\alpha = 0.8)$ erabiliz.
- **GLEN-Large:** BART-Large ereduan oinarrituta, **GLEN(Lex)+GLEN(Gloss)+AU** eredua erabiltzen da, konbinaketarako $\alpha = 0.5$ eta $\beta = 0.05$ erabiliz. Eredu honen ezaugarriak azaldu dira 4.10 atalean eta 4.6 irudian.

Taulan lehenik, Semcor erabiliz kalkulatutako adiera usuena aurkezten da. Honen helburua nolabaiteko oinarri-lerroa aurkeztea da; oso modu sinplean kalkulatu daitekeen arren, urte askoz ataza zaila izan da oinarri-lerro hau gainditzea. Jarraian, arloaren egoerako hiru sistema erakusten dira, aurreko atalean azalduak. Azkenik 3.2 atalean aurkeztu den Generationary sistemak eskaintzen dituen emaitza onenak agertzen dira, GLENen emaitzekin guztiz konparagarriak antzeko ikuspegiaz baliatzen baita.

GLENen emaitzei dagokionez, ikus daiteke emaitza lehiakorrak eskaintzen dituela, baina arloaren egoerako emaitza onenen urrun geratzen da. Hala ere, arloaren egoerako sistema sortzaileetan oinarritutako emaitzak gainditzen ditu zabalki, Generationary ia bi puntuz gainditzuz. Bi BART eredu desberdin entrenatuz eta bien probabilitateak konbinatuz emaitzak hobetzea lortu da, Generationary-ren maila berdinera iritsiz BART bertsio txikiago

Sistema	SE2	SE3	SE07	SE13	SE15	ALL
Adiera usuena	65.2	62.0	54.5	63.8	67.1	65.5
SVC	79.7	77.8	73.4	78.7	82.6	79.0
EWISER	80.8	79.0	75.2	80.7	81.8	80.1
ConSeC†	82.7	81.0	78.5	85.2	87.5	83.2
Generationary	78.0	75.4	71.9	77.0	77.6	76.7
GLEN-Base	77.9	75.0	69.7	77.4	78.4	76.6
GLEN-Large	**79.9**	**77.0**	**73.0**	**78.8**	**79.9**	**78.5**

5.1 Taula: Ingelesezko datu-multzoen gaineko emaitzak, GLENen emaitzak Generationary sistemarekin alderatzen dira. Beste Semcor+WNGE erabiliz entrenatutako sistemak ere aurkezten dira erreferentzia moduan, arloaren egoeraren ordezkari gisa. Gure emaitza onenak beltzez. (†) arloaren egoerako sistema onena.

batekin. Honek abantaila asko ditu gure sistemaren inplementazio aldetik, baliabide gutxiagoko makinetan exekutatzea baimenduz. BART eredu handiago bat erabiliz emaitzak nabarmenki hobetzen dira ebaluazio datu-multzo guztietan.

5.3 Emaitza eleanitzak

Atal honetan, GLENen bertsio eleanitza aztertzen da ebaluazio datu-multzo desberdinen gainean. Bertsio hau sortzeko mBART eredu eleanitzen abantailaz baliatzen da *zero-shot transfer learning* teknika erabiliz. Teknika honetan mBART.cc100 (ikusi 2.3.4 atala) **eredua ingelesezko datuekin birdoitzen da** orain arteko esperimentuetan bezala, baina **ebaluazio garaian euskaraz idatzitako esaldi bat ematen zaio sarrera gisa, eta sistemak ingelesez idatzitako glosak iragarri behar ditu.** Ideia hau erabilgarria da entrenamendu datu kopuru gutxi eskuragarri dauden hizkuntzetan; hizkuntza handi batekin lortutako ezagutza beste hizkuntzetara eramatea posible baita hizkuntzaren prozesamenduko ataza askotan. Hau oso erabilgarria suertatzen da euskara bezalako hizkuntzetan. Emaitza hauek lortzeko hiru muga garrantzitsu identifikatu dira funtzionamendurako:

1. Ez dago euskararaz bakarrik aurre-entrenatutako BART bertsiorik.
2. Ez dago nahikoa daturik sistemaren entrenamendurako. EuSemcor [Pociello et al., 2011] corpusa da bakarra eta adibide kopurua ingelesezko Semcor baino 10 aldiz gutxiago izateaz gain, 400 izen ohikoen adibideak bakarrik biltzen ditu. Entrenamendu eta ebaluazio datu-multzoak partizio berdinetik datozenez, emaitzak ez

lukete islatuko mundu errealeko errendimendua, gabeziak ez lirateke agerian geratuko.

3. Ez dago hitz guztientzako euskarazko glosarik eskuragarri. Ingelesez adiera guztien estaldura lor daiteke, euskaraz ez bezala.

Honetaz gain, sistema eleanitz bat sortzean ezin daitezke hizkuntza bateko adiera-kodeak erabili beste hizkuntza bateko adierak identifikatzeko. Nahiz eta kontzeptuak hizkuntza arteko elementuak izan (ez daude hizkuntzaren menpe), lemak ez dira baliokideak, eta ondorioz, hitzen eta kontzeptuen arteko erlazioak ezin daitezke gauzatu. Hori dela eta, WordNeteko kontzeptu-kodeak, *synset*ak, erabili behar dira adiera-kodeen partez. Esperimentuetan ikusitako bi eredu desberdin konbinatzearen metodoa bateraezin bihurtzen da, **GLEN(Gloss)** ereduak lemak erabiltzen baititu irteerako adiera-hautagaiak identifikatzeko. Eredu hau aldatzeak inplementazio kostu handiak ditu, eta ezin daitekeenez ziurtatu emaitzetan hobekuntzarik eskainiko duen ala ez, GLENen bertsio sinpleago bat erabiltzen da: **GLEN(Lex+Gloss)** egiten duen eredu bakarra birdoitzen da 125 hizkuntzetan aurre-entrenatu den mBART eredu bat erabiliz (ikusi 2.3.4 atala). mGERNE sisteman [Cao et al., 2021b], BART eredu bat aurre-entrenatu zuten *cc100* corpusa erabiliz [Wenzek et al., 2020]. Jarraian, eredu hau birdoitu zuten Wikipediako datuak erabiliz 125 hizkuntza desberdinetan entitateen berreskurapena egiteko. Eredu honek, beraz, euskarazko datuak ikusi ditu aurre-entrenamenduan beste hainbat hizkuntzekin batera, baina ez du euskarazko ezagutzarik jaso HAD atazan birdoitzean.

Emaitza eleanitzak 5.2 taulan ikus daitezke. Euskarazko emaitzak alderatzeko adiera-usuena, eta [Aguirre Blanco, 2020] laneko emaitzak erabiltzen dira euskarazko arloaren egoera definitzeko. GLEN eleanitzaren emaitzak oso baxuak dira: ausaz adiera bat aukeratzeak F1 altuagoa lortzen du. Hau gertatzeko hipotesi desberdinak aztertzen dira. Hipotesi hauek aztertzeko, beste bi esperimentu eleanitz gauzatu dira 5.2 taulan ikusgai.

Sistema	**Euskara (EuSemcor)**	**Gaztelera (SemEval2013-es)**	**Ingelesa (Dev2001)**
Ausaz	28.5	35.6	39.6
Adiera usuena	67.1	69.5	76.9
Arloaren egoera	**61.8**	**78.8**	-
GLEN-Large	-	-	**81.1**
GLEN (mBART.cc100)	24.2	32.4	79.4

5.2 Taula: Datu-multzo eleanitzen ebaluazioa GLENen bertsio elebakarra eta eleanitza erabiliz. Euskarazko emaitzak nire gradu amaierako lanean lortutako emaitzekin eta gaztelerakoak EWISER sistemarekin alderatzen dira.

Entrenatutako sistema eleanitzean arazorik ez dagoela baieztatzeko, ingelesezko garapen faseko Dev2001 datu-multzoaren gainean ebaluatu da. Garapeneko ingelesezko errendimendua baino apur bat okerragoa da bertsio eleanitzean, baina inolaz ez euskarazkoan bezala. Honek erabilitako ereduan arazorik ez dagoela baieztatzen laguntzen digu: 125 hizkuntzatako eredua gai da HAD gauzatzeko, eta ingelesez emaitza lehiakorrak eskaintzeko.

Ingelesezko emaitzek ez dute laguntzen, ordea, arazoa euskarazko ezagutza edo *zero-shot transfer learning* teknika den jakiten. Kontuan izan behar da mBART ereduak gaztelerazko informazio askoz gehiago jaso duela aurre-entrenamenduan. Hau egiaztatzeok, gaztelerazko ebaluazio datu-multzo bat erabili da sistema ebaluatzeko. Emaitza hauetan ere oso errendimendu txarra ikus daiteke, euskarazko emaitzen parekoak, eta ausaz adiera bat aukeratzearen azpitik. Modu honetan, sistemaren errendimendu eleanitz eskasaren arrazoia *zero-shot transfer learning* teknikan aurki daitekeela esan daiteke. GLENen sarrera (hitz anbiguoa) hizkuntza batean, eta irteera (glosa) hizkuntza desberdin batera mugatzeak arazoak ekartzen ditu, eta ez da gai euskarazko HAD egiteko, beharrezkoa den ingelesezko ezagutzaz baliatzeko.

Nahiz eta gure kasuan funtzionatu ez izan, *transfer learning*a maiz erabili da HPko beste atazetan, emaitza onak lortuz. Arazoa, beraz, *zero-shot* egitean aurkitu daiteke; hau da, euskarazko esaldi bat emanik, ingelesezko glosa bat sortzeko ahalmenean. Sistemak, HAD egiteko beharrezkoa den ezagutzaz gain, itzulpen automatikoko sistema batek bezala funtzionatu behar du: euskarazko esaldiak kodetzen duen informazioa ingelesera itzuliz. GLENen bertsio eleanitzak HAD egiteko ezagutza ikasi du birdoitze prozesuan, baina ez du inoiz ikusi euskara-ingeles daturik, ez aurre-entrenamenduan, ez birdoitze prozesuan.

6. KAPITULUA

Ondorioak eta etorkizuneko lanak

Kapitulu honetan, proiektua garatu bitartean lortu diren ondorioak eta emaitzen hausnarketa azaltzen dira. Honez gain, proiektuari jarraipena eman nahi izanez gero, etorkizunean egin daitezkeen hobekuntzak eta bestelako esperimentuen proposamenak aurkezten dira.

6.1 Ondorioak

Proiektu hasieran zehaztutako helburuak betetzea lortu da: entitateen berreskurapena gauzatzeko diseinatu zen sistema bat HAD gauzatzeko eraldatzea, eta honen emaitzak hobetzeko proposamenak aztertzea. Esperimentu anitz gauzatzeko aukera izan da, bururatutako hobekuntza posible guztiak aztertuz, eta hauei esker antzeko sistemen emaitzetara hurbiltzea eta gainditzea. Sistema sortzaileak oinarri bezala erabiltzen dituen arloaren egoerako sistema onena gainditzea lortu da, GLEN ekarpen garrantzitsua izanik. Entrenamendu datu-multzoen formatua aldatuz, eta hiztegiko glosak identifikatzaile moduan erabiliz, sistema sortzaileen ahalmena frogatu da modu honetako atazak gauzatzeko. Gainera, hiztegiek eskaintzen duten bestelako informazioa erabilgarria dela ikusi da, desanbiguatu beharreko adieraren informazio gehigarri moduan *lexname*ak erabiliz.

Nahiz eta duela urte bateko arloaren egoerarekin alderatuta (% 80 F1 gainditu berri zen *ALL* datu-multzoan) GLENen emaitzak oso lehiakorrak izan (% 78 F1), azken hilabeteetako argitalpenek muga oso goian jarri dute, F1 % 3 puntutan hobetuz denbora gutxian. Honek zalantzan jar dezake sistema sortzaileen ahalmena etorkizunera begira HAD gau-

zatzeko, hizkuntza-eredu berrien erabilerari esker (DeBERTa, ikusi 5.1 atala) emaitzak asko hobetu baitira.

Emaitza eleanitzei dagokionez, ez zen espero euskarazko emaitza aipagarririk lortzea, eta lortutako emaitzek argi utzi dute ikuspegi desberdinak beharrezkoak direla ataza hau ebazteko. Argi geratzen da euskarazko sistema egoki bat lortzeko beharrezkoa dela BART eredu bat euskarazko datuekin birdoitzea, euskarazko HAD datu-multzo bat erabiliz. Hau lortzea ez da ataza erraza, entrenamendu datu kopuru handia beharrezkoa baita, honek dakarren kostu ekonomiko garrantzitsuarekin. Euskarazko HAD egiteko gai den sistema bat izatea baliabide preziatua izango litzateke eredu sortzaileen ikuspegitik antzeko atazak ebatzi ahal izateko. Nahiz eta gure emaitzetan arrakastarik ez izan, posible da eredu eleanitz bat ingelesezko datuekin birdoituz gero, eta euskarazko datu gutxi batzuk gehituz birdoiketa prozesu horretan, emaitzak hobetu ahal izatea. Esperimentu hau etorkizuneko lanetarako utziko da, ikerketa bide bat irekita utziz.

Nahiz eta, orokorrean, entitateekin lan egitean sistema sortzaileek prozesamendu ahalmen gutxiago eskatu; txartel grafikoen erabilgarritasuna, eta entrenamendu zein ebaluazio atazak gauzatzeko ordu kopuru handiak botila-lepo handia izan dira esperimentuak gauzatu ahal izateko, lanaren lehen pausuak asko motelduz. Hala ere, garapen datu-multzo bat eraikiz (Dev2001) eta BART-Base bertsioa erabiliz, eginiko esperimentuak modu azkarrago batean exekutatzea lortu da, emaitzak alderatzeko erabilgarria izaten jarraituz, eta sistema doitzeko nahikoa informazio eskainiz.

6.2 Etorkizuneko lanak

Nahiz eta hasiera batean ezarritako helburuak bete ahal izan, mota honetako proiektuek ez dute bukaera finkorik izaten. Teknologiak aurrera egin ahala, eta teknologia berriak ustiatzeko teknika berriak garatu ahala, garrantzitsua da etorkizunean azter daitezkeen bideak irekita mantentzea. Hori dela eta, honako hobekuntzak proposatzen dira jarraipen moduan.

1. **Euskarazko sistema lehiakor bat lortu:** Emaitzak azaltzean aipatu den moduan, BART eredu bat euskarazko datuekin birdoituz gero posible izango litzateke datu hauek ingelesezko entrenamendu datu anitzez aberastea. Honela, HAD gauzatzeko nahikoa informazio ikasiko luke ingelesezko datuetatik, eta euskarazko informazioa ingelesera itzultzen ikasiko luke euskarazko datu urri erabiliz. Gainera, HAD

sistema bat bestelako atazak gauzatzeko egokitu daitekeenez modu erraz batean, euskarazko HAD sistema lehiakor bat ekarpen oso interesgarria izango litzateke etorkizunera begira.

2. **GLEN beste atazetan ebaluatu:** Generationary azaltzean ikusi den moduan, sistema jatorri batean glosa-sorkuntza gauzatzeko garatu zen, eta ataza hau ebaluatzeko zailtasuna kontuan izanik, HAD egiteko egokitu zen. Interesgarria izango litzateke gure sistemak erabiltzen dituen bilaketa zuhaitzak ezabatzea, eta sorkuntza librea egiten uztea. Honela, sistema glosa-sorkuntza bezala ebaluatu ahalko litzateke, eta glosa egokiak sortzeko duen gaitasuna aztertu.

3. **ConSeC-en ikuspegia erabili:** Aurreko kapituluan, HAD gauzatzeko egungo sistema onena laburki azaldu da, ikuspegi desberdin bat erabiliz ataza ebazteko. Sistema honetan esaldi bateko hitz anbiguo guztiak identifikatzen dira, eta errazenetik zailenera desanbiguatzen dira (hautagai kopurua kontuan izanik), desanbiguatzen den bakoitzaren informazioa erabiliz hurrengoetan.

   GENREren azalpenean, sistema entrenatzen den atazetako bat Muturretik-Muturrerako Entitateen Ezagutze eta Desanbiguazioa dela ikusi da. Ataza honetan, esaldi bateko izen-aipamen guztiak identifikatzen eta desanbiguatzen dira. Hau oso erabilgarria izan daiteke ConSeC-en funtzionamendua erreplikatzeko, sistemaren hitz anbiguoak identifikatzeko gaitasunaz baliatuz zuzenean.

   Oso posible da sistema sortzaileen arloaren egoerako emaitzak hobetu ahal izatea esaldian agertzen diren beste hitz anbiguoen glosak erabili ahalko balira HAD gauzatzeko. Ikuspegi hau aztertzea oso interesgarria litzateke; sistema sortzaileak oso tresna ahaltsuak izanik, etorkizuneko sistemak garatzerako orduan kontuan izan beharreko tresna bat baitira.

Eranskinak

A. ERANSKINA

Fitxategi lexikografikoen izenak

Lexname id	Identifikatzaile testua
01	This is an adjective
02	This is a relational adjective
03	This is an adverb
04	This is a noun
05	This is an action
06	This is an animal
07	This is a man made object
08	This is an attribute of people and objects
09	This is a body part
10	This is a cognitive process
11	This is a communicative process
12	This is a natural event
13	This is a feeling
14	This is a food
15	This is a group
16	This is a location
17	This is a goal
18	This is a natural object
19	This is a person
20	This is a phenomenon

21	This is a plant
22	This is a possession
23	This is a natural process
24	This is a quantity
25	This is a relation
26	This is a shape
27	This is a state of affair
28	This is a substance
29	This is a temporal relation
30	This is a verb about dressing and bodily care
31	This is a verb about change
32	This is a verb about cognition
33	This is a verb about communication
34	This is a verb about competition
35	This is a verb about consumption
36	This is a verb about contact
37	This is a verb about creation
38	This is a verb about emotion
39	This is a verb about motion
40	This is a verb about perception
41	This is a verb about possession
42	This is a verb about social
43	This is a verb about stative
44	This is a verb about weather
45	This is a participal ajective

A.1 Taula: Lan honetan WordNeteko *lexname*ak identifikatzeko erabiltzen den testua.

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.

[Agirre and Soroa, 2009] Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece. Association for Computational Linguistics.

[Aguirre Blanco, 2020] Aguirre Blanco, T. (2020). Adiera desanbiguazioa euskararako ikasketa sakona erabiliz. Online.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.

[Barba et al., 2021] Barba, E., Procopio, L., and Navigli, R. (2021). ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Barrena et al., 2018] Barrena, A., Soroa, A., and Agirre, E. (2018). Learning text representations for 500K classification tasks on named entity disambiguation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 171–180, Brussels, Belgium. Association for Computational Linguistics.

[Bevilacqua et al., 2020] Bevilacqua, M., Maru, M., and Navigli, R. (2020). Generationary or: "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

[Bevilacqua and Navigli, 2020] Bevilacqua, M. and Navigli, R. (2020). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

[Broscheit, 2019] Broscheit, S. (2019). Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

[Cao et al., 2021a] Cao, N. D., Izacard, G., Riedel, S., and Petroni, F. (2021a). Autoregressive entity retrieval. In *International Conference on Learning Representations*.

[Cao et al., 2021b] Cao, N. D., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., and Petroni, F. (2021b). Multilingual autoregressive entity linking. In *arXiv pre-print 2103.12528*.

[Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

[Edmonds and Cotton, 2001] Edmonds, P. and Cotton, S. (2001). SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

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

[He et al., 2021] He, P., Liu, X., Gao, J., and Chen, W. (2021). {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

[Hoffart et al., 2011] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[Lazic et al., 2015] Lazic, N., Subramanya, A., Ringgaard, M., and Pereira, F. (2015). Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.

[Le and Titov, 2018] Le, P. and Titov, I. (2018). Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

[Lewis et al., 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, pages 74–81.

[Liu et al., 2020] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation.

[Moro and Navigli, 2015] Moro and Navigli (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. 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.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th*

Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Petroni et al., 2021] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. (2021). KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

[Piccinno and Ferragina, 2014] Piccinno, F. and Ferragina, P. (2014). From tagme to wat: A new entity annotator. In *Proceedings of the First International Workshop on Entity Recognition and Disambiguation*, ERD '14, page 55–62, New York, NY, USA. Association for Computing Machinery.

[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.

[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.

[Radford and Narasimhan, 2018] Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

[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.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

[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.

[Vial et al., 2019] Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

[Wenzek et al., 2020] Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

[Wu et al., 2020] Wu, L., Petroni, F., Josifoski, M., Riedel, S., and Zettlemoyer, L. (2020). Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.