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

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Hizkuntzen arteko transferentzia analisia Euskarazko Informazio-Erauzketan

Egilea

Mikel Zubillaga

2023

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

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

Hizkuntzen arteko transferentzia analisia Euskarazko Informazio-Erauzketan

Egilea

Mikel Zubillaga

Zuzendariak

Oier Lopez de Lacalle eta Oscar Sainz

Laburpena

Hizkuntzaren prozesamendua goren une batean dagoen gai bat da. Izan ere, azken aurrerapenekin, konputagailuek hizkuntza naturala prozesatzeko gaitasun asko irabazten hari dira. Gaitasun horiek garatzen lan egiten duen hizkuntzaren prozesamenduko arlo garrantzitsuetako bat da Informazio Erauzketa. Honen helburua, testu batetik informazio garrantzitsuena ateratzea izanik.

Gure lana informazio erauzketan zentratuko den arren, haratago joateko intentzioa dago, hizkuntzen arteko transferentzia neurtuz. Hau da, IE hizkuntza jakin batean aplikatzeko prestatuta dagoen eredu bat hartu eta beste hizkuntza batean ebaluatuko da. Hau egitean, hizkuntzen arteko transferentzia neurtzeaz gain, hizkuntza desberdinen ezaugarri linguistikoak transferentzian duten eragina aztertu nahi da. Gure kasuan euskara erabiliko dugu transferentzia neurtzeko, honek dituen ezaugarri topologiko partikularrengatik.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Gizakiontzako hizkuntza informazioa partekatzeko biderik sinpleena den arren, konputazionalki arazo asko ekartzen ditu: anbiguotasuna eta hizkuntzaren aldakortasuna besteak beste. Arazo horien konponbideak ikertzen ditu hizkuntzaren prozesamenduak. Honen arloan gai garrantzitsuetako bat Informazio Erauzketa da. IE edo Informazio Erauzketa testu batetik nahi den informazioa automatikoki lortzen datza, testua formatu egituratu batera bihurtuz. IE-ren ataza nabarmenetako bat Gertaera Erauzketa edo EE (ingeleseko siglengatik) da, zeinak helburu moduan testu batetik entitate, gertaera eta gertaeren argumentuak lortzea du.

Historikoki EE hizkuntzaren prozesamenduko ataza garrantzitsuenetako bat izan da. Hori dela eta, asko ikertu da honek zuzen funtziona dezan Nguyen et al. (2016); Ahn (2006). Hala ere, ikerketa eta garapen gehienak hizkuntza handietan (bereziki ingelesean) zentratuta daude eta hauentzako garatzen dira erabiltzen diren datu-multzo gehienak. Honen adibide gisa ACE Walker et al. (2006) datu-multzoa jarri genezake, zeina IE egiteko gehien erabili den datu-multzoa izanik soilik ingelesez, txineraz eta arabieraz dago anotatua. Hau dela eta beste hizkuntza txikiago batzuk bigarren plano batean gelditzen dira. Hizkuntza txiki hauen taldean sartzen da euskara, ez baitaude EE ereduak euskaraz entrenatu ahal izateko datu-multzoak.

Euskararen egoera hau dela eta, euskararentzako hizkuntzen arteko transferentzi aztertu nahi izan da. Hau da, beste hizkuntza batzuetako datu-multzoak erabiltzen dituzten EE ereduak garatu dira eta ondoren euskaraz probatuak izan dira. Transferentziako hizkuntzei dagokionez, ondorengo 8-ak erabili dira: ingelesa, poloniera, gaztelera, portugesa,

japoniera, hindiera, koreera eta turkiera. Ebaluaketa hau aurrera eramateko euskarazko lehendabiziko gertaera-erauzketako ebaluaketarako datu-multzo bat bereziki sortu da (lan honetik kanpo).

Kasu honetan hizkuntzen transferentzia soila egitetik haratago joan nahi izan da, oinarri moduan erabili diren hizkuntzen eta euskararen arteko ezaugarri linguistikoen azterketa bat eginez. Azterketa horren helburua ezaugarriek transferentzian duten eragina ikustea izan da.

Egindako ataza guzti hauek azaltzeko memoria hau gauzatu da, kapitulu desberdinetan banandua. Lehenik, IE eta EE-ren inguruko hainbat kontzeptu azaltzen dira 3 kapituluan. Ondoren, erabili diren datu-multzoetan sakontzen da 4 kapituluan. Jarraian, 5 kapituluan, transferentzia egiteko erabili diren ereduak nola gauzatu diren ikusten da. Hauek lortzen dituzten emaitzak eta ezaugarrien linguistikoen azterketa 6 kapituluan ikusten dira. Eta azkenik, 7 kapituluan, proiektutik ateratako ondorioak eta etorkizunean egin daitezkeen hobekuntza posibleak jasotzen dira.

2. KAPITULUA

Proiektuaren Helburuen Dokumentua

Kapitulu honek lau atal nagusi izango ditu. Lehenengoan, eginiko proiektua deskribatu eta helburuak azalduko dira. Ondoren plangintza azalduko da, bide batez lanaren deskonposaketa eredua (LDE) ikusi eta egingo diren atazetan sakonduko da. Jarraian, aurreikusitako denboraren kudeaketa eta Gantt diagrama azalduko dira. Azkenik, arriskuen analisia burutuko da eta hauek ekiditeko prebentzio-plana aurkeztuko da.

2.1 Proiektuaren deskribapena eta helburuak

Proiektu honen helburu nagusia MEE datu-multzoan Veyseh et al. (2022) dauden hizkuntzen eta euskararen arteko transferentzi analisia egitea izan da informazio-erauzketa atazan. Gure lan nagusia, jatorrizko sistemaren emaitzetara hurbiltzen diren ereduak sortzea, eta eredu horiek ikasitako ezagutza euskarara eramatea izango da, euskarazko datu-multzoaren gainean informazio-erauzketa atazak burutu ahal izateko. Hizkuntza guztiak datu kopuru berdina izateko helburuarekin datu murrizketa batzuk ere gauzatuko dira, hizkuntza desberdinek lortzen dituzten emaitzak beraien artean konparagarriak izateko.

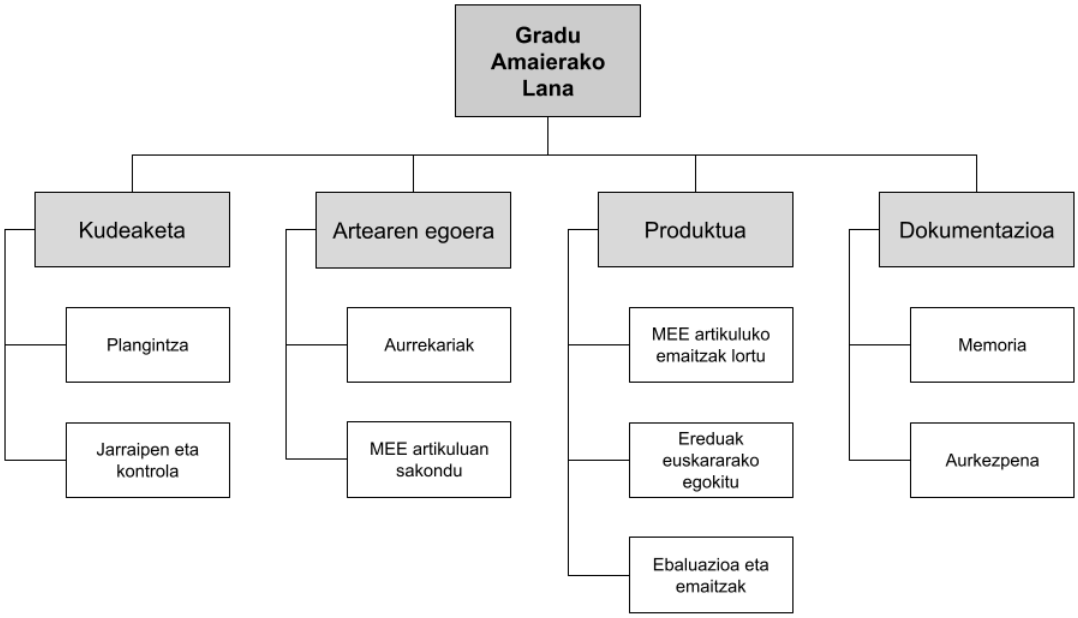
Lehenik, artearen egoera aztertuko da, erabiliko diren baliabide desberdinak azalduz. Ondoren, erabilitako datu multzoetan sakonduko da. Jarraian, ereduen sorkuntza eta hauek euskaran erabiltzeko burutuko diren pausuak azalduko dira. Azkenik, lortutako emaitzak aztertu eta ondorioak erauziko dira.

2.2 Plangintza

Azpiatal honetan proiektua garatzeko eginiko plangintza azalduko da. Proiektua egin ahal izateko faseak aurkezteaz gain, fase bakoitza burutzeko aurreikusitako denbora, eta egutegia ere ikusiko dira.

2.2.1 LDE diagrama

Proiektuan zehar egin beharreko lanaren deskonposaketa agertzen da 2.1 irudian, LDE diagrama batean.



2.1 Irudia: Lanaren Deskonposaketa Eredua.

2.2.2 Lan-paketeak

LDE diagraman agertzen diren lan-paketeen deskribapena egiten da azpiatal honetan.

Ikerkuntzako proiektuetan zaila izaten da ordu kopurua zehazki aurreikustea. Hori dela eta ordu kopurua ez da guztiz zehatza izango. Ondorioz, aurreikusitako ordu kopurutik asko urruntzen garela nabarituz gero, proiektuaren helburuetan gehiago edo gutxiago sakontzeko aukera egongo da, ahal den neurrian guztizko 300 orduen inguruan mantenduz.

Ataza bakoitzari emandako denborak 2.1 taulan ageri dira.

Kudeaketa

- **Plangintza:** Ataza honetan proiektuaren planifikazioa garatuko da; helburuak, atazak eta lan-metodologia definitzeaz gain, proiektuaren bideragarritasuna eta arriskuen analisia burutuko da.

- **Jarraipena eta kontrola:** Ataza honetan proiektuaren helburuak betetzen direla bermatuko da, zuzendariekin egingo diren bilerei esker.

Artearen egoera

- **Aurrekariak:** MEE artikuluarekin lanean hasi aurretik IE-ren inguruko kontzeptu nagusiak ulertu eta ikasiko dira.

- **MEE artikuluan sakondu:** IE-ren inguruko kontzeptuak garbi edukita MEE artikulua sakonean aztertuko da, bertan azaltzen den prozesua ondo ulertzeko.

Produktua

- **MEE artikuluko emaitzak lortu:** Ataza honetan MEE datu-multzoan oinarritzen diren ereduak sortzeko kodea egingo da. Ereduak, MEE artikuluko emaitzetara gerturatzeko intentzioarekin.

- **Ereduak euskararako egokitu:** Ereduak sortzeko kodea edukita, eredu berriak sortuko dira euskararekin neurketak egiteko egokituak.

- **Ebaluazioa eta emaitzak:** Ataza honetan esperimentuetan lortutako emaitzak aztertuko dira, hizkuntza desberdinen transferentzia neurtuz.

Dokumentazioa

- **Memoria:** Proiektuaren azalpenak eta lortutako emaitzen analisia biltzen dituen dokumentua garatuko da.
- **Aurkezpena:** Proiektuaren defentsarako aurkezpen bat prestatu beharko da, eginiko memoria laburtzen duen gardenki batzuk sortuz.

Lan-paketea	**Iraupena (orduak)**
Kudeaketa	**30**
Plangintza	10
Jarraipena eta kontrola	20
Artearen egoera	**25**
Aurrekariak	15
MEE artikuluan sakondu	10
Produktua	**145**
MEE artikuluko emaitzak lortu	70
Ereduak euskararako egokitu	60
Ebaluazioa eta emaitzak	15
Dokumentazioa	**100**
Memoria	80
Aurkezpena	20
Guztira	**300**

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

2.2.3 Emangarriak eta mugarriak

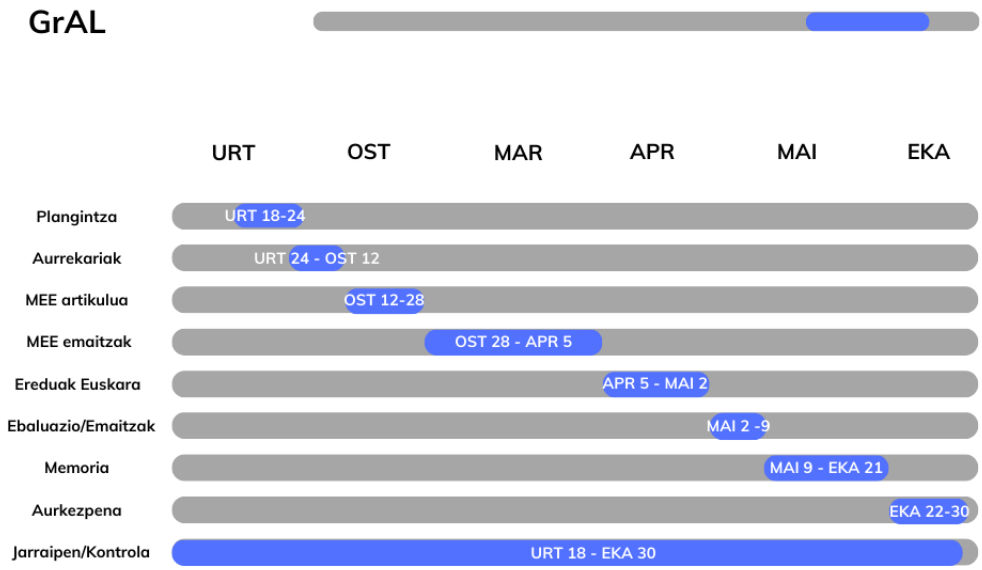
Proiektuan zehar garatu beharreko emangarrien mugarriak 2.2 taulan agertzen dira.

Emangarria	**Data**
Memoria	2023-06-25
Aurkezpena	2023-07-03

2.2 Taula: Proiektuko emangarrien datak.

2.2.4 Gantt diagrama

Proiektuko Gantt diagrama 2.2 irudian agertzen dena da.



2.2 Irudia: Proiektuko kronograma Gantt diagrama batean.

2.3 Lan Metodologia

Proiektua garatzeko etxetik lan egingo da, ordenagailu pertsonala erabiliz IXA-ko zerbitzarietara konektatzeko. Ez da ordutegi finkorik ezarriko eta beharrezkoak diren orduak modu egokienean banatuko dira. Garrantzitsuak izango dira ere burutzen ari diren ikasgaien lan karga kontuan hartzea, horiek izango baitira orduen lehentasuna izango dutenak.

2.3.1 Bilerak

Ikasle eta zuzendarien arteko komunikazioa bileren bidez egingo da. Bilera hauetan jarraipena eta kontrola burutuko da, hala nola garapenean zehar sortutako zalantzak argitu. Bilerak fakultatean bertan burutuko dira, ikaslearen beharren araberakoak izanik. Hauek normalean ostiraletan 9:30etan egingo dira.

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

2.4 Arriskuak eta prebentzioa

Proiektuan zehar egon daitezkeen arriskuak identifikatzen eta prebentzio-plana aurkezten dira atal honetan.

2.4.1 Zerbitzariarekin arazoak teknikoak egotea

- **Deskribapena:** Proiektuaren konputazio behar handiak direla eta IXA-ko zerbitzariak erabiliko dira exekutatzeko. Honek, arriskuak ekar ditzake: alde batetik, zerbitzariarekin lotutako zenbait arazo tekniko ekar ditzake; bestalde, exekutatu nahi den momentuan zerbitzariak beteak egotea posible da.
- **Prebentzioa:** Arrisku hau saihesteko, zerbitzariak erabiltzen ongi ikasteaz gain, denbora tarte batekin bidaliko dira gauzak exekutatzera. Modu honetara, zerbait gaizki ateraz gero ez gara epeetatik irtengo.

2.4.2 Euskarako datu-multzoa prest ez egotea

- **Deskribapena:** Euskarako datu-multzoa proiektuko lehen atazak egin ahala paraleloan prestatuko da. Hala ere, posible izango da datu-multzoa hau behar den eperako prest ez egotea.
- **Prebentzioa:** Hori gertatuz gero, hizkuntzen transferentzi analisia MEE datu-multzoko hizkuntza batekin egin beharko da.

2.4.3 Zailtasun maila handiegia

- **Deskribapena:** Mota honetako proiektuetan ez da erraza izaten proiektuak izango duen zailtasun maila aurreikustea. Hori dela eta, posible da proiektu honen atazaren bat gradu amaierako lan batek eskatzen duen ezagutza mailatik haratago geratzea.
- **Prebentzioa:** Jarraipen eta kontroleko bileretan zailtasunaren inguruko analisia burutuko da proiektuaren lehen asteetan. Zailtasun maila handiegia dela ikusten bada, helburuak egokitzeko edota irismena aldatzeko aukera egongo da.

3. KAPITULUA

Artearen Egoera

Kapitulu honek hiru atal nagusi izango ditu: lehenengoan, Informazio-Erauzketaren (IE) inguruan sakonduko da. Ondoren, IE ebaluatzeko metodoak ikusiko dira. Azkenik, hirugarren atalean, hizkuntzaren prozesamenduan erabiltzen diren ikasketa sakoneko metodoak azalduko dira.

3.1 Informazio-Erauzketa

IE edo Informazio-Erauzketa (Information Extraction ingelesez) hizkuntzaren prozesamenduko arloan gai garrantzitsuetako bat da. Honen helburua testu baten informazioa datu-mota egituratuetara bihurtzea da.

Testua gizakiontzako informazioa elkarbanatzeko modu naturalenetako bat den arren, konputagailuentzat zaila da bertatik informazio zuzenean lortzea. Izan ere, testua ez da datu-mota egituratu bat. Hori, hainbat arrazoiengatik gertatzen da: anbiguotasuna, hizkuntzaren aldakortasuna, etab. Aipatutako arrazoi horiek dira IE zaila bilakatzen dutenak.

IE-k zenbait azpi-ataza ditu. Azpi-ataza hauek bilatu nahi den informazioaren araberakoak dira. Jarraian horietako batzuk ikusiko ditugu.

3.1.1 Izendun Entitateen Erauzketa

Izendun Entitateen Erauzketa edo NER (Named Entity Recognition ingelesez), IE-ren azpi-ataza bat da. Honen helburua testu batean dauden entitateak, hau da, munduan existitzen diren pertsonak, erakundeak, kokalekuak, etab. Domeinuaren arabera aurredefinitutako kategorietan sailkatzea da. NER sekuentzia-etiketatze ataza bat bezala planteatu ohi da.

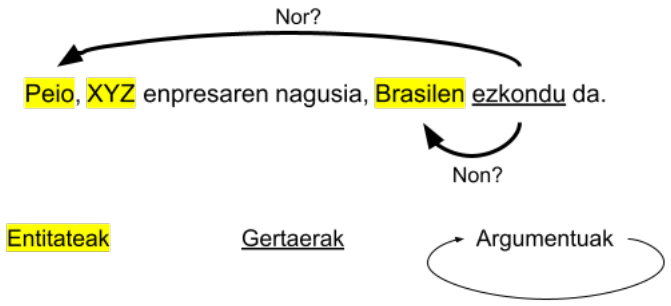
[Tim Cook]$_{\text{PER}}$ is the CEO of [Apple]$_{\text{ORG}}$

3.1 Irudia: NER-en adibide bat.

Adibide moduan 3.1 irudian definitutako kategoriak pertsonak eta erakundeak dira.

3.1.2 Gertaera Erauzketa

IE-ren ataza garrantzitsuenetako bat Gertaera Erauzketa edo EE (Event Extraction ingelesez) da, zeinak testu batetik automatikoki gertaerak eta gertaera horien argumentuak lortzea du helburu. Detektatu nahi diren gertaerak aurredefinituta egongo dira, adibidez: jaio, hil, ezkondu etab. Hau hobeto ulertzeko, adibide moduan 3.2 irudia sortu da, bertan Gertaera Erauzketaren adibide bat azaltzen da entitate, gertaera eta argumentuak detektatuz.



3.2 Irudia: Gertaera Erauzketaren adibide bat.

Gaur egun EE hainbat arlotan garrantzitsua izaten hasi da, esaterako: segurtasun kontuetarako, hauetan erakunde baten eta pertsona baten arteko erlazioak jakitea interesgarria da; analisi geoespaziala egiteko, kasu horretan garrantzizkoa da jakitea gertaera bat non gertatzen hari den, etab.

Gertaera-erauzketa ataza, normalean, hiru azpi-atazetan banatzen da: Entitate Aipamen Detekzioa, Gertaeren Aipamen Detekzioa eta Gertaeren Argumentuen Erauzketa. Ondorengo azpi-ataletan hauen inguruan sakonduko dugu.

Entitate Aipamen Detekzioa

Entitate Aipamen Detekzioa (Entity Mention Detection ingelesez) edo EMD bere ingeleseko siglengatik, esaldi batean dauden entitateak detektatzea du helburu, NER ataza baten modura definitu daiteke.

Adibidez, 3.2 irudian "Peio" (pertsona), "XYZ" (erakundea) eta "Brasilen" (lekua) detektatuko lituzke.

Gertaeren Aipamen Detekzioa

Gertaeren Aipamen Detekzioa (Event Detection ingelesez) edo ED atazak gertaerak detektatu eta gertaera horiek zein motatakoak diren sailkatzen ditu. Aurreko azpi-atazaren antzera, NER ataza baten modura definitu daiteke. Detektatutako gertaerak hurrengo ataza gauzatzeko baliatu daitezke.

Aurreko adibide berdinarekin jarraituz, detektatuko lukeen gertaera "ezkondu" (ezkontzea) izango litzateke.

Gertaeren Argumentuen Erauzketa

Gertaeren Argumentuen Erauzketa (Event Argument Extraction ingelez) edo EAE atazak gertaera bat emanik, gertaera horren argumentuak detektatzea du helburu. Gertaera bi modu nagusitan lortu daiteke: aurreko atazarekin lotuz kate baten moduan edo guk emanez.

Adibidea mantenduz, detektatuko lituzkeen argumentuak "Peio" (pertsona) eta "Brasilen" (lekua) izango lirateke.

3.1.3 EE-ren egoera

EE egiteko modua aldatzen joan da urteetan zehar: lehen lanek, ezaugarrietan oinarritutako ereduak erabiltzen zituzten (Ahn, 2006; Ji and Grishman, 2008). Bestalde, aurrerago,

ikasketa sakoneko (*Deep Learning* ingelesez) teknikan oinarritutako ereduak erabiltzen hasi ziren (Nguyen et al., 2016; Sha et al., 2018) gaur egungo artearen-egoera izanik.

Hala ere, naiz eta aurrera pausu handiak eman diren azken urteetan EE-ren ikerkuntzan, muga nagusietako batek berdin jarraitzen du: hizkuntza handi gutxi batzuentzat garatzen da EE ia osoa. Ondorioz, hizkuntza txikiago batzuk bigarren maila batean geratzen dira; bertan eredu hauek ez baitute hain ondo funtzionatzen hizkuntza handietan bezala. Honen arrazoi nagusia hizkuntza txiki horietan egindako kalitatezko EE datu-multzoen falta izaten da. Esaterako, EE atazaren datu-multzo nagusienetariko bat ACE 2005 (Walker et al., 2006), bakarrik hiru hizkuntza desberdinetan dago erabilgai: ingelesez, txineraz eta arabieraz.

3.1.4 Hizkuntzen arteko transferentzia

Aurreko atalean aipatu bezala, hizkuntza asko ez daude EE egiteko datu-multzoetan. Hori dela eta beste bide batzuk aztertu behar dira horietan EE egiteko. Bide horietako bat hizkuntzen arteko transferentziaz baliatzea da.

Hizkuntzen arteko transferentzia egitea hizkuntza batean entrenatu eta beste batean ebaluatzeari deritzo, hizkuntza baten ikasitakoa beste hizkuntza batera aplikatuz. Entrenamenduko datu-multzoak askoz ere handiagoak izaten direnez testekoak baino, posible da entrenamenduan ez dauden hizkuntzentzat datu-multzo txiki batzuk sortzea test moduan funtzionatuko dutenak.

Egoera hau ikusita, gaur egun abiatu dira zenbait lan honen inguruan ikertzeko helburuarekin, esaterako Dolicki and Spanakis (2021) lanean, hizkuntza arteko transferentzian ezaugarri linguistikoek EE-ko ataza desberdinetan duten eragina aztertzen da.

3.2 Ebaluazioa

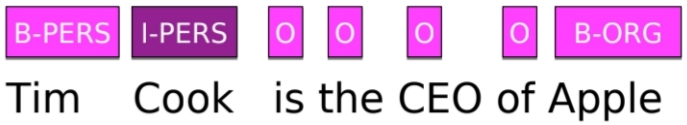
IE-k duen zailtasuna dela eta, ez dira izaten sistema perfektuak. Horren ondorioz, sistema hauen funtzionamendua ebaluatzea ezinbestekoa da. Ebaluazio horretan zentratuko gara atal honetan.

3.2.1 Anotazio formatua (BIO)

Erauzitako datuak etiketa konkretu batekin anotatuta izateak IE sistemen ebaluazioa errazten du. Etiketa horiek errepresentatzeko BIO edo Beginning-Inside-Outside formatua erabili ohi da.

BIO formatuan token (hitz) bakoitza etiketatzen da 3 etiketa posiblerekin:

- O: Ez da detektatu nahi den zerbait
- B-*KATEGORIA*: Detekzio baten hasiera eta kategoria adierazten ditu. Adibidez, "Apple" entitatea baldin badugu B-ORG etiketa hartuko du.
- I-*KATEGORIA*: Detekzio baten jarraipena eta kategoria adierazten ditu. Etiketa hau markatu nahi dugun entitatea hitz batez baino gehiagoz osatuta dagoenean erabiltzen da. Esaterako entitatea "Tim Cook" bada, bere etiketak B-PERSON, I-PERSON izango dira.



3.3 Irudia: BIO formatuaren adibide bat.

Formatu honekin esaldi bateko hitz guztiak etiketatzen dira 3.3 irudian ikus daiteken bezala.

3.2.2 Ebaluazio-neurriak

EE egiten duen eredu bat ebaluatzeko neurri ezberdinak daude. Erabiltzen den neurrietako bat asmatze-tasa da.

Asmatze-tasa

Asmatze-tasa zuzen iragarri diren elementu kopuruaren eta guztira dauden elementu kopuruaren arteko zatiketa da. Hau da:

$$\text{Asmatze-tasa} = \frac{\text{Zuzenak}}{\text{Totalak}}$$

Neurri hau ez da oso esanguratsua Informazio Erauzketa atazetan. Izan ere, elementu gehienak O BIO etiketa dute, eta beraz eredu batek beti O iragartzen badu asmatze-tasa altua izango du; baina urrun egongo da eredu egoki bat izatetik. Hori dela eta, beste ebaluatze neurri bat erabiltzen dugu: F1.

F1 neurria

F1 ulertzeko lehenengo bi kontzeptu azaldu behar dira: doitasuna eta estaldura. Doitasuna (precision ingelesez) hautemandako elementu guztien artean zuzen hautemandako elementuen proportzioa da, eta estaldura (recall ingelesez) hauteman beharreko elementu guztietatik hautemandako elementuen proportzioa da. Honen adibide bat 3.4 irudian ikus daiteke. Doitasuna eta estalduraren formulak hurrengoak dira:

$$\text{Doitasuna} = \frac{\text{PositiboZuzenak}}{\text{PositiboZuzenak} + \text{PositiboOkerrak}}$$

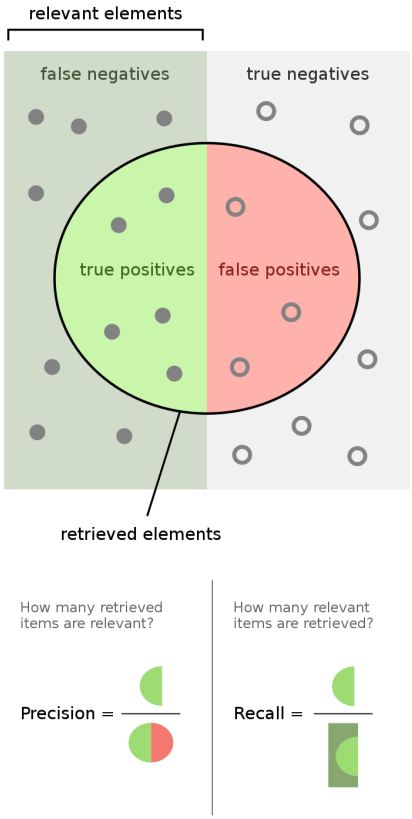
$$\text{Estaldura} = \frac{\text{PositiboZuzenak}}{\text{PositiboZuzenak} + \text{NegatiboOkerrak}}$$

Doitasuna eta estalduraren balioetan oinarritzen da F1 neurria, 0-tik 1-era doan balio bat bueltatuz. Zehazki, doitasuna eta estalduraren bataz-besteko harmonikoa da F1. Hurrengo formula betetzen du:

$$\text{F1} = 2 * \frac{\text{Doitasuna} * \text{Estaldura}}{\text{Doitasuna} + \text{Estaldura}}$$

Horretaz gain, posible da klase batzuk F1 puntuazioaren kalkulutik baztertzea. Gure kasuan, interesgarria da O klasea alde batera uztea. Modu honetara, asmatze-tasarekin dugun arazo nagusia konpondu dezakegu. Bestalde, klase anitzeko sailkapen bat badugu (gure kasuan bezala), klase bakoitzeko F1 kalkulatu beharrean F1 orokorra kalkulatzea interesgarriagoa da. Hau lortzeko, mikro bataz-bestekoa [1] erabiltzen dugu. Mikro bataz-bestekoak F1 globala kalkulatzen du, horretarako positibo zuzen, negatibo oker eta positibo okerren guztizkoa zenbatzen du eta kopuru horrekin kalkulatzen ditu doitasuna eta estaldura orokorrak.

[1] Mikro bataz-bestekoaren eta orokorrean F1-aren inguruko informazio gehiago: towards data science.



3.4 Irudia: Doitasuna (precision) eta estaldura (recall) azaltzeko adibide bat.

Azkenik, aipagarria da, F1 puntuazioaren kalkuluan doitasunak eta estaldurak 0 balioak hartzen badituzte, $\frac{0}{0}$ moduko indeterminazio bat gertatuko dela. Gure kasuan, modu horretako indeterminazioak, 0 balira bezala tratatuko ditugu.

3.3 Ikasketa Sakona

Ikasketa sakona azken urteetan indarra hartzen joan den adimen artifizialeko diziplina da. Gaur egun, hizkuntzaren prozesamenduan gehien erabiltzen den metodoa izanik.

Hizkuntza prozesatzeko ikasketa sakonen metodoen barnean hainbat aldaketa egon dira urteetan zehar: Lehenik, MLP-ak erabiltzen hasi ziren. Baina hauek ez zuten lortzen hizkuntza ondo modelatzea. Hori dela eta RNN-ak edo "Recurren Neural Network" (David E. Rumelhart and Williams, 1985) sortu ziren. RNN-en hizkuntza prozesatzeko ahalmena MLP-ena baina handiagoa izanda ere, arazo nagusi bat zuten: esaldi baten hasieran zeuden tokenak (hitzak) 'ahaztu' egiten zituzten. Hori dela eta, RNN-en hobekuntza bat sortu zen LSTM (Hochreiter and Schmidhuber, 1997) deitua, honek arazo hori hobetzen lagundu zuen. Hobekuntzekin jarraitzeko *atentzioa* deituriko mekanismo bat LSTM-etara sartzen hasi zen eta pixkanaka gero eta indar gehiago lortu zuen, Transformerren iritsiera gauzatuz.

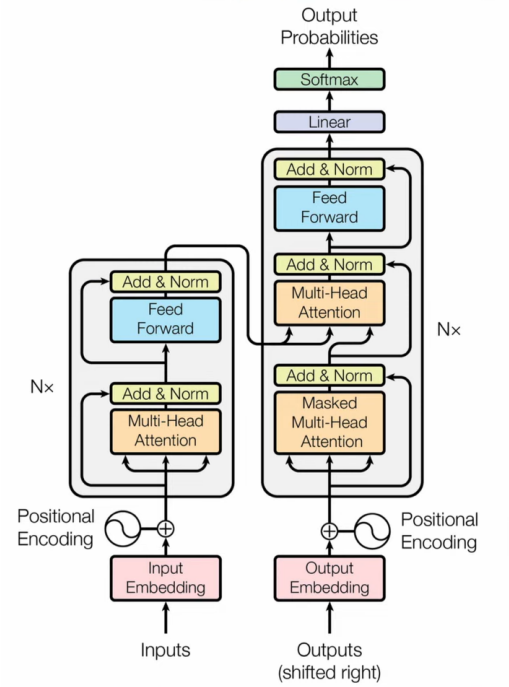
3.3.1 Transformerrak

Gaur egun hizkuntzaren prozesamenduaren arloan gehien erabiltzen diren ereduak Transformerrak dira. 2017-an publikatutako Vaswani et al. (2017) lanean aurkeztu eta gero hizkuntzaren prozesamenduko artearen egoera guztiz aldatu zuten (Min et al., 2021). Hauek, RNN-ak ez bezala ez dituzte tokenak sekuentzialki prozesatzen, hau da, esaldi bat prozesatu nahi bada, amaierako tokena prozesatzeko ez litzateke beharrezkoa izango hasierakoa prozesatu izana. Honek paralelizazio ahalmen handia ematen die, eredu oso handiak entrenatu ahal izateko ezinbestekoa.

Transformerren oinarri nagusia bere buruarekiko *atentzioa* da. Arkitektura honek, LSTM-ak gainditu zituen hasierako informazioaren galera oraindik ere gehiago txikituz. *Atenzioa* erabiliz sekuentzia bateko elementu garrantzitsuenak identifikatzen dira, informazio hori sekuentzia luzeagoetan mantenduz.

Atentzio egitura hau hainbat geruza eta blokeetan errepikatzen da. Sakontasun honek hizkuntzaren hainbat aspektu modelatzeko gaitasuna ematen dio.

Transformerrak 3.5 irudian ikus daitekeen kodetzaile-dekodetzaile arkitektura erabiltzen dute. Arkitektura hau erabiltzen duten hainbat eta hainbat eredu daude, esaterako: BERT (Devlin et al., 2019), GPT (Alec Radford and Sutskever, 2018) eta T5 (Raffel et al., 2020),



3.5 Irudia: Transformerraren arkitektura. Vaswani et al. (2017).

besteak beste. Kasu honetan BERT ereduan jarriko dugu arreta, hori baita gure lanerako interesgarriena.

3.3.2 BERT

BERT edo "Bidirectional Encoder Representations from Transformers" (Devlin et al., 2019) Transformerren kodetzailearen arkitekturan oinarritzen den hizkuntza eredua da. Honek, bi norabidezko atentzio mekanismoa erabiltzen du. Hau da, eskuin eta ezkerreko informazioa edukitzen du kontuan. Eredu hau datu askorekin aurre-entrenatzen da ataza jakin baterako entrenamendu espezifikoa eduki aurretik. Atal honetan BERT ereduan pixka bat sakonduko dugu, lan honetan erabiliko dugun eredua honetan oinarritzen baita.

Aurre-Entrenamendua

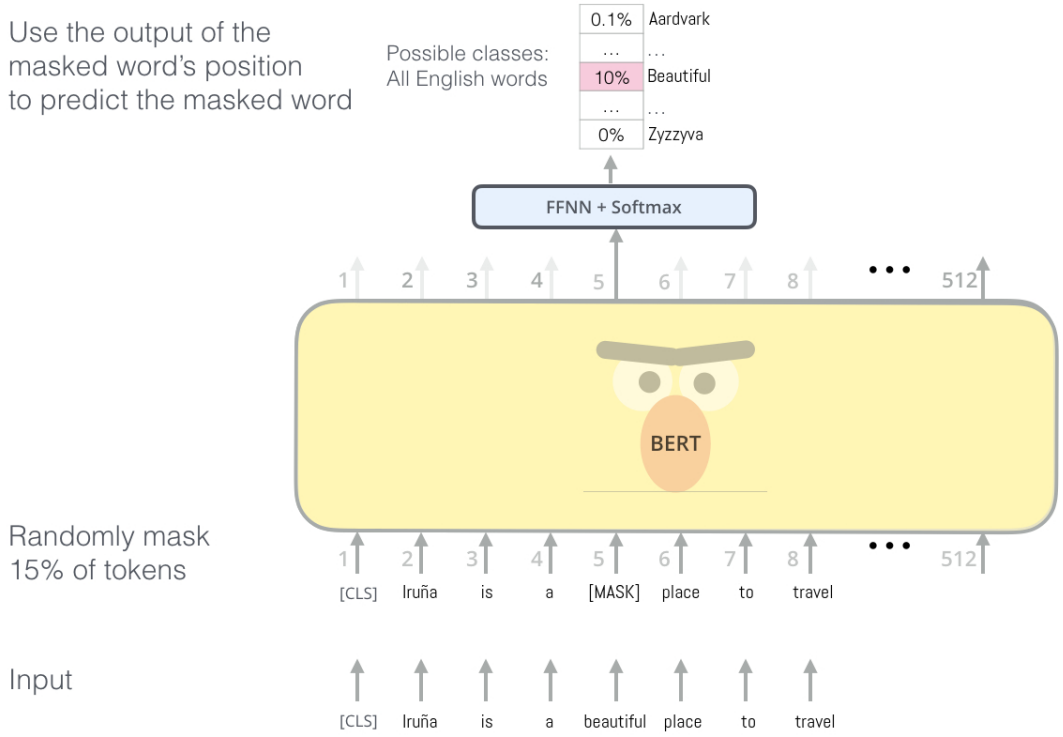
BERT bi norabideko atentzio mekanismoetan oinarritzen denez, eskuineko zein ezkerreko informazioa dauka. Hori dela eta, ez dauka zentzurik hurrengo tokena zein den iragartzeko entrenamendua egitea. Horrenbestez, autoreek gainbegiratu gabeko bi iragarpen ataza proposatzen dituzte entrenamendua gauzatzeko: "Masked Language Model" eta "Next Sentence Prediction".

Masked Language Modeling atazan, sarrerako tokenen %15 ezkutatzen dira, eta sistemak jatorrizko tokena zein zen iragarri behar du. Tokenak ezkutatzeko [MASK] tokena erabiltzen da. Hala ere, ez da beti aplikatzen [MASK] tokena ezkutatu nahi den tokenean:

- Kasuen %80-an [MASK] tokena erabiltzen da. Adibidez "Iruña is a beautiful place to travel" "Iruña is a [MASK] place to travel" bihurtzen da.
- Kasuen %10-ean ausazko hitz batekin ordezten da: "Iruña is a apple place to travel".
- Kasuen %10-ean ez da hitza aldatzen.

Prozedura honen abantaila da BERT-ek ez duela soilik [MASK] sarrerako hitzak iragartzen ikasten, baizik eta hitz guztiak iragartzen saiatuko da. Hori dela eta, sarrerako token guztien testuinguruaren menpeko errepresentazioak mantentzera behartuta geratzen da. Gainera, hitzen ordezkapena token guztien %1.5-etan (hau da, % 15aren %10-a) gertatzen da soilik eta horrek ez dirudi kalte egiten dionik ereduak hizkuntza ulertzeko duen gaitasunari.

Ataza honen adibide bat ikus daiteke 3.6 irudian.

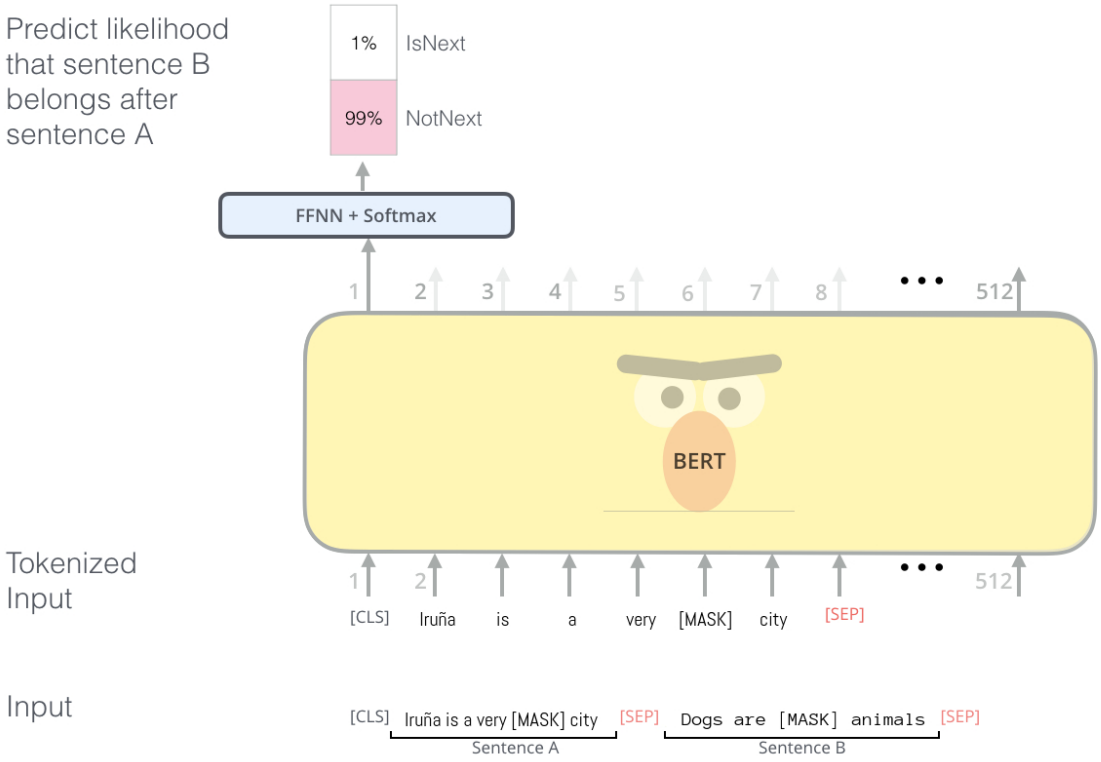


3.6 Irudia: Masked Language Model-en adibide bat

Next Sentece Prediction atazan esaldien arteko erlazioa ikastea da helburua. Hau lortzeko bi esaldi aukeratzen dira corpusetik: Kasuen %50-ean jarraian doazen esaldiak aukeratzen dira, gainontzeko kasuetan elkarrekin erlaziorik ez duten ausazko bi esaldi aukeratzen dira. Sistemak, jasotako esaldiak elkarren jarraian doazen edo ez iragarri behar du. Adibidez:

```
Input = [CLS] Iruña is a very [MASK] city [SEP] Dogs are [MASK] animals [SEP]
Label = NotNext
Input = [CLS] Iruña is a very [MASK] city [SEP] You should visit it [SEP]
Label = IsNext
```

Ataza honen adibide bat ikus daiteke 3.7 irudian.



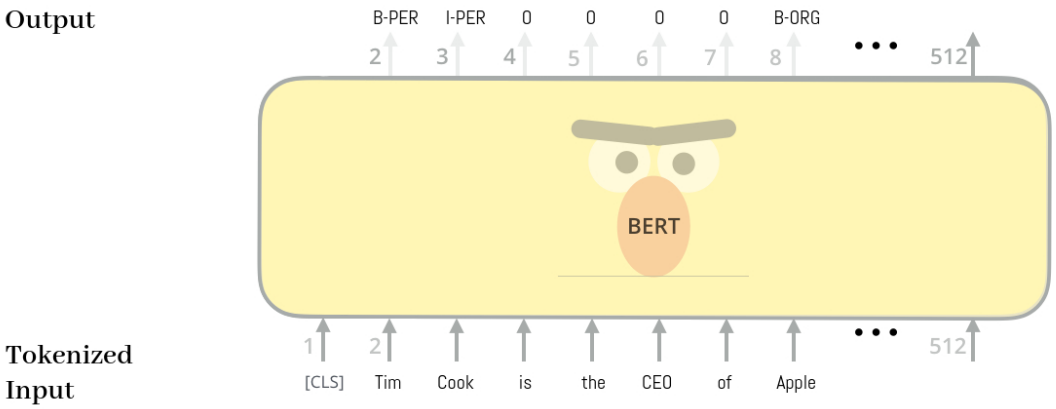
3.7 Irudia: Next Sentece Prediction-en adibide bat

Birdoiketa

Behin BERT eredua aurre-entrenatuta daukagula, nahi dugun atazarako entrenamendu espezifikoa egiten da. Bigarren entrenamendu honi birdoiketa edo “fine-tuning” deritzo.

Lan honen kasuan, BERT-ek sekuentzia etiketatze ataza egiten ikasteko egingo da fine-tuning-a. Hau da, BERT-ek token (hitz) bakoitzari BIO etiketa bat eman beharko dio. Adibidez:

```
Input = Tim Cook is the CEO of Apple
Outpout = B-PERS I-PERS O O O O B-ORG
```



3.8 Irudia: BERT NER egiten

Honen adibide bat 3.8 irudian ikus daiteke.

3.3.3 Hizkuntza-Eredu Eleanitzak

XLM-RoBERTa (Conneau et al., 2020) BERT-en antzera, Transformer kodetzaile arkitekturan oinarritzen den eredua da. Baina BERT-ekin alderatuz entrenamendu korpus handiago ikusi du eta hainbat hizkuntzetan entrenatuta dago. Azken ezaugarri hau dela eta, hizkuntza-eredu eleanitza da.

Ondorengoak dira XLM-roBERTa-ren ezaugarri zehatzak:

- 2.5TB testu datuekin entrenatuta.
- 100 hizkuntza desberdinetan entrenatuta.
- 125M parametro base bertsioan.

Eredu honen base bertsioa da erabiliko duguna gure lana gauzatzeko.

4. KAPITULUA

Datu-multzoak

Lan hau gauzatu ahal izateko bi datu-multzo erabili dira: MEE eta EusIE. Atal honetan bi datu-multzo hauen inguruan sakonduko da.

4.1 MEE datu-multzoa

MEE datu-multzoa (Veyseh et al., 2022) erabili da lan honetako ereduak entrenatzeko. Ondorengo atalean datu-multzo honen inguruan gehiago sakonduko da.

4.1.1 Ezaugarriak

MEE datu-multzoa 8 hizkuntzaz osatzen da: ingelesa, japoniera, portugesa, turkiera, gaztelera, poloniera, koreera eta hindiera. Hizkuntza horietako bakoitzarentzat hiru banaketa daude: entrenamendukoa, garapenekoa eta testekoa. Horretaz gain, banaketa hauetan dauden datuak aztertzen baditugu, segmentuetan banatuak daudela ikus dezakegu. Segmentu bat 5 esaldiz osatutako testua da. Segmentu horietako bakoitzarentzat EE ereduak sortu ahal izateko anotazioak daude, ataletan banatuak:

- *tokens*: Segmentuko hitzak tokenizatuta.
- *entities*: EMD eredua entrenatzeko datuak.
- *triggers*: ED eredua entrenatzeko datuak.

- *arguments*: EAE eredua entrenatzeko datuak.

EMD, ED eta EA 3.1.2 atalean aipatutako EE-ren atazak izanik. Ataza hauetan lan egin ahal izateko, anotatu beharrekoa zehaztua egon behar da ataza bakoitzeko: alde batetik, *entities*-eko entitatei dagokionez, ACE 2005 (Walker et al., 2006) datu-multzoan dauden berdinak dira; 4.1 taulan ikus daitezke.

Mota	**Azalpena**	**Adibidea**
PERSON	Giza entitateak	Tim Cook
ORGANIZATION	Enpresa, agentzia eta beste pertsona talde batzuk	Apple
GPE	Talde politikoek edota gizarte-taldeek definitutako eskualde geografikoak	Brasil
LOCATION	Entitate geografikoak, hala nola lur-masak edo ur-masak	Zelai
FACILITY	Gizakiak eraikitako eraikin eta beste egitura iraunkor batzuk	Suezko kanala
VEHICLE	Objektu bat leku batetik bestera mugitzeko diseinatutako gailu fisikoak	Tren
WEAPON	Min fisikoa egiteko tresna gisa erabiltzen diren gailu fisikoak	Lehergailu

4.1 Taula: Entitate motak azalpen eta adibideekin

Bestalde, *triggers*-eri dagokionez, gertaera motak daude, ACE 2005-eko gertaera mota batzuk soilik erabiltzen ditu (hizkuntza artean anbiguoak ez direnak). Azkenik, *arguments*-ri dagokionez, argumentuen rolak daude, ACE 2005-eko 23 rolak mantentzen ditu. *Triggers* eta *arguments*-eko gertaera motak eta argumentuak 4.2 taulan ikus daitezke. 4 atal hauen adibide bat 4.1 irudian ikus daiteke.

ID	**Gertaerak**	**Argumentuak**
1	Life_Be-Born	Person, Time, Place
2	Life_Marry	Person, Time, Place
3	Life_Divorce	Person, Time, Place
4	Life_Injure	Agent, Victim, Instrument, Time, Place
5	Life_Die	Agent, Victim, Instrument, Time, Place
6	Movement_Transport	Agent, Artifact, Vehicle, Price, Origin, Destination, Time
7	Transaction_Transfer-Ownership	Buyer, Seller, Beneficiary, Price, Artifact, Time, Place
8	Transaction_Transfer-Money	Giver, Recipient, Beneficiary, Money, Time, Place
9	Business_Start-Organization	Agent, Organization, Time, Place
10	Conflict_Attack	Attacker, Target, Instrument, Entity, Time, Place
11	Contact_Meet	Entity, Time, Place
12	Contact_Phone-Write	Entity, Time
13	Personnel_Start-Position	Person, Entity, Position, Time, Place
14	Personnel_End-Position	Person, Entity, Position, Time, Place
15	Justice_Arrest-Jail	Person, Agent, Crime, Time, Place

4.2 Taula: Gertaera motak eta gertaera horietan egon daitezken argumentuen rolak

Aipatzekoa da ere hizkuntza guztietan ez dagoela datu kopuru berdina. Hori dela eta, hizkuntza bakoitzaren datu kopuruak jasotzen dituen 4.3 taula egin da. Datu hauen %80-a entrenamendu multzoan sartuko da eta beste bi %10-ak garatze eta test multzoen artean

banatuko dira. Taula honetan aipagarria da hizkuntza askok dituzten argumentu kopuru urria. Izan ere, segmentu askok ez dituzte argumentuak anotatuak.

Hizkuntza	#Seg.	Bataz-Besteko Luzera	#Entitateak	#Gertaerak	#Argumentuak
Ingelesa	13,000	123	190,592	17,642	13,548
Gaztelera	3,268	112	48,001	6,064	802
Portugesa	1,500	102	25,463	1,953	12,329
Poloniera	4,479	108	62,971	10,875	3,395
Turkiera	4,480	117	38,469	8,390	1,416
Hindiera	1,499	98	18,797	1,810	2,117
Japoniera	1,500	99	19,174	2,152	3,399
Koreera	1,500	103	12,508	1,125	1,742
Guztira	31,226	-	415,975	50,011	38,748

4.3 Taula: MEE datu-multzoko datuak, hizkuntzaka banatuak

#Seg zutabeak segmentu kopurua adierazten du eta bataz-besteko luzerak zutabeak segmentu bakoitzak bataz-bestean dauzkan token kopurua. Beste zutabeek dagokion atazaren datu kopurua adierazten dute.

Azkenik, segmentu guztiak Wikipediatik atera dira, zehazki ondorengo gaiak erabiliz: ekonomia, politika, teknologia, delinkuentzia, natura, eta militarrak. Modu honetara, datuak publikoak eta anitzak izatea bermatzen da.

4.2 EusIE datu-multzoa

EusIE EE egiteko euskarako datu-multzo bat da. Honek, MEE datu-multzoaren antzeko ezaugarriak ditu baina soilik garapeneko eta testeko banaketekin. Izan ere, 295 segmentu bakarrik ditu. Hau da, datu-multzo honekin ezin dira entrenamenduak egin, asko jota beste datu-multzo batekin egindako entrenamenduak findu daitezke (garapenerako datuak erabiliz). Beraz, datu-multzo honen helburu nagusia ebaluazioak egitea da.

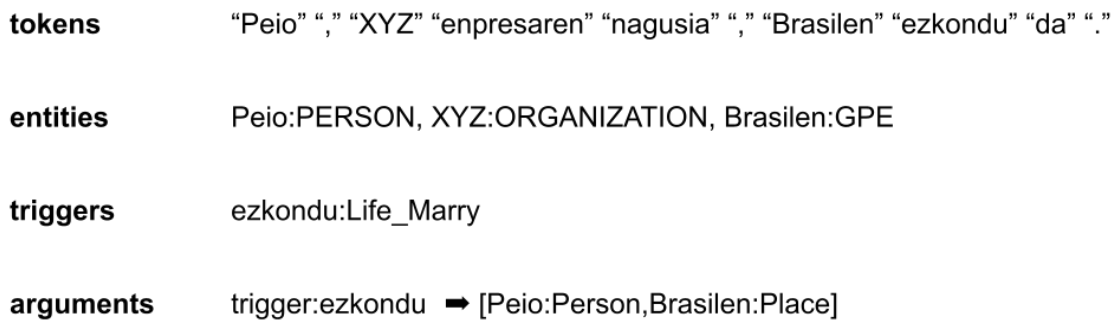
Datu-multzo hau erabiliko da MEE datu-multzoko beste hizkuntzen eta euskararen artean transferentzia egiteko.

4.2.1 Ezaugarriak

Aurretik aipatu bezala euskarako datu-multzoa oso urria da segmentu kopuruan. Hala ere, segmentu kopuru hori edukitzeko, jasotzen dituen datu kopurua ez da batere txikia, bereziki aipagarria izanik dituen argumentu kopurua. Honen datuak 4.4 taulan ageri dira. Bertako segmentuak lortzeko MEE datu-multzoaren berdina egin da: Wikipediatik atera.

Hizkuntza	#Seg.	Bataz-Besteko Luzera	#Entitateak	#Gertaerak	#Argumentuak
Euskara	295	94	4867	642	1326

4.4 Taula: Euskarako datu-multzoaren datuak



4.1 Irudia: MEE eta EusIE datu-multzoek dituzten anotazio atalen adibide bat.

5. KAPITULUA

Ingurune Esperimentala

Lan honetan gertaera-erauzketa ataza egiten duten ereduak sortu dira, MEE datu-multzoko 8 hizkuntza desberdinekin (ingelesa, japoniera, portugesa, turkiera, gaztelera, poloniera, koreera eta hindiera). Honen helburua, hizkuntza horiek eta euskararen arteko transferentzia egitea izan da, hizkuntzen ezaugarriak nola eragiten duten ikusteko ezagutza-transferentziari. Hori egin ahal izateko proiektua 3 fasetan banatu da: Oinarri-lerroa, Datu kopurua berdintzea eta Hizkuntzen arteko transferentzia euskaraz. Erabilitako kodeari dagokionez ondorengo GitHub-eko estekan dago.

5.1 Oinarri-lerroa

Lehen fase honetan helburu nagusia MEE artikuluko (Veyseh et al., 2022) ereduak inplementatzea izan da.

5.1.1 Datuak prestatzea

Lehen lana MEE eta EusIE datu-multzoetako datuak prestatzea izan da. Alde batetik, datuak BIO formatura bihurtu dira ereduak entrenatu ahal izateko. Bestalde, argumentuen kasuan, ereduei gertaera markatzea beharrezkoa da ereduak jakin dezan konkretuki zein gertaeren argumentuak lortu behar dituen. Horretarako, markaketa hori ere gauzatu behar da.

Datu-multzoak BIO formatuan

MEE eta EusIE datu-multzoen jatorrizko formatua ez da BIO. Hori dela eta, BIO formatura bihurketa bat egin da. Honen helburua, besteak beste, entitateen eta gertaeren atalak jatorrizko formatutik BIO formatura bihurtzea izan da, 4.1.1 atalean azaldutako atalak mantenduz, baina kasu honetan BIO formatura egokituak.

Argumentuak

Datuak prestatzeko garaian izandako beste ataza garrantzitsuetako bat gertaeraren markaketa izan da. Markaketa honen bidez jakinarazi diezaiokegu ereduari zein gertaeren argumentuak bilatu behar dituen.

Kontuan izanik ereduen sarrera 3.3.2 atalean ikusitakoaren antzekoa dela, markaketa sarrera moduan sartzen diren tokenetan egon beharko du. Hori dela eta, markatzeko modua gertaera token berezien (kasu honetan $$$ ikurren) artean inguratzea izan da. Informazio berri hau jasotzeko fitxategi berriak sortu dira, bat banaketa bakoitzeko. Hau da, lehen entrenamenduko, garapeneko eta testeko fitxategiak bagenitu, orain hauetako bakoitzarentzat fitxategi berri bat egongo da. Fitxategi hauetako lerro bakoitzean gertaera bakar bat egongo da markatua, gertaera horretako argumentuak BIO formatuan etiketatuta daudelarik. Gertaeraren markaketako adibide bat 5.1 irudian ikus daiteke.

Entrenamendua (tokenak)	“Peio” “eta” “Maria” “Brasilen" “ezkondu” “dira”
Entrenamendua_Arg (tokenak)	“Peio” “eta” “Maria” “Brasilen" “$$$” “ezkondu” “$$$” “dira”
Entrenamendua (gertaerak)	“O” “O” “O” “O" “B-Life_Marry” “O”
Entrenamendua_Arg (argumentuak)	“B-Person” “O” “B-Person” “B-Place" “O” “O” “O” “O”

5.1 Irudia: Argumentuen fitxategiko gertaeraren markaketako adibide bat.

Esaldi batek gertaera bat baina gehiago baditu orduan gertaera adina aldiz azalduko da esaldi hori fitxategi berrian, bakoitzean gertaera desberdin bat markatua duelarik.

5.1.2 Ereduak

Ereduak ataza bakar baterako entrenatu dira. Hori dela eta, 3.1.2 atalean ikusi ditugun EE-ren 3 azpi atazak (EMD, ED eta EAE) kontuan izanik, bakoitzarentzako eredu bat entrenatu beharko da. Hori, MEE datu-multzoko 8 hizkuntzetarako egin beharko da. Hau da, 3 ataza eta 8 hizkuntza daudenez guztira 24 XLM-RoBERTa eredu desberdin entrenatu beharko dira. Horretaz gain, beste hizkuntza bat bazen moduan, *denak* eredua ere entrenatu da. Hau, hizkuntza guztien datuekin entrenatutako EE sailkatzaile orokorra izango da. Beraz, 3 ataza eta 9 hizkuntza (aurretik aipatutakoak eta *denak*) guztira 27 eredu desberdin izango dira entrenatu beharrekoak.

5.1.3 Entrenamendua

Entrenamendua gauzatzeko ahal izateko bi elementu nagusi behar izan dira: alde batetik, eredu horiek entrenatzeko gai den hardwarea. Eta bestalde, entrenamendu script-a, hardware hori erabiliz ereduak aurreprozesatutako datuekin entrenatzen jarriko duena.

Hardwarea

Eredu hauen tamaina dela eta, entrenamendua gauzatu ahal izateko ahalmen handiko GPU-ak behar dira. Hori dela eta, IXA taldeko zerbitzariak erabiltzea erabaki da entrenamendua egin ahal izateko. Nahiz eta bertako hainbat zerbitzari erabili diren, gehienen erabili dena *xirimiri* izeneko da; zeina, 2 Nvidia Tesla A30 GPU-z osatzen da.

Entrenamendu Script-a

Entrenamendu script-aren oinarri handiena Hugging Face-eko run_ner.py script-a izan da. Script hori erabiliz eta Huggin Face-eko trainer moduluko parametroak doituz jartzen dira ereduak entrenatzen. Behin entrenamendua hasieratuta, gainbegiratzea posible da Weights & Biases-en web orria erabilita.

Hiper-parametroak

Ikasketa sakoneko edozein eredu entrenatzerakoan oso garrantzitsuak dira hiper-parametro egoki batzuk erabiltzea. Gure entrenamenduaren kasuan ondorengoak izan dira:

- *Pisuen gainbehera*: 0.001
- *Batch-aren tamaina*: 32

Hiper-parametro hauetaz gain aipatzeko da ere erabilitako *aro* kopurua. Gure kasuan *aro* kopurua atazaro aldatu egin da. Kopuru hauek ikusteko 5.1 taula sortu da.

EMD	**ED**	**EAE**
16	32	64

5.1 Taula: Aro kopurua atazaren arabera oinarri lerro fasean.

5.1.4 Testa

Behin ereduak entrenatuta daudela ebaluazioa falta da. Hori egiteko, EE-ko ataza bakoitza independenteki ebaluatzen da, test multzoko datuak erabiliz. Hori dela eta, argumentuen kasuan ez da katerik erabiltzen baizik eta test multzotik lortzen dira gertaerak.

Atal honetako gure helburua MEE artikuluko emaitzetara gerturatzea denez, eredu bakoitza bere test fitxategiarekin ebaluatu da. Adibidez, koreeraren eredua ebaluatzeko koreeraren test fitxategia erabiltzen da. Modu honetara, eredu guztiak ebaluatzen dira (*denak* eredua barne) eta emaitzak fitxategi batean jaso. Ebaluazio hau gauzatzeko 3.2.2 atalean azaldu den F1 neurria erabili da.

Horretaz gain, bigarren ebaluazio bat egin da hizkuntzen arteko transferentzia neurtzeko. Hau egin ahal izateko eredu guztiak interesatzen zaigun hizkuntzaren (gure kasuan euskararen) testarekin probatu dira.

5.2 Datu kopurua berdintzea

Behin oinarri-lerroa eginda, euskararen transferentziarekiko lehen datuak lortu nahi dira. Hori lortzeko aurreko fasean egindakoaren berdina egin da fase honetan, baina kasu honetan murrizketa bat aplikatuz.

5.2.1 Murrizketa

Hizkuntzen arteko esperimentuak egiterakoan datu gehien dituzten hizkuntzak ez nagusitzeko helburuarekin egin da murrizketa. Kasu honetan, entrenamenduko, garapeneko

eta testeko multzoak kontuan hartuta 1500 segmentuetara mugatu dira hizkuntza guztiak (1200 entrenamendu, 150 test eta 150 garapen). Horrela, hizkuntza guztiek segmentu kopuru berdina dute. Aukeraketa hau ausaz egin da 1500 segmentu baino gehiago dituzten hizkuntzetatik. Lehenik, ausazko aukeraketa argumentuak dituzten segmentuen artean egiten da, eta gero argumenturik gabekoak hartzen dira, ausaz ere. Modu honetara egiteak argumentuen informazio galerak ekiditen dizkigu. Izan ere, 4.1.1 atalean ikusi den bezala, hizkuntza batzuk argumentuak dituzten segmentu gutxi dituzte.

Hala ere, murrizketa honetan lortu den informazioa ez da oso orekatua. Izan ere, hizkuntza batzuetako segmentuek etiketa kantitate handiagoa dute beste hizkuntza batzuena baino. Gainera, segmentu askotan argumentuak ez daudenez etiketatuak, hizkuntza asko ez dira iristen 1500 argumentudun segmentuetara.

5.2.2 Entrenamendua

Entrenamenduari dagokionez aurreko fasearen berdina egin da, baina entrenatzeko murriztutako datuak erabili dira. Hiperparametroen atalean aldaketa txiki bat egin da ere.

Hiper-parametroak

Fase honetako hiper-parametroak aurreko fasearen berdinak izan dira, aro-kopurua izan ezik. Hauek, 5.2 taulan jaso dira. Aldaketa hau, murrizketa eta gero hizkuntza batzuen argumentu kopuru urria konpentsatzeko egin da.

EMD	**ED**	**EAE**
16	32	80

5.2 Taula: Aro kopurua atazaren arabera datu-kopuru berdintze fasean

5.2.3 Testa

Fase honetan ebaluatu nahi duguna aurreko fasearen berdina denez, ebaluazio berdinak mantendu dira. Berriro ere bi ebaluazio eginez: orokorra eta euskararena.

5.3 Hizkuntzen arteko transferentzia euskaraz

Azken fase honetan, hizkuntza guztietako ereduak euskararekin neurtzea da helburua. Horretarako, beste murrizketa bat aplikatu da, aurreko murrizketarekin alderatuz datu orekatuagoak lortzen dituena.

5.3.1 Murrizketa

Etiketa kantitatea orekatuta mantentzeko helburuarekin, bigarren murrizketa bat egitea erabaki da. Kasu honetan murrizketa segmentuen gainean egin beharrean etiketa kopuruaren gainean egingo da. Hau egiteko etiketa kantitate maximo bat jarri da ataza bakoitzeko. Maximo hori baina etiketa gehiago dituzten hizkuntzetan segmentuak ausaz aukeratzen joango dira, etiketa kopurua maximo horrekin berdindu arte. Entitate eta gertaeren kasuan etiketa kopuru txikien duen hizkuntzarekin berdindu da aipatutako maximoa. Bi kasuetan koreera da, 12508 entitate izanik eta 1125 gertaera. Bestalde, argumentuen kasuan etiketa gutxien dituen hizkuntza gaztelera da 802-rekin, baina maximo hau txikiegia izango zenez, turkieraren kopurua hartu da maximo moduan, 1416 argumentuekin. Aipatutako kopuru hauek ikusgai daude 4.3 taulan.

Horretaz gain, bigarren murrizketa honetan hizkuntza bakoitzaren garapen eta testeko multzoak alde batera utzi dira, biak entrenamendu multzoari batuz. Multzoen batuketa horrekin lortzen dira hizkuntza txikienetan lehen aipatutako kantitateak.

Hau horrela egiteak badu bere zergatia: azken murrizketa honetan soilik euskarako datumultzoa probatuko da. Hori dela eta, euskara ez den hizkuntzen testeko multzoak ez ditugu behar, eta optimizatu nahi duguna euskarazko testa denez euskarazko garapeneko multzoa erabiltzea da egokiena, beste hizkuntzena alferrikakoa izanik.

5.3.2 Entrenamendua

Fase honetan eredu bakoitza bere garapeneko datu-multzoarekin entrenatu beharrean euskarako garapenarekin entrenatu dira. Horretaz gain, 3 aldiz entrenatu dira ereduak, bakoitzean ausazko hazi desberdin bat erabiliz; zehazki, 16, 85 eta 44 zenbakiak erabili dira hazi moduan. Honen helburua, 3 emaitza desberdin lortzeaz gain, hauen bataz-besteko eta desbideratze estandarra lortzea da.

Hiper-parametroak

Fase honetako hiper-parametroak aurreko fasearen berdinak izan dira, aro-kopurua izan ezik. Hauek, 5.3 taulan jaso dira.

EMD	**ED**	**EAE**
64	64	64

5.3 Taula: Aro kopurua atazaren arabera hizkuntzen-arteko transferentzia fasean

5.3.3 Testa

Kasu honetan ez da test orokorrik egiten, soilik euskarako test-a probatzen da. Hori lortzeko, eredu guztiak euskararekin ebaluatu dira. Kasu honetan F1 puntuazio bueltatzeaz gain, doitasuna eta estaldura ere kalkulatu dira. Kalkulu hau 3 aldiz egingo da, hirugarren faseko ereduen hazi bakoitzeko behin, bakoitzean ematen duen emaitza jasoz. Emaitza horiekin bataz-bestekoak eta desbideratze estandarrak kalkulatu dira.

6. KAPITULUA

Emaitzak

Kapitulu honek bi zati izango ditu: Lehenengoan, fase desberdinetan gure sistemak lortzen dituen emaitzak azalduko dira. Ondoren, azken fasean lortutako emaitzetatik habiatuz, hizkuntza bakoitzaren ezaugarriak kontuan hartzen dituen azterketa sakonago bat egingo da.

6.1 Esperimentuen emaitzak

Atal honetan 5 atalean azaldutako hiru faseen emaitzak erakutsiko dira, bakoitzean gertatutakoa adieraziz.

6.1.1 Oinarri-lerroa

Lehen fase honetan hizkuntzen datuak ez dira murrizten. Fase honen helburu nagusia Veyseh et al. (2022)-ren emaitzetara gerturatzea izan da. Hala ere, euskarako datu-multzoarekin ere egin dira probak. Bi proba hauen emaitzak jasotzen ditu 6.1 taulak, bi zutabetan bereizita: domeinuan (hizkuntza propioan) eta euskaraz. Zutabe hauetako bakoitza hiru azpizutabetan banandua dago, bat ataza bakoitzeko. Bestalde, errenkadetan hizkuntzak daude.

Lehenik, domeinuan zutabean, hizkuntza bakoitza bere testarekin ebaluatzen da, 6.2 taularekin alderatzeko helburuarekin. Emaitzak nahiko desberdinak izan dira: Entitateen arloan hobeak izan diren arren, gertaerak eta argumentuak detektatzeko garaian ez dira lortu

	Domeinuan			Euskaraz		
Hizkuntzak	Entitate	Gertaera	Argumentu	Entitate	Gertaera	Argumentu
Ingelesa	0.80	**0.79**	0.66	**0.61**	0.48	**0.42**
Poloniera	0.81	0.69	**0.78**	0.59	**0.53**	0.22
Gaztelera	**0.85**	0.64	0.31	0.58	0.49	0.03
Portugesa	0.80	0.58	0.70	0.58	0.13	0.28
Japoniera	0.44	0.37	0.54	0.5	0.05	0.14
Hindiera	0.76	0.47	0.44	0.59	0.40	0.12
Koreera	0.71	0.42	0.38	0.48	0.07	0.17
Turkiera	0.72	0.58	0.24	0.48	0.48	0.13
Denak	0.80	0.68	0.61	0.52	0.52	0.28

6.1 Taula: Oinarri lerroaren emaitzak F1 erabiliz

hain emaitza onak. Bereziki txarrak izanik, turkierak, koreerak eta gaztelerak argumentuetan ematen dituzten emaitzak, gertaera kopuru gutxien dituzten hizkuntzak izanik 4.3 taulan ikus daitekeen moduan. Bi taulen arteko aldaketa hau hiper-parametroen desberdintasunagatik gerta daiteke, ez baitira artikuluko berdinak erabili.

Hizkuntzak	Entitate	Gertaera	Argumentu
Ingelesa	0.70	**0.71**	0.61
Poloniera	0.69	0.59	0.60
Gaztelera	0.70	0.66	0.60
Portugesa	**0.75**	**0.71**	**0.69**
Japoniera	0.68	0.68	0.68
Hindiera	0.66	0.58	0.58
Koreera	0.57	0.61	0.68
Turkiera	0.72	0.66	0.56

6.2 Taula: MEE artikuloko emaitzak F1 erabiliz.

Bestalde, nahiz eta fase honen helburua ez zen hizkuntzen arteko transferentzia egitea, lehen hurbilketa bat egiteko euskarako datu-multzoaren test-a erabili da aurretik entrenatutako ereduak probatzeko; emaitzak 6.1 taulako euskaraz zutabean jasoz. Domeinuko emaitzekin alderatuz, emaitzak asko jaitsi dira, normala den modura ereduak ez baitzeuden prest euskara tratatzeko. Euskaraz zutabeko emaitzetan zentratzen bagara ikus daiteke datu gehien dituzten hizkuntzek (ingelesa esaterako) lortzen dituztela emaitza hoberenak. Kasu honetan emaitza oso txar batzuk egon arren (japoniera eta koreeraren gertaerak), ez dira oso garrantzitsuak, fase honen helburua ez baizen euskarako datuak detektatzea.

6.1.2 Datu kopurua berdintzea

Fase honetan lehen murrizketa aplikatzen da. Honen helburua, segmentu kopuru berdina izateak hizkuntzen arteko transferentziako emaitzetan duen eragina ikustea da. Hala ere, test orokor bat ere egin da, hizkuntzei datu galerak nola eragin dien ikusteko. Datu guzti hauek 6.3 taulan jaso dira, aurreko fasean bezala bi zutabeetan banatuta: domeinuan eta euskaraz.

	Domeinuan			Euskaraz		
Hizkuntzak	Entitate	Gertaera	Argumentu	Entitate	Gertaera	Argumentu
Ingelesa	**0.81**	**0.78**	0.64	**0.61**	0.43	**0.40**
Poloniera	0.78	0.68	**0.76**	0.60	**0.46**	0.16
Gaztelera	0.84	0.63	0.36	0.58	0.43	0.02
Portugesa	**0.81**	0.56	0.67	0.56	0.12	0.26
Japoniera	0.45	0.36	0.52	0.52	0.14	0.14
Hindiera	0.76	0.46	0.45	0.59	0.25	0.11
Koreera	0.72	0.42	0.34	0.50	0.23	0.15
Turkiera	0.68	0.54	0.22	0.58	0.43	0.11
Denak	0.73	0.52	0.63	0.52	0.46	0.27

6.3 Taula: Datu-kopuru berdintze fasearen emaitzak F1 neurtuz

Domeinuko emaitzei dagokionez esperotakoa ikusi da: 6.1 taularekin alderatuz domeinuko emaitzak okertu dira. Hala ere, aldaketa hau ez da oso handia izan bataz-bestean soilik 0.02 F1 puntu galdu baitira. Atazari dagokionez puntuazio galera handien eduki duena gertaerak izan dira 0.03-ko galera batekin. Azkenik, entrenamenduko hizkuntzei dagokionez gehien galdu duena denak izan da 0.07 puntuko galera batekin. Denak ereduak hizkuntza guztien murrizketak pairatzen dituenez, galera hau zentzuzkoa dirudi.

Lehen murrizketa honekin hizkuntzen arteko transferentziaren lehen emaitza esanguratsuak lortzea zen helburua. Emaitza horiek euskaraz zutabean agertzen dira. Bertan 5.2.1 atalean azaldutakoaren ondorioa ikus daiteke: anotazio dentsitate handiena duten hizkuntzek abantaila erakusten dute beste hizkuntzekiko. Anotazio dentsitatea eta kopurua 4.3 taulatik atera daitezke, taula horretako argumentuen ataleko bi hizkuntza nabarienak lortu dituzte emaitza hoberenak: ingelesak, 1500 argumentudun segmentu baina gehiago dituelako, eta portugesak segmentu kopuru txikian anotatutako argumentu kantitate handia duelako. Bestalde, emaitza okerrenak ere 4.3 taularekin erlazioa dauka: gaztelera baita argumentu kopuru gutxien dituen hizkuntza.

6.1.3 Hizkuntzen arteko transferentzia euskaraz

Anotazio kopuruak guztiz parekatzeko 5.3.1 atalean azaldutako murrizketa aplikatzen da. Hori dela eta soilik euskaraz ebaluatu dira ereduak. Horretaz gain, 5.3.2 atalean azaldu den bezala, 3 eredu ezberdin entrenatu dira, emaitzen fidagarritasuna handitzeko asmoz. Aipatzeko da ere, fase honetan soilik F1 puntuazioa kalkulatu beharrean, estaldura eta doitasunak ere kalkulatu direla ataza bakoitzerako. Eredu hauen emaitzak 6.4, 6.5, 6.6 taulatan banatu dira, bat ataza bakoitzeko. Taula hauek lortzeko 3 ereduek eman duten bataz-bestekoak eta desbideratze estandarrak kalkulatu dira

Hizkuntzak	Doitasuna	Estaldura	F1
Ingelesa	0.57 ± 0.01	0.62 ± 0.00	0.60 ± 0.01
Poloniera	0.55 ± 0.01	**0.65 ± 0.01**	0.59 ± 0.01
Gaztelera	0.51 ± 0.02	0.60 ± 0.01	0.55 ± 0.02
Portugesa	**0.63 ± 0.01**	0.61 ± 0.00	**0.62 ± 0.01**
Japoniera	0.51 ± 0.03	0.61 ± 0.00	0.56 ± 0.02
Hindiera	0.52 ± 0.01	0.63 ± 0.02	0.57 ± 0.01
Koreera	0.38 ± 0.02	0.60 ± 0.01	0.47 ± 0.01
Turkiera	0.51 ± 0.01	0.64 ± 0.02	0.56 ± 0.01
Denak	0.50 ± 0.03	0.69 ± 0.03	0.58 ± 0.03

6.4 Taula: Hizkuntza arteko transferentzia entitate detekzioan

Hizkuntzak	Doitasuna	Estaldura	F1
Ingelesa	0.40 ± 0.02	0.58 ± 0.01	0.47 ± 0.02
Poloniera	**0.44 ± 0.07**	0.53 ± 0.03	0.47 ± 0.03
Gaztelera	0.43 ± 0.01	0.58 ± 0.02	**0.49 ± 0.01**
Portugesa	0.16 ± 0.08	0.48 ± 0.08	0.23 ± 0.11
Japoniera	0.07 ± 0.06	0.20 ± 0.03	0.10 ± 0.06
Hindiera	0.27 ± 0.03	**0.65 ± 0.03**	0.38 ± 0.02
Koreera	0.15 ± 0.04	0.55 ± 0.11	0.23 ± 0.07
Turkiera	0.39 ± 0.02	0.40 ± 0.02	0.40 ± 0.00
Denak	0.43 ± 0.02	0.60 ± 0.00	0.50 ± 0.01

6.5 Taula: Hizkuntza arteko transferentzia gertaera detekzioan

Taula hauetan ikus daiteke orokorrean datu guztiekin entrenatutako ereduak (**denak**) funtzionatzen duela egokien. Hala ere, hizkuntza bakoitza begiratzen hasten bagara, entitate detekzioan **portugesak** lortzen ditu emaitzarik altuenak. Gertaeren detekzioan, aldiz, **gaztelerak**. Eta azkenik, argumentuen detekzioak **turkiera** da nagusitzen dena. Badirudi hizkuntza bakoitzaren ezaugarri linguistikoak garrantzia irabazten dutela datu-kopuruak berdintzean, hurrengo atalean honen azterketa sakonago bat egingo da.

Desbideratzeei dagokionez, orokorrean nahiko txikiak dira. Beraz, 3 ereduek nahiko emaitza antzekoak bueltatzen dituztela ondorioztatu daiteke.

Hizkuntzak	Doitasuna	Estaldura	F1
Ingelesa	**0.18 ± 0.02**	0.13 ± 0.02	0.15 ± 0.01
Poloniera	0.12 ± 0.01	0.06 ± 0.00	0.08 ± 0.00
Gaztelera	0.03 ± 0.02	0.03 ± 0.01	0.03 ± 0.02
Portugesa	0.14 ± 0.02	0.10 ± 0.02	0.12 ± 0.02
Japoniera	0.11 ± 0.01	0.15 ± 0.03	0.13 ± 0.02
Hindiera	0.08 ± 0.02	0.14 ± 0.03	0.10 ± 0.03
Koreera	0.14 ± 0.02	0.20 ± 0.01	0.16 ± 0.02
Turkiera	0.17 ± 0.00	**0.21 ± 0.02**	**0.18 ± 0.01**
Denak	0.25 ± 0.01	0.34 ± 0.00	0.29 ± 0.01

6.6 Taula: Hizkuntza arteko transferentzia argumentu erauzketan

6.2 Ezaugarrien Gaineko Azterketa

Aurreko atalean azaldu diren emaitzen inguruan gehiago sakontzeko, hizkuntza bakoitzaren ezaugarri linguistikoak kontuan hartzen dituen azterketa bat egin da. Ezaugarri horiek 6.7 taulan azaltzen direnak dira.

Hizkuntza	Morfologia	Morfosintaxia	Hitzen Ordena	Alfabetoa	Kokaleku Geografikoa
Euskara	Aglutinatzaile	Ergatibo-Absolutibo	SOV	Latin	Europa Mendebaldea
Ingelesa	Fusionatzaile	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Gaztelera	Fusionatzaile	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Portugesa	Fusionatzaile	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Poloniera	Fusionatzaile	Nominatibo-Akusatibo	SVO	Latin	Europa Ekialdea
Turkiera	Aglutinatzaile	Nominatibo-Akusatibo	SOV	Latin	Europa Ekialdea/Asia Mendebaldea
Hindiera	Fusionatzaile	Ergatibo Banandua	SOV	Devanagari	India
Japoniera	Aglutinatzaile	Nominatibo-Akusatibo	SOV	Kanji eta Kana	Asia Mendebaldea
Koreera	Aglutinatzaile	Nominatibo-Akusatibo	SOV	Hangul	Asia Mendebaldea

6.7 Taula: Hizkuntza bakoitzaren ezaugarriak

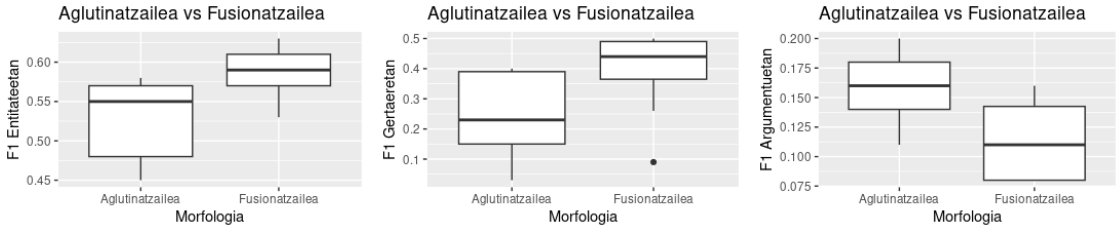
Gure hipotesian hizkuntzen ezaugarriak euskararen antzekoak izatea positiboa iragingo duela da. Zehazki ondorengoak izango dira, gure ustez, ezaugarri bakoitzaren eraginak:

- Alde batetik, ataza lexikoetan (EMD eta ED) kokalekua eta alfabetoa izango dira garrantzitsuak.
- Bestalde, ataza sintaktikoetan (EAE) morfologia, morfosintaxia eta hitzen ordena izango dira nabarmenak.

Hurrengo ataletan ezaugarri hauek aztertuko dira.

6.2.1 Morfologia

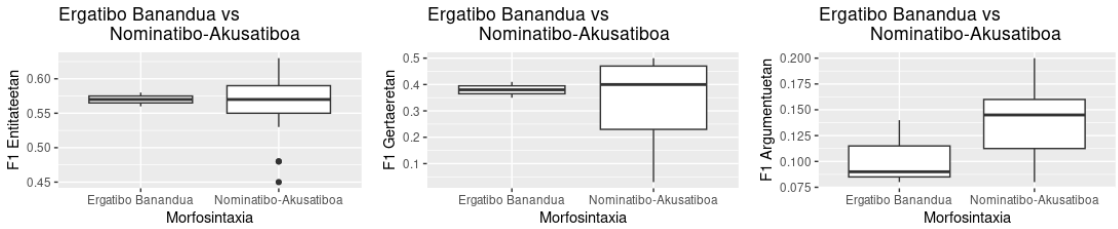
Morfologiak hitzen barneko estruktura aztertzen du. Hori dela eta, ataza sintaktikoekin lotura garrantzitsu bat du; beraz, argumentuen erauzketari eragingo die gehienbat. Gure kasuan bi morfologia mota ditugu: fusionatzailea eta aglutinatzaileak. Euskararena aglutinatzailea da, beraz gure hipotesia hizkuntza aglutinatzaileak EAE atazan hobeto funtzionatzen dutela izango da. Konparaketa hau ikusteko 6.1 irudia sortu da. Honek, gure hipotesia zuzena izan daitekeela erakusten du.



6.1 Irudia: Hizkuntza desberdinen morfologiaren konparazioa euskararentzat

6.2.2 Morfosintaxia

Morfosintaxiak morfologiaren antzeko azterketa bat egiten du, baina kasu honetan perpausen azterketa ere gehituz. Hau, morfologia bezala, sintaxiarekin lotura duen ataza bat da, hori dela eta argumentuen azterketa izango da garrantzitsuena ere kasu honetan. Horretaz gain, ez dugu euskararen morfosintaxi berdina duen hizkuntzarik, eta desberdina den morfosintaxi bakarra hindierarena da. Hau dela eta, konparaketa hau ez da guztiz aipagarria izango. Konparaketa 6.2 irudian ikus daiteke.

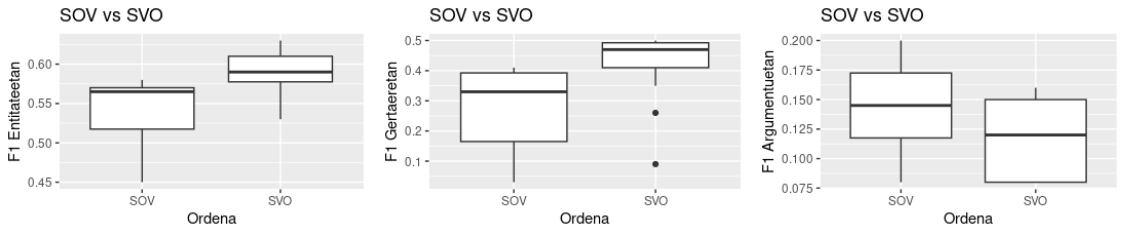


6.2 Irudia: Hizkuntza desberdinen morfosintaxiaren konparazioa euskararentzat

6.2.3 Hitzen Ordena

Hitzen ordenari dagokionez 6.7 taulan bi mota ikus daitezke: Alde batetik, SOV (Subject Object Verb) mota dutenak, subjektua, objektua eta aditza orden horretan dituzten hizkuntzak dira. Eta bestalde, SVO (Subject Verb Object) motakoak, subjektua, aditza eta objektua ordena dutenak dira.

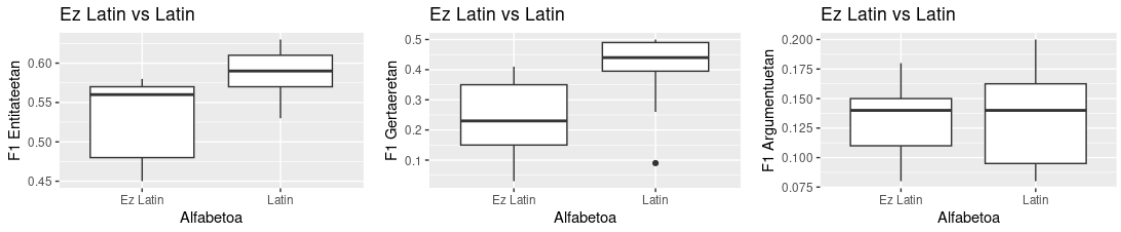
Hitzen ordena garrantzi hartzen du argumentuak detektatzeko garaian. Izan ere, ataza honetan ematen da aditzaren informazioa. Beraz, argumentuak detektatzeko erabiltzen dugun hizkuntzak euskararen orden berdina edukitzea positiboki eragin beharko luke. Hori konprobatzeko 6.3 grafikoa egin da. Bertan ikus daiteke gure hipotesia bete daitekeela.



6.3 Irudia: Hizkuntza desberdinen hitz ordenaren konparazioa euskararentzat

6.2.4 Alfabetoa

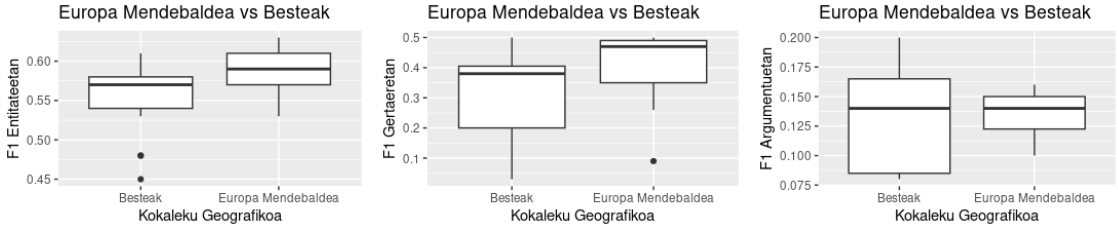
Analizatu den beste ezaugarria alfabetoa da. Alfabetoak ataza guztiei eragiten dien arren, ataza lexikoei (entitate eta gertaerei) egiten die eragin nabarmenena. Honen hipotesia aurrekoaren antzekoa da: euskararen alfabeto berdina edukitzea positiboki eragin beharko luke F1 puntuazioan. Hori frogatzeko 6.4 grafikoa sortu da. Bertan, latindar alfabetoa erabiltzen dituzten hizkuntzak beste alfabetoak erabiltzen dituztenekin konparatzen dira. Irudia ikusiz, badirudi eragin hori existitzen dela.



6.4 Irudia: Hizkuntza desberdinen alfabetoen konparazioa euskararentzat

6.2.5 Kokalekua

Azken ezaugarria kokalekua da. Kokalekua ezaugarri moduan aztertzea ez da erraza. Hizkuntza askoren kokaleku zehatza ezartzea oso zaila baita. Hala ere, 6.7 taulan azaltzen diren kokalekuak hartuz konparaketa egin da, 6.5 irudia sortuz. Kokalekuak bereziki ataza lexikoa eragiten dienez, entitateak eta gertaerak izango dira aipagarrienak. Irudia ikusiz badirudi euskararen inguruko hizkuntza izateak positiboki eragiten duela ataza lexikoetan.



6.5 Irudia: Hizkuntza desberdinen kokalekuaren konparazioa euskararentzat

7. KAPITULUA

Ondorioak eta etorkizuneko lana

7.1 Ondorioak

Kapitulu honetan proiektuan zehar lortu ditugun ondorioak azaltzen dira: emaitzei dagokienak eta ondorio pertsonalak. Azkenik, proiektuaren sakontasuna handitzeko eman daitezkeen hobekuntzak azaltzen dira.

7.1.1 Emaitzen ondorioak

Proiektuak zituen helburuak bete dira. Alde batetik, MEE datu-multzoa oinarri bezala hartuz lehen ereduak entrenatzea lortu da; nahiz eta emaitza guztietan ez den lortu (Veyseh et al., 2022)-ren emaitzetara iristea, gure helburua betetzeko nahikoak izan dira. Bestalde, hizkuntza hauen eta euskararen arteko transferentzia egitea ere lortu da, gure helburu nagusia betez.

MEE datu-multzoko hizkuntzen eta euskararen arteko transferentzian sakontzen badugu, ikusi da hasiera batean segmentu gehien zuten hizkuntzak nagusitzen zirela. Hori dela eta, hizkuntza guztiak segmentu kopuru berdinarekin entrenatzea erabaki da.

Entrenamendu berri hau eta gero, emaitzak ez dira hain desorekatuak izan. Hala ere, etiketa kopuru gehien zuten hizkuntzak ziren nagusitzen zirenak. Hizkuntzen arteko datu oreka guztiz bermatzeko, etiketa kopurua murriztu da, denak etiketa kantitate berdina izateko helburuarekin. Murrizketa hau egin eta gero azken entrenamendu bat egin da.

Bertako emaitzetan ikusi da datu-kopuruak berdintzean, hizkuntza bakoitzaren ezaugarri linguistikoak garrantzia irabazten dutela.

Azkenik, azken entrenamendu honen emaitzak hizkuntza bakoitzaren ezaugarri linguistikoekin erlazionatu dira, ondorio interesgarri batzuk lortuz. Alde batetik, ataza lexikoetan (EMD eta ED) euskararen alfabeto berdina izateak eta kokaleku antzekoan egoteak positiboki eragiten duela ikusi da 6.4 eta 6.5 irudiak aztertu eta gero. Bestalde, ataza sintaktikoetan (EAE) morfologia eta hitzen ordena euskararen berdina izateak; hau da, morfologia aglutinatzaile bat izateak eta SOV ordena edukitzeak, positiboki eragiten duela ikusi da 6.1 eta 6.3 irudiak aztertu ostean. Azkenik, morfosintaxiaren kasuan, ez da lortu ondorio aipagarririk. Hau gertatzeko bi arrazoi nagusi egon dira: alde batetik, morfosintaxi desberdina duten hizkuntzen falta; eta bestalde, euskararen morfosintaxi berdina duen hizkuntzarik ez egotea.

7.1.2 Ondorio Pertsonalak

Lan hau maila pertsonalean oso aberasgarria izan da hainbat arrazoiengatik: alde batetik, sakontasun honetako proiektu bat egiteak hizkuntzaren prozesamenduko ezagutza handitzen lagundu du. Bestalde, luzera honetako proiektuen kudeaketa teknikak ikasteko onuragarria izan da.

Horretaz gain, prozesu guzti honetan oso lagungarriak izan dira IXA-ko zerbitzariak. Hauek exekuzioak asko azkartzeaz gain, mota honetako zerbitzarien erabilera ikasten lagundu didate.

Azkenik, eskertzekoa izan da zuzendarien partetik jasotako laguntza; posta eta bilera bidez edozein zalantza edo arazo argitzeko beti prest egon baitira. Laguntza hori esker lortu da proiektu hau epeen barruan egin ahal izatea. Pertsonalki, oso gustura aritu naiz beraiekin lanean.

7.2 Etorkizuneko lana

Proiektu honen helburuak lortu diren arren, badaude jarraipen edo hobekuntza posible batzuk.

Alde batetik, onuragarria izango litzateke euskararen entrenamenduko datu-multzo bat

izatea. Izan ere, modu honetara, euskararekin entrenatutako eredu baten emaitzak lortuko ziren, transferitutako ereduekin konparatzeko interesgarria izanik.

Bestalde, hizkuntzen transferentzia euskararentzako bakarrik egin beharrean MEE datumultzoko beste hizkuntzentzako egitea interesgarria litzateke ere. Hau egitean ezaugarri linguistikoak hizkuntza transferentzian duten eragina sakonago aztertu ahalko litzateke, hainbat hizkuntza egongo baitziren ezaugarri ezberdinak alderatzeko.

Eranskinak

Bibliografia

Ahn, D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Alec Radford, KarthikÑarasimhan, T. S. and Sutskever, I. (2018). Improving language understanding by generative pre-training.

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.

David E. Rumelhart, G. E. H. and Williams, R. J. (1985). Learning internal representations by error propagation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dolicki, B. and Spanakis, G. (2021). Analysing the impact of linguistic features on cross-lingual transfer. *CoRR*, abs/2105.05975.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Ji, H. and Grishman, R. (2008). Refining event extraction through cross-document inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey.

Nguyen, T. H., Cho, K., and Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Sha, L., Qian, F., Chang, B., and Sui, Z. (2018). Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

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

Veyseh, A. P. B., Ebrahimi, J., Dernoncourt, F., and Nguyen, T. H. (2022). Mee: A novel multilingual event extraction dataset.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium*.