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

Hizkuntza-arteko distilazioa eta antzekotasun semantikoa maiz egiten diren galderak erantzuteko

Egilea

Ander Berrondo Urruzola

Zuzendariak

Ander Barrena Madinabeitia eta Arantxa Otegi Usandizaga

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2022ko ekaina

Sailak: Lengoaia eta sistema informatikoak

Laburpena

Azken urteotan gure artean egon den *COVID-19aren* harira, gaixotasunari buruzko dudak argitzeko sistema bat garatu nahi izan da, Osakidetzak eskainitako Maiz Eginiko Galderen (MEG) multzo batean oinarrituta; honela, sistemaren helburua, erabiltzaileak galdera bat eginik, MEGean haren semantikoki antzekoena dena bilatzea da, sistemak iragarritako galderaren erantzunak duda argitzeko balioko duela suposatuz. Hori horrela, ideia nagusia euskarazko zein gaztelaniazko sistema bana garatzea izan da, eta hizkuntza hauen baliabideak ingelesekoekin alderatuz nahikoa mugatuak direnez, distilazio bidezko ikasketa erabili da ahalik eta eredu eraginkorrenak sortzeko, ingelesez birdoitutako ereduetatik ikasiz. Hala, distilazioak ekartzen dituen onurak aztertu nahi izan dira, distilatu gabeko ereduekin alderatuz domeinuko zein domeinuz kanpoko ebaluazioaren bitartez. Emaitzei erreparatuta, helburu nagusia bete dela ikus daiteke, aipatutako bi ebaluazio motetan distilazioaren bidez hobekuntza nabarmenak lortu baitira distilatu gabeko ereduekin alderatuz, kasu batzuetan emaitza benetan lehiakorrak lortuaz.

Abstract

With regards to the *COVID-19* that has been among us these last years, a system which solves doubts about the disease has been created, based on the information provided in Osakidetza's Frequently Asked Questions (FAQs); in other words, the aim of the system is, having received a question by an user, to find the most semantically similar one in the FAQs, assuming that the answer of the question predicted by the system serves to solve the doubt. So, the principal idea has been to develop a system for both Basque and Spanish, and given that the resources of these languages are quite limited compared to those of English, knowledge distillation has been used in order to create the most effective models possible, learning from models fine-tuned in English. Thereby, the benefits of distillation have been analyzed, comparing them with the non-distilled models by means of both in-domain and out-of-domain evaluation. Based on the results, it can be stated that the main objective has been achieved, as significant improvements in both types of evaluation have been obtained through distillation compared with non-distilled models, getting really competitive results in some cases.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Azken urteotan gure gizartean izan dugun gertakari garrantzitsuena, ezbairik gabe, *SARS-CoV-2* birusak sortutako ***COVID-19*** gaixotasuna izan da. Bi urte badira jada lehen itxialdia agindu zutenetik, eta geroztik bizi izandako egoera (nahiz eta, zorionez, pixkanaka normaltzen ari den) guztiz berria zein bitxia izan da denontzat. Denbora guzti honetan, pandemiaren bilakaeraren arabera, lege jakin batzuk ezarri izan dira, etxeratze ordua, mugikortasuna, batera egon zitekeen pertsona kopurua... bezalako kontzeptuak mugatu izan dituztenak; gainera, kutsatu zein ospitalizazio kopurua maiztasun handiarekin aldatzen denez, oso denbora tarte txikian murriztapen ezberdinak argitaratu izan dituzte agintariek. Horregatik, sarritan konplexua izan da zer (eta zer ez) egiteko baimenduta geunden jakitea: Joan al naiteke ondoko herrira erosketak egitera? Atera al naiteke kalera gaueko 22:00etatik aurrera? Mota guztietako galderak izan ditugu, eta haiei erantzun bat bilatzea ez da inondik inora lan erraza izan adituetara jo gabe.

Hain zuzen ere, nahiz eta gutariko askok jakin ez, Osakidetza [1] une bakoitzean aktibo egon diren murriztapenekin erlazionatutako **Maiz Eginiko Galderak** (MEG [2], edo ingelesez FAQ) argitaratzen joan da bere webgunean; modu honetan, era orokor batean bada ere, herritarren artean sortutako duda gehienei soluzio bat emateko baliabide bat eskaini ahal izan dute. Lan honetan, aipatutako baliabidea oinarri hartuta, zalantza hauek argitzeko prozesua automatizatu nahi izan da adimen artifizialeko teknikak baliatuz; hau da, proiekturako garatutako sistemaren helburua herritarrek egindako galdera bakoitza prozesatu eta MEGean haren (semantikoki) antzekoena bilatzea da. Honela, sistemak egiten

[1] https://www.osakidetza.euskadi.eus/ataria/

[2] MEG: testuinguru jakin batean eta gai zehatz bati buruzko galdera-erantzun zerrenda bat.

duen iragarpena jatorrizko galdera ordezkatzeko gai dela suposatzen dugu, MEGeko galdera horren erantzunak herritarrak egindako galderan azaldutako zalantza argitzeko balio dezakeelako.

Horretaz gain, nahiz eta proiektuaren jomuga pandemiaren inguruko galderak erantzutea den, sistemak orokortzeko duen gaitasunari esker beste edozein arlotara eraman daiteke. Gaur egun, erakunde zein enpresa askok MEGak erabiltzen dituzte, adibidez, haien politikaren inguruko dudak argitzeko; ondorioz, honelako sistema bat garatzea onuragarria da esparru askotariko langileentzako, asko errazten baita desiratutako informazioa aurkitzeko prozesua. Hain zuzen ere, sistema hau txertatuta duen plataforma batek oso zerbitzu erabilerraz eta intuitiboa eskain diezaieke erabiltzaileei: galdera idatzi hutsarekin (edozein mota zein zailtasunekoa) baliozko erantzun bat jasoko dute normalean, MEG osoa goitik behera irakurri behar izan gabe helburu bera lortzeko.

Lehen aipatu bezala, proposatutako duda edozein mota zein zailtasunekoa izan daiteke; eta honek, ezbairik gabe, sistemaren eraginkortasuna baldintzatzen du, esaldi konplexuegiak nahasketak ekar baititzakete. Horretaz gain, galderen idazkera gainbegiratuta ez dagoenez, akatsez betetako esaldien aurka ere lehiatu beharko da askotan sistema. Kasu honetan, gainera, *COVID-19* gaixotasunarekin erlazionatutako duden esparrua oso zabala da, galdera anbiguoak, informazio gehigarrikoak (geografikoa, adibidez) zein konplexuak aurki baitaitezke:

- *Zenbat elkartu gaitezke?*
- *¿Urkullu, que vive en Durango, puede subir al Gorbea mañana?*
- *Nire dentista, gorrian dagoen beste herri batean dago. Joan naiteke han dudan hitzordura?*

Era berean, proiektu honen oinarrian **antzekotasun semantikoa** izeneko kontzeptua dago, lehen esan bezala, antzekoena den galdera bilatzea baita sistemaren helburua; laburki azalduta, hitz edo esaldi ezberdinen arteko erlazioa aztertzen du, haien esanahia ulertuz eta interpretatuz. Nahiz eta gizakion begietara erraza eman dezakeen, hizkuntzaren prozesamenduko arloan azken urteotan izan den ataza landuenetariko bat izan da, sistema batentzako oso korapilatsua izan baitaiteke esaldi baten zentzua aztertzea (are gehiago, adibidez, anbiguoa bada). Horrexegatik, erronka polita da lan honetan aurkezten dena, batez ere kontuan hartuta, besteak beste, euskara bezalako baliabide gutxiko hizkuntza batekin eraman dela aurrera.

Horretaz gain, proiektura hoberen egokitzen den aukera zein den aztertu ondoren, **distilazio** teknika erabiltzea erabaki da, nagusiki, euskararaz antzekotasuna neurtzeko eredu egokirik ez dagoelako eskuragarri. Motzean azalduta, bi eredu ezberdin edukirik (irakaslea eta ikaslea), eredu ikaslearen helburua irakaslearen funtzionamendua imitatzea da, esaldien errepresentazio bektorialak ahalik eta antzekoenak izan daitezen saiatuz. Honela, irakaslea ingelesezko eredu elebakarra izango da, eta ikaslea, ordea, ingeles zein euskarazko eredu eleanitza, irakaslearen errendimendu bera eskuratzen saiatuko dena. Proiektu honen ardatza euskara da eta, hortaz, hasierako asmoa ingelesa-euskara hizkuntza-parearekin soilik lan egitea zen; hala ere, azkenean ingelesa-gaztelania ere jorratzea erabaki da, baliabide gehiagoko hizkuntza batekin sistemaren erantzuna zein den aztertzeko. Horretaz gain, emaitzen kalitatea neurtu ahal izateko, ikasle eredua zuzenean (distilazio bidezko ezagutza jaso gabe) testatzean ezarri da oinarri-lerroa; eta, gehigarri gisa, *zeroshot* teknika ere frogatu da, euskarazko eredua gaztelaniazko datuekin (eta alderantziz) testatzean datzana.

Txostenaren egiturari dagokionez, sarrera txiki honen ondoren, 2. kapituluan arloaren egoera azalduko da, batez ere hitz zein esaldi-errepresentaziorako ereduen, antzekotasun semantikoaren eta distilazio teknikaren inguruko informazioa emanaz, gaur egungo egoeran zein aurrekarietan zentratuz. Honen ostean, 3. kapituluan metodologia azaltzeari ekingo zaio, proiektu hasieratik gaurdaino eman diren urratsak zehaztuz; adibidez, ereduen aukeraketa, datu-multzoen lorpena edo ebaluazio metriken hautaketa nola eman diren azalduko da. Ondoren, 4. kapituluan garapeneko prozesuan burututako esperimentuak aurkeztuko dira, domeinuko datu-multzo bat erabiliaz eskuratu direnak; segidan, 5. kapituluan, domeinuz kanpoko testeko (*COVID-19aren* ingurukoa) emaitzak erakutsiko dira, aurreko kapituluan hautatutako ereduekin lortu direnak. Amaitzeko, 6. kapituluan ondorio batzuk azalduko dira, proiektutik bereganatutako ezagutza aintzat hartuz eta alde positibo zein negatibo garrantzitsuenak azpimarratuz; gainera, denbora edo baliabide arazoengatik aurrera eraman ezin izan diren atazak aurkeztuko dira, aukera izanez gero etorkizun batean bukatzeko asmoz.

2. KAPITULUA

Arloaren egoera

Atal honen helburua proiektuan zehar erabilitako baliabide ezberdinen bilakaera aurkeztea da, haien sorkuntzaz geroztik nola eboluzionatu diren aztertuz eta, azkenik, gaur egun zein egoeratan dauden ikusiz. Hasteko, gaur egun ezagutzen ditugun esaldi-errepresentazio teknika zein eredu konplexuetaraino iristeko eman behar izan diren pausuak analizatuko dira; hau da, errepresentazio tradizionaletatik hasi, gaur egungo testuinguruan sartu hitz-errepresentazio ereduen eskutik (eleanitzak bereziki aztertuaz) eta esaldi-errepresentazio ereduekin amaitu. Horretaz gain, proiektu honetan berebiziko garrantzia duen antzekotasun semantikoa aurkeztuko da, eta azkenik lanaren muina den distilazio teknikaren analisi sakon bat egingo da, aipamen berezia eginez proiektua garatzeko ezinbestekoa izan den artikulu zientifikoari.

2.1 Esaldien errepresentazioaren aurrekariak

Hizkuntzaren Prozesamendua (HP) arloaren historian egon den erronkarik handienetariko bat, ezbairik gabe, gizakien arteko komunikazioaren oinarri diren hizkuntzak konputazionalki prozesatzea izan da; hau da, nolabait, testu edo esaldi batean inplizituki datorren informazio (semantiko) osoa era numeriko batean irudikatu ahal izatea. Honela, behin idatzizko ezagutza guztia zenbakietara pasata, eskuragarri dauden baliabide matematikoekin posible da ataza oso konplexuak ebaztea.

Atal honetan, lehen aipatu bezala, gaur egun ditugun eredu zein teknika garatuak lortzeko zein prozesu eman den azalduko da, kronologikoki ordenatuta.

2.1.1 Esaldi mailako hitzen errepresentazio tradizionala

Hasteko, aipatu beharra dago azken urteotan eman den garapenaren ondorioz lortu direla gaur egun eskuartean ditugun emaitzak (batez ere, BERT teknikaren sorreraz geroztik); hasiera batean, ordea, baliabide mugatuagoak zeuden eta metodoak ez ziren hain eraginkorrak, baina beharrezkoak dira esaldien errepresentazioaren historia ulertzeko. Hasiera batean hitz mailan soilik egiten zen lana, eta teknika erabilienen artean *One-Hot* Hitz-Zakua [Zhang et al., 2010] dago, non hitz bakoitza bektore-indize batera lotuta dagoen, 1 edo 0 gisa markatuta esaldi jakin batean azaltzen bada edo ez, hurrenez hurren (2.1 irudian ikus daitekeen bezala). Metodo hau, nahiz eta oso azkarra izan, oso mugatua ere bada, ez baititu ez hitzen ordena, ez garrantzia ez eta informazio semantikoa jasotzen.

2.1 Irudia: Hitz-zakuaren errepresentazio grafikoa. Bertan, hitz bakoitza bektorearen dimentsio jakin batean errepresentatzen da (*the* lehenengoan, adibidez); honela, hitza esaldi jakin batean azaltzen bada 1 balioa izango du bektoreko dimentsio horretan, eta ez bada agertzen, 0 izango da.

Iturria: https://torrellesdefoix.cat/

Beste teknika tradizional bat TF-IDF (Terminoen Maiztasuna–Alderantzizko Dokumentu Maiztasuna) litzake, hitz batek dokumentu batean duen garrantzia kalkulatzeko balio duena, pisu jakin bat emanez azaltzen den testuinguruaren arabera. Honela kalkulatzen da:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{2.1}$$

non:

i = aztertutako hitza.

j = aztertutako dokumentua.

$\mathrm{tf}_{i,j}$ = i-ren agerpen kopurua j-n.

df_i = i azaltzen den dokumentu kopurua.

N = dokumentu kopuru totala corpusean.

Kasu honetan, azkarra eta sinplea izateaz gain, hitzaren garrantzia ere kontuan hartzen du teknikak; hala ere, oraindik ez da ez hitzen ordena ez eta informazio semantikoa jasotzen.

2.1.2 Hitzen errepresentaziorako eredu elebakarrak

Azaldutako errepresentazio tradizionaletan dimentsio handiko bektoreak erabiltzen dira, adibide bakoitza errepresentatzeko dimentsio gutxiren beharra izanik (balio gehienak, ondorioz, 0 dira); horregatik, sakabanatutako (*sparse*) bektore gisa definitzen dira, beren arteko eragiketak oso garestiak (eta askotan bideraezinak) izanik dimentsionaltasun handi hori dela eta. Arazo honi aurre egiteko, aurrerapauso nabarmen bat eman zen atal honetan azalduko diren bektore dentsoekin; kasu honetan, dimentsio askoz txikiagoko bektoreekin egiten da lan, eta dimentsio horietako bakoitza erabiltzen da errepresentazioetan informazio semantikoa jasotzeko, 0 balioak saihestuz.

Azpiatal honetan aipatu beharreko lehen eredu mota (eta sinpleena) errepresentazio **estatikoak** sortzen dituena da; errepresentazio hauek globalak kontsideratzen dira, ez baitute hitzaren testuingurua kontutan hartzen. Teknika honetarako hiztegi global bat sortzen da testu guztiko hitzak jasoaz (bakoitza soilik behin), eta ondoren, testuan ondoz-ondo sarrien azaltzen diren hitzei errepresentazio bektorial antzekoak esleitzen zaizkie. Honen arazoa, esanda bezala, errepresentazio horietan testuinguruak ematen duen zentzua (gertuko hitzetatik eratortzen dena) galdu egiten dela da. Adibidez, demagun azpiko esaldiko hitzentzako errepresentazio bektoriala lortu nahi dugula teknika honen bidez:

*Bihar **arte** dago **arte** berezien argazkien **arte** erakusketa.*

Kasu honetan, *arte* hitzak hiru esanahi ezberdin ditu (postposizioa, zuhaitz mota eta gizaekintza), eta ondorioz, hiru errepresentazio bektorial ezberdin eduki beharko lituzke; hala ere, errepresentazio estatikoen bidez ezinezkoa da hau lortzea. Modalitate honetan, bereziki nabarmendu beharreko hiru eredu daude, oso ongi ordezkatzen baitituzte errepresentazio estatikoak lortzeko modu nagusiak; hona hemen, laburki azalduta, haien ezaugarri garrantzitsuenak:

- **Word2Vec** [Mikolov et al., 2013]: Bektore dentsoen aitzindaria, mota honetako errepresentazioak erabiltzen lehen eredua izan baitzen. Testua erabiltzen du zuzenean sare neuronal bat entrenatzeko eta lortutako errepresentazioek hitz ezberdinak testuinguru lokal berean azaltzen diren (edo ez) jakiteko balio dute.

- **GloVe** [Pennington et al., 2014]: Corpus osoko hitzen ko-okurrentzietan zentratzen da eta lortutako errepresentazioek bi hitz batera zer probabilitatez azaltzen diren adierazten dute.

- **FastText** [Joulin et al., 2016]: Word2Vecen hobekuntza bat da, harekiko (eta baita ere GloVerekiko) duen alde nagusia datuak prozesatzeko erabiltzen den unitate txikiena izanik: lehenengo bien kasuan hitza bera da; FastTexten kasuan, ordea, hitz bakoitza *n-grama* txikiagoetan banatzen da, horiek izanik unitate txikiena.

Errepresentazio estatikoek dituzten mugak gainditzeko helburuz (hitz polisemikoen prozesatzea, adibidez), beste errepresentazio mota garatuago bat sortu zen, hitz bakoitzaren errepresentazioa lortzeko esaldiko gainerako hitzak kontuan hartzen dituena: **testuinguruaren araberako errepresentazioak**.

Errepresentazio mota hauek, orokorrean, *Transformer* [Vaswani et al., 2017] izeneko arkitektura baten bidez lortzen dira, atentzio mekanismoan oinarritzen dena; honela, azken teknika hau da esaldiko hitz ezberdinen arteko erlazioak aztertzeaz arduratzen dena. *Transformerra*, laburki definitzeko, ikasketa sakoneko eredu bat da non irteera elementu bakoitza sarrera elementu bakoitzarekin konektatuta dagoen, beren arteko pisuak dinamikoki kalkulatuz duten loturaren arabera (prozesu hau bera da, zehazki, atentzioa). Arkitekturari dagokionez, bi mekanismo nagusitan banatzen da: kodetzailea, sarrerako testua irakurtzeaz arduratzen dena, eta dekodetzailea, ataza jakin baterako iragarpenak sortzen dituena.

Horretaz gain, *Transformerretan* oinarritutako ereduen artean, HP arloan azken urteotan izan den aurkikuntza garrantzitsuenetako bat aipatu behar da: BERT, *Transformerretako Kodeketa Bidirekzionalen Errepresentazioa* [Devlin et al., 2018]. Historikoki, aurretik plazaratutako eredu gehienek testua sekuentzialki irakurtzeko gaitasuna soilik zuten (eskuinetik ezkerrera, edo alderantziz), baina ez biak aldi berean; horregatik, BERTen ezaugarri nagusietako bat aldi berean bi noranzkoetan irakurri ahal izatea da (bidirekzionalitatea), *Transformerrak* emandako aukera zabalei esker. Honela, hitz bakoitzaren errepresentazio bektoriala kalkulatzeko orduan, alde bietara dituen hitzak har daitezke kontuan (hau da, testuinguru osoa).

Era berean, BERT aurre-entrenamendua erabiltzen duen teknika bat da, eta transferitzeko gaitasun altua da aurretik argitaratutako eredu bidirekzionaletatik (batez ere, LSTMetatik) bereizten duena; horretarako, etiketatu gabeko testu laua erabiltzen da, ingelesezko Wikipediatik eta Brown Corpusetik (biak osorik) lortutakoa. Honela, etiketatu gabeko testutik gainbegiratu gabe ikasten du, eta azkenean, ezagutza handiko oinarrizko geruza bat lortzen da gainontzeko atazak burutzeko abiapuntu gisa erabiliko dena; ondorioz, ebatzi nahi den atazaren araberako datuekin sistema birdoitzea besterik ez litzake faltako. Prozesu hau, transferentzia bidezko ikasketa (*transfer learning*) gisa definitzen da. Gainera, aurre-entrenamendurekin jarraituz, bi modu ezberdinetan egiten dela aipatu behar da:

- **Hizkuntza-eredu maskaraduna (*Masked Language Model*, MLM):** Hitz-sekuentziak ereduari sartu baino lehen, horietako %15 inguru *[MASK]* deituriko token batengatik ordezkatzen dira. Helburua testuinguruko informazioan oinarrituz, maskara bidez gordetako hitz horien benetako balioa iragartzea da.
- **Hurrengo esaldiaren iragarpena (*Next Sentence Prediction*, NSP):** Ereduak sarrera bezala bi esaldi ezberdin jasotzen ditu, helburua bigarrena lehenengoaren jarraipena den erabakitzea izanik. Entrenamenduan zehar, kasuen erdietan bi esaldiak ondoz ondokoak dira, eta beste erdietan, ordea, ez; azken hauetan, argi bereiz daiteke bi esaldiak loturarik ez dutela.

Erabilpenari dagokionez, BERT HP ataza desberdin askotan erabil daiteke, beti ere birdoitze prozesua egoki eginda (are gehiago, horietako gehienetan artearen egoerako emaitzak lortu ditu); batzuk nabarmentzekotan, generazio atazak (hala nola, galdera-erantzun sistemak edo esaldi iragarpenak) zein hizkuntza naturalaren interpretaziokoak (adibidez, hitzen desanbiguazioa edo sentimenduen sailkapena) aipatu beharko lirateke.

Azkenik, kontuan hartu behar da BERT kode irekikoa dela; hau da, edonorentzat eskuragarri dagoela sarean. Ondorioz, ikerketa talde ugarik eredu originalaren aldakuntza oso interesgarriak argitaratu dituzte, gainbegiratutako datuekin birdoitzearen teknika aplikatuz; honela, alde batetik, eredu efizienteagoak lor daitezke eta, beste aldetik, ataza jakin batzuetarako espezializatu daitezke testuinguruaren araberako errepresentazio zehatz batzuen bidez birdoituaz. Amaitzeko, hona hemen plazaratu diren aldakuntza garrantzitsuenetako batzuk, lehenengo hiruetan ereduaren ezaugarri batzuk editatzen direlarik (aurre-entrenamenduko prozesua edo parametro kopurua, adibidez), azken lauetan BERT bera datu-multzo ezberdinetan birdoitzen den bitartean (ataza jakin batzuetarako prestatuz):

- **RoBERTa** [Liu et al., 2019]: BERTen aldakuntza ezagunenetako bat, aurre-entrenamendua optimizatzea helburu duena; horretarako, maskaratze dinamikoa, NSP aurre-entrenamendu atazaren ezabapena, datu gehiagoren erabilpena eta *batch* tamainaren handitzea burutzen ditu.
- **ALBERT** [Lan et al., 2019]: BERTen aldakuntza sinplifikatuago bat (entrenamendu denbora asko murrizten duena), 110 milioi parametro eduki beharrean 12 milioi bakarrik dituena; sinplifikazio hori lortzeko, geruzek parametroak partekatzen dituzte beren artean.
- **DeBERTa** [He et al., 2020]: BERTen aldakuntza handiago bat, bere bertsio garatuenean 1.5 bilioi parametro edukiz; ezaugarri garrantzitsuenen artean, errepresentazio posizionalen bektoreak gehitu beharrean kateatu egiten dituela da.
- **patentBERT** [Lee and Hsiang, 2019]: Patente sailkapenean espezializatutako birdoitutako aldakuntza.
- **docBERT** [Adhikari et al., 2019]: Dokumentu sailkapenean espezializatutako birdoitutako aldakuntza.
- **bioBERT** [Lee et al., 2019]: Biomedikal hizkuntza errepresentaziorako birdoitutako eredua, testu biomedikoen meatzaritzarako prestatua.
- **sciBERT** [Beltagy et al., 2019]: Testu zientifikoekin lan egiteko birdoitutako eredua.

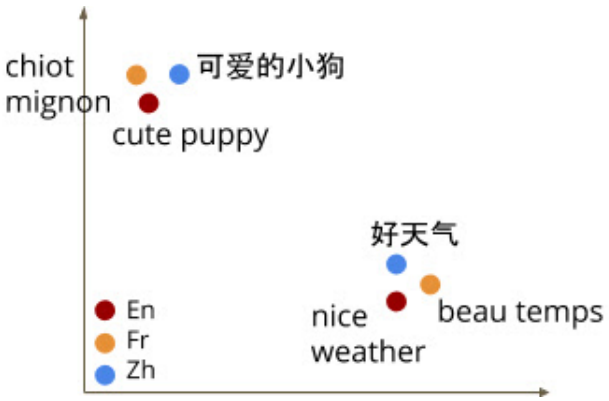
2.1.3 Hitzen errepresentaziorako eredu eleanitzak

Aurretik azaldutako ereduei erreparatzen badiegu, ikus dezakegu denak ingelesez lan egiteko prestatuta daudela, ezagutzen ditugun hizkuntza guztietatik baliabide gehien dituena, inolako dudarik gabe (eta alde handiarekin, gainera). Ondorioz, HP arloko ikerlari asko, ingelesarekin lortutako arrakasta ikusita, gainerako hizkuntzetara orokortzen saiatu dira.

Alde batetik, hizkuntza bakoitzerako eredu elebakarrak sortzeko aukera dago; hau ideia ona izan liteke baliabide handiak dituzten hizkuntzentzako, baina baliabide gutxikoentzako ez da oso erabilgarria, ez baitago nahikoa informazio eredua ongi entrenatzeko. Beste aldetik, ordea, hainbat hizkuntzetako datuak bateratu eta eredu bakar bat entrenatzeko aukera dago, **eleanitza** izango dena. Emaitzetan sartu ere egin gabe, hainbat abantaila ikus

daitezke hasieratik; adibidez, eredu bakarra entrenatu behar dela edo hizkuntza ezberdinetan idatzitako testuak prozesatu daitezkeela. Horretaz gain, eredu eleanitzen eraginkortasunari erreparatuta, onuradun nagusiak baliabide gutxiko hizkuntzak direla ikus daiteke, hobekuntza handiak lortuz eredu elebakarrarekin alderatuz. Argitaratutako lehen eredu eleanitza mBERT izan zen, eta beste hainbat eredu garatu ondoren, uneko eredu eleanitz nabarmenetako bat den XLM-R (azpiatal honen bukaeran azalduko da sakonki) gai izan da ingelesezko BERTek lortutako emaitzekin konparagarriak kontsidera daitezkeen balioak eskuratzeko.

Funtzionamenduari dagokionez, hizkuntza-eredu eleanitz gehienak ideia berdinean oinarrituta daude: sarrerako hizkuntza edozein dela ere, esanahi berdina duten hitzak errepresentazio bektorial berdina (edo ahalik eta antzekoena) eduki behar dute, 2.2 irudian ikus daitekeen bezala.



2.2 Irudia: Eredu eleanitz baten funtzionamenduaren errepresentazio grafikoa; ingelesez, frantsesez zein txinatarrez esanahi bera duten hitzak ahalik eta gertuen egotea da helburua, errepresentazio bektorial berdintsuak lortuz.

Iturria:
https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html

Behin kontzeptu nagusiak ulertuta, sarean eskuragarri dauden hiru eredu oso ezagun azalduko dira jarraian: mBERT, XLM eta XLM-R; horretaz gain, euskaraz erreferentea den IXAmBERT eredua ere aztertuko da. Kasu honetan, proiektuan erabili diren ereduak XLM-R eta IXAmBERT direnez, haietan sakonduko da gehien:

- **mBERT** [Devlin et al., 2018]: BERT originalarekin batera argitaratuta, 104 hizkuntza ezberdin hartzen ditu kontuan, ideia oso sinple bat aurkeztuz: originalaren mekanismo berdina erabiltzen du, baina hizkuntza ezberdinetako testuekin entrenatuaz

(dena Wikipediatik erauzitakoa) eta hizkuntza guztiek partekatzen duten hiztegi bat sortuz. Baliabide handi eta txikiko hizkuntzen datu kopurua oso ezberdina denez Wikipedian ere, txikientzako handiagotze teknikak erabiltzen dira (eta alderantziz) hizkuntza guztiak maila antzekoan errepresentatuta gera daitezen.

- **XLM** [Lample and Conneau, 2019]: *Transformerrean* oinarritutako sistema bat da eta, BERT bezalaxe, MLM bidez entrenatzen da. Horretaz gain, Itzulpenetan Oinarritutako Hizkuntza Eredua (*Translation Language Modeling*, TLM) ere erabiltzen da aurre-entrenamendua burutzeko, hizkuntza ezberdinentzako errepresentazio antzekoak ikastea helburu duena. Azken prozesu hau nahiko sinplea da: esaldi berbera bi hizkuntza ezberdinetan eman eta token batzuk maskaratzen dira (bi hizkuntzetakoak); ondoren, esaldiak paraleloak direnez, bi hizkuntzetako informazioa erabiltzen da maskaratutako tokenen iragarpenak egiteko, hizkuntza-arteko ikasketa burutuaz. Entrenamenduko datu-multzoei dagokienez, MLMrako Wikipediatik erauzten da informazioa XNLIko [1] 15 hizkuntzetarako (transferentzia linguistikoa zein hizkuntza-arteko esaldi-sailkapen atazak ebaluatzeko datu-multzo ezaguna, MultiNLI korpuseko esaldiak eskuz itzulita lortu dena); TLMrako, ordea, data multzo ezberdinak erabiltzen dira hizkuntzaren arabera, hizkuntza-pareetako esaldi paraleloen beharra baitu (hizkuntza-pare batzuetarako eskuratzeko zailak izan daitezkeenak).

- **XLM-RoBERTa (XLM-R)** [Conneau et al., 2019]: XLMren aldakuntza bat da, MLM eta TLMren bidez entrenatu ordez lehenengoa soilik erabiltzen duena; hain zuzen ere, hortik datorkio RoBERTa [Liu et al., 2019] izena, eredu elebakar honen entrenamendu prozesua horrelakoa baita. XLM-R ereduaren ezaugarri nagusia entrenamendurako erabiltzen duen datu kopurua da: 100 hizkuntza ezberdinentzako gainbegiratu gabeko testua erauzten du Common Crawletik [2], guztira 2.5TB testu lortuz. Beste ezaugarri azpimarragarri bat hiztegiaren dimentsioa da, 250.000 token ezberdin jasotzeko gaitasuna baitu; konparazio bat egite aldera, mBERTek 110.000 token besterik ezin ditu prozesatu.

  Dimentsionaltasunarekin lotutako ezaugarriak alde batera utzita, XLM eta XLM-R ereduen arteko alde handiena azkenekoa guztiz gainbegiratua dela da, lehenengoak esaldi-pare paraleloak behar dituen bitartean (hizkuntza batzuetarako, eta batez ere eskala handian, lortzeko zailak izanik).

[1] https://github.com/facebookresearch/XNLI
[2] https://commoncrawl.org

XLM-R ereduaren eraginkortasuna ebaluatze aldera, [Conneau et al., 2019] artikuluan aipatutako XNLI datu-multzoan testatzen da. Bertan, bai mBERTek eta bai XLMk baino emaitza hobeak lortzen ditu, eta, hain zuzen ere, hori izan da proiektuan modelo hau aukeratzearen arrazoietako bat. Horretaz gain, ezin aipatu gabe utzi GLUE *benchmarkean* [3] (HP arloko ingelesezko erreferentzia bat) egindako ebaluazioa, non XLM-R gai den eredu elebakarrekin lehiatzeko, nahiz eta 100 hizkuntza ezberdinekin lan egiteko prestatuta egon.

- **IXAmBERT** [Otegi et al., 2020]: Ingelesa, gaztelania eta euskararako aurre entrenatutako eredu eleanitz baten aurrean gaude, bereziki ezaguna dena euskara bezalako baliabide gutxiko hizkuntza batean lortutako emaitza onengatik.

  Eredu hau sortzearen arrazoia, sarean eskuragarri dauden ereduek euskarari behar adina garrantzi ez ematea da; hau da, hizkuntza askorekin lan egiteko prest daudenez, euskarak beste hizkuntza handiagoek baino errepresentazio txikiagoa du. Ondorioz, BERTeus [Agerri et al., 2020] oinarri hartuta eta bere konfigurazioa mantenduaz, eredu berri honek aipatutako hiru hizkuntzetara soilik mugatzen du bere jarduna.

  Entrenamendurako datu-multzoei dagokienez, euskararen kasuan 2.1 taulan erakutsitako informazio-iturriak erabiliz sortu da corpusa.

Iturria	Testu mota	Token kopurua
Euskal Wikipedia	Entziklopedia	35M
Berria egunkaria	Albisteak	81M
EiTB	Albisteak	28M
Argia aldizkaria	Albisteak	16M
Herri aldizkariak	Albisteak	224.6M

2.1 Taula: IXAmBERT entrenatzeko erabili diren datu euskarazko datu-multzoak, bakoitza zenbat tokenez osatuta dagoen adieraziz.

Ikus daitekeen moduan, Wikipediak eta sareko baliabide ezberdinetatik erauzitako albisteek osatzen dute entrenamendurako datu-multzoa. Gaztelania eta ingelesaren kasuan, ordea, Wikipediako informazioa soilik erabiltzen da; hala ere, kontuan hartu behar da ingelesezko Wikipedia 80 aldiz handiagoa dela (2.5G token) euskarazkoa baino, eta gaztelaniazkoa 20 aldiz handiagoa (650M token).

[3] https://gluebenchmark.com/

Eredua testatze aldera, egileek Elkarrizketa Bidezko Galdera Erantzun (CQA, *Conversational Question Anwering*) ataza aukeratu zuten; honela, artikuluan erakutsitako emaitzetan oinarrituta, ataza honetan IXAmBERTek mBERTek baino errendimendu hobea lortzen du, eta ondorioz, hori izan da proiekturako aukeratu izanaren arrazoia.

2.2 Esaldien errepresentaziorako eredu elebakarrak

Behin hitz-errepresentazioaren historia eta gaur eguneko eredu garrantzitsuenak aztertuta, urrats bat haratago joan eta haren ideia nagusietan oinarrituta dagoen teknika garatuago bat erakutsi behar da: **esaldi-errepresentazioak**. Izenak argi iradokitzen duen bezala, kasu honetan esaldi mailan egiten da lan, errepresentazio bektorialak lortuz haietako bakoitzerako; honek ikaragarrizko aukera sorta zabala ematen du, esaldien arteko erlazioak modu ezberdin askotara prozesatu zein ebaluatzeko aukera baitago.

Esaldi-errepresentazioak lortzeko modu asko daude, baina horien artean ezagunena eta sinpleena hitz-errepresentazio eredu baten bidez esaldiko hitz guztiak kodetu eta, azkenik, bektore guztien batezbestekoa ateratzea da; gainera, konplexutasun handirik ez duen teknika bat izateaz gain, oso erabilgarri zein eraginkorra da ataza askotarako.

Ereduei dagokienez, nahiz eta ataza nahiko berria den, aukera ezberdin ugari garatu dira azken urteotan; esate baterako, ezin dira aipatu gabe utzi Doc2Vec [Le and Mikolov, 2014], lehenago aztertutako Word2Vec ereduaren moldaera bat, edo InferSent [Conneau et al., 2017], errepresentazioak era gainbegiratuan entrenatutako sare neuronal batzuen bidez lortzen dituena (esaldien arteko erlazio semantikoak aztertzeko helburuz). Hala ere, aipatutako ereduak oso garrantzitsuak diren arren, beste bi artikulu daude nahitaez sakonki aztertu behar direnak: USE eta sBERT; gainera, azkeneko hau izango da proiektuan zehar erabiliko den ereduaren oinarri, ondorioz azterketa are garatuago bat egingo da haren inguruan.

2.2.1 USE (*Universal Sentence Encoder*)

Sarrera txiki batekin hastearren, USE [Cer et al., 2018] bi eredu ezberdinez osatuta dago, hainbat ataza ezberdinetan ikasten dutenak generikotasun altuko errepresentazioak lortzeko. Ondoren, transferentzia bidezko ikasketaren bidez, errepresentazio bektorial hauek oso erraz molda daitezke HP arloko ataza ezberdinetara.

Esan bezala, bi eredu ezberdin azaltzen dira, bata *Transformerrean* eta bestea *Deep Averaging Network* batean oinarrituta; honela, bakoitzaren errendimendua (eta baliabideen erabilpena) aztertzen da, beren artean eta hainbat hitz-errepresentazio ereduekin alderatuz. Horretarako, bi ataza mota ezberdin erabiltzen dira, sailkapenean eta antzekotasunean oinarritzen direnak. Hasteko, ereduen ezaugarri garrantzitsu batzuk erakusten dira, beren arteko ezberdintasun nagusiak ulertzeko:

- ***Transformerran* oinarritutakoa** [Vaswani et al., 2017]: Eredu honek *Transformerraren* zati kodetzailea erabiltzen du, atentzioaz baliatuz testuingurua kontutan hartzen duten hitz-errepresentazioak sortzeko. Ondoren, errepresentazio bektorial hauek luzera zehatz bateko bektore bakar batera pasatzen dira, elementukako batura eginaz. Honela, ereduak PTB (*Penn TreeBank*) erara tokenizatutako esaldi bat hartzen du sarrera gisa, eta 512 dimentsioko bektore bat lortzen du esaldiaren errepresentazio moduan.

  Entrenamendu prozesuari dagokionez, aipatu bezala, hainbat ataza ezberdinetan burutzen da: hasteko, ausazko testua erabiliz gainbegiratu gabeko ikasketa helburu duen ataza batean; ondoren, galdera-erantzun elkarrizketan oinarritutako beste ataza batean, analizatutako elkarrizketa datuak kontuan hartzeko helburuz; eta, azkenik, gainbegiratutako datuak erabiliz sailkapen ataza batean.

  Amaitzeko, eredu honen bidez lortzen dira emaitza onenak, baina errekurtso gehien behar dituen dituena ere bada, batez ere esaldi luzera handitzen denean.

- **DAN (*Deep Averaging Network*) batean oinarritutakoa** [Iyyer et al., 2015]: Eredu honetan, hasteko, hitz-errepresentazioen zein *2-gramen* batezbestekoa egiten da, ondoren DNN (*Deep Neural Network*) bati pasaz esaldi-errepresentazioak lortzeko. *Transformerraren* moduan, PTB erara tokenizatutako esaldi bat hartzen du sarrera gisa eta 512 dimentsioko bektore bat eraikitzen du; gainera, entrenamendu prozesua ere oso antzekoa da.

  Kasu honetan, nahiz eta emaitzak apurtxo bat kaxkarragoak diren, exekuzio-denbora linealki proportzionala da sarreraren luzerarekiko, denbora asko aurreztuz esaldiak oso luzeak direnean.

Entrenamendurako gainbegiratu gabeko datuak sareko hainbat iturrietatik eskuratu dituzte, nagusia Wikipedia izanik; horretaz gain, gainbegiratutako datu batzuk ere erabili dituzte, SNLI [4] *(Stanford Natural Language Inference)* corpusetik erauzi direnak.

[4] https://nlp.stanford.edu/projects/snli/

2.2.2 SBERT (SentenceBERT)

Izenetik iradoki daitekeen bezala, artikulu honetan [Reimers and Gurevych, 2019] BERT ereduaren moldaketa bat aurkezten da, esaldi mailan lan egiteko prestatua; horretaz gain, RoBERTa ereduarekin ere prozesu berbera burutzen da. Egileek azaltzen duten moduan, erregresio atazak gauzatzeko sistema berberak bi esaldiak prozesatu behar ditu, konputazionalki oso garestia dena; esate baterako, esaldi baten antzekoena bilatu nahi badugu 10.000 esaldiko corpus batean, BERTek 65 ordu inguru beharko lituzke. Ondorioz, BERT horrelako atazak burutzeko prestatuta ez dagoenez, haren moldaketa bat proposatzen dute egileek, sare siamdar eta hirukoitzetan oinarritzen dena esaldien errepresentazioak lortzeko; modu honetan, aurreko adibidera bueltatuz, 65 orduko lana 5 segundotara jaisteko aukera dago.

Arkitekturari dagokionez, ezaugarri garrantzitsuenetako bat BERTek emandako irteerari aplikatutako *pooling* operazioa da, hitz bakoitzaren errepresentazio independentetik luzera jakin bateko esaldi-bektore bakar bat sortzea ahalbidetzen duena, hitz-bektore bakoitzaren informazio semantikoa mantenduz. Hiru aukera ezberdin probatu dira operazio hau burutzeko: [CLS] tokena erabilita, hitz-bektoreen (dimentsio bakoitzeko) batezbestekoa kalkulatuta edo dimentsio horietako bakoitzeko maximoa jasota; azkenik, hainbat froga egin ondoren, ikusi da denetatik eraginkorrena batezbestekoa dela. Arkitekturarekin jarraituz, aipatu bezala, sare siamdarrak eta hirukoitzak sortu dira, burutu nahi den atazaren arabera aldatzen direnak:

- **Sailkapena:** Bi esaldiak kateatzen dira $|u - v|$ erabiliaz (hainbat aukeren artean, emaitza onenak eman dituena), eta ondoren $W_t \in \mathbb{R}^{3n \times k}$ pisuarekin biderkatzen da, non n esaldi-bektorearen dimentsioak eta k sailkapeneko etiketa kopurua diren; azkenik, lortutako emaitzari *softmax* sailkatzailea aplikatzen zaio, ondorengo formulan zein 2.3 irudian ikus daitekeen bezala.

$$o = \text{softmax}(W_t(u, v, |u - v|)) \tag{2.2}$$

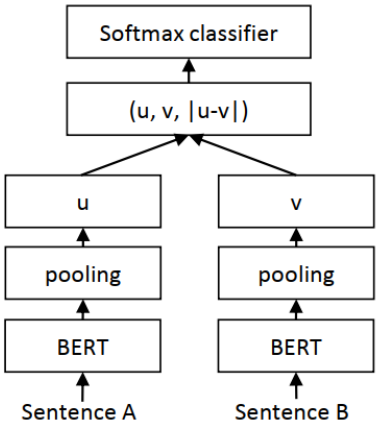
non:

u = lehen esaldiaren errepresentazioa

v = bigarren esaldiaren errepresentazioa

W_t = pisu matrizea

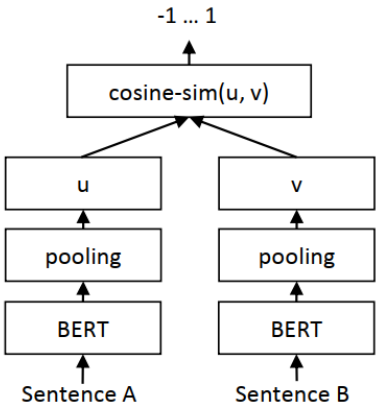
Erroreari dagokionez, entropia gurutzatuko galera-funtzioa optimizatzea da helburua.



2.3 Irudia: SBERT arkitektura helburua sailkapen ataza bat ebaztea denean. Esaldiak BERT bidez prozesatu eta irteerari *pooling* teknika aplikatuta *u* eta *v* esaldi-bektoreak lortzen dira; ondoren, $|u - v|$ kalkulatu eta sailkapena burutzen da *softmax* aplikatuta.

Iturria: https://arxiv.org/pdf/1908.10084.pdf

- **Erregresioa:** Kasu honetan, 2.4 irudian erakusten den bezala, *u* eta *v* errepresentazioen arteko kosinu antzekotasuna kalkulatzen da; oraingoan, galera-funtzioa, batezbesteko errore koadratikoa izango da, eta helburua, hura minimizatzea.



2.4 Irudia: SBERT arkitektura helburua erregresio ataza bat ebaztea denean. Esaldiak BERT bidez prozesatu eta irteerari *pooling* teknika aplikatuta *u* eta *v* esaldi-bektoreak lortzen dira; ondoren, kosinu antzekotasuna aplikatzen da bi esaldien arteko antzekotasun semantikoa neurtzeko.

Iturria: https://arxiv.org/pdf/1908.10084.pdf

- **Hirukoitza:** *a* aingura esaldi bat, *p* esaldi positibo bat eta *n* esaldi negatibo bat edukita, helburua *a* eta *p* esaldien arteko distantzia *a* eta *n* esaldien artekoa baino txikiagoa izatea da; matematikoki, hurrengo galdera-funtzioa minimizatzen da.

$$\max(||s_a - s_p|| - ||s_a - s_n|| + \varepsilon, 0) \tag{2.3}$$

non:

s_x = *a/n/p* esaldien errepresentazio bektoriala

$||\cdot||$ = distantzia metrika

ε = marjina

ε marjinak s_p s_a-tik gutxienez s_n-tik baino ε gertuago egotea ziurtatzen du. Esperimentuetarako metrika distantzia euklidearra da eta marjinaren balioa $\varepsilon = 1$.

Behin arkitektura mota ezberdinak ulertuta, entrenamenduari buruzko hainbat ezaugarri aipatzen dira. Prozesu hau burutzeko, SNLI eta MultiLNI [5] datu-multzoak erabiltzen dira, helburua premisa baten arabera hipotesi jakin bat egia, kontraesana edo neutrala den ondorioztatzea izanik. Lehena 570.000 parekatutako esaldi-parez osatuta dago, premisak Flickr30k plataformako irudi-oinak izanik eta hipotesiak eskuz sortuta egonik; bigarrena, ordea, 430.000 esaldi-parek osatzen dute eta lehenengoaren aldaera bat da, 10 genero ezberdinetako idatzizko zein ahozko testuaren bidez eratu dena.

Ebaluazioaren atalera igaroaz, artikuluan esperimentu ugari egiten dira, hiru arkitektura motak probatzeko helburuz; hala ere, gure proiektuan antzekotasuna (erregresioa) erabili dugunez, haren inguruko emaitzak dira gehienbat kontuan hartu behar ditugunak. Horregatik, SBERTek ataza honetan dituen balioak USEk lortutakoekin alderatuz gero, lehenengoaren errendimendua dezente hobea dela ikus daiteke; eta, hain zuzen ere, hori izan da SBERT proiektuan erabiltzeko erabakia hartu izanaren arrazoia.

2.3 Esaldien errepresentaziorako eredu eleanitzak

Jada ikusi dugu hitz-errepresentazio eredu elebakarretatik eleanitzetara igarotzeko prozesua, ondoren hitz mailatik esaldi mailara pasatzekoa, eta orain, amaitzeko, azken urratsa geratzen zaigu: esaldi-errepresentazio eredu elebakarretatik eleanitzetara igarotzea.

[5] https://cims.nyu.edu/~sbowman/multinli/

Esaldi-errepresentazio ereduak eleanitz bihurtzeko prozesua hitzarentzat egindakoaren oso antzekoa da kasu gehienetan, semantikoki antzekoak diren hizkuntza ezberdinetako esaldiak espazio bektorial berean kokatzen saiatuz. Hori horrela, hainbat eredu daude euskararako baliagarriak direnak, hauek izanik garrantzitsuenak:

- **LASER** [Artetxe and Schwenk, 2018]: Hizkuntzarekiko agnostiko diren esaldi-errepresentazioak kalkulatzen ditu, esaldi-pare paraleloetan eta partekatutako hiztegi batean oinarrituta (ataza ezberdinetan erabil daitekeena).
- **LaBSE** [Feng et al., 2020]: Edozein hizkuntzetarako balio duten BERT esaldi-errepresentazioak sortzen ditu, entrenamenduan ikusi ez dituen hizkuntzetara (edo oso esaldi-pare gutxirekin entrenatutakoetara) orokortzeko gaitasuna edukiz.

2.4 Antzekotasun semantikoa

Ataza honen helburua bi hitz edo esaldi zenbateraino diren antzekoak definitzea da; ikusi dugun moduan, prozesu hau aurrera eramateko moduetako bat errepresentazioak lortu eta beren artean, antzekotasun metrika bat erabiliz, konparatzea da. Ondorioz, atal hau bi zati nagusitan banatuko da, lehenik metrika ezberdinak azalduz (batez ere, proiektuan erabilitakoa) eta azkenik atazaren inguruko informazioa emanez.

2.4.1 Metrikak

Aipatutako konparazioa burutzeko modu ezberdin asko daude, eta ondorioz, ataza bakoitzerako egokiena den metrika aukeratu behar da; adibidez, distantzia euklidearra edo *Jaccarden* indizea bezalako teknikak eskuragarri daude, baina ezagunena eta gure proiekturako eraginkorrena dena kosinu antzekotasuna da:

- **Distantzia euklidearra:** Bi esaldi-errepresentazioen (u eta v) bukaeren arteko distantzia motzena kalkulatzen du, Pitagorasen teorian oinarrituz. Emaitzak distantzia adierazten du, 0 izanik bi bektoreak berdinak direnean:

$$\sqrt{\sum_{i=1}^{k}(u_i - v_i)^2} \tag{2.4}$$

- ***Jaccarden* indizea:** Honen bidez bi errepresentazioek zenbat elementu dituzten komunean neurtzen da; ondorioz, geroz eta elementu komun gehiago, geroz eta antzekoagoak dira. 0 eta 1 arteko eskala batean neurtzen da, 1 izanik bi bektoreak berdinak direnean:

$$\frac{|u \cap v|}{|u \cup v|} = \frac{|u \cap v|}{|u| + |v| - |u \cap v|} \tag{2.5}$$

- **Kosinu antzekotasuna:** Aipatu bezala, honako hau da ataza gehienetarako erabiltzen dena (baita gure proiekturako ere), bere propietateak oso baliotsuak direlako. Bertan, bi errepresentazioen arteko angelua kalkulatzen da; honela, orientazio berbera badaukate, 1 izango da antzekotasunaren balioa:

$$\cos(\theta) = \frac{u \cdot v}{||u||\,||v||} \tag{2.6}$$

  Garrantzitsua da kontuan hartzea metrika honek ez digula bektoreen magnitudearen inguruko informaziorik ematen, esan bezala, errepresentazioek orientazio berdina duten adierazten du soilik. Hala ere, propietate hau oso baliagarria zaigu antzekotasun semantikoa neurtzeko, errepresentazio bateko osagai ezberdinek (edo beren arteko konbinazio linearrek) hitz edo esaldiaren informazio semantikoa jasotzen baitute. Esate baterako, demagun $u = [-1, 2, -3]$ eta $v = [-3, 6, -9]$ bektoreak ditugula, non $v = 3u$ den; magnitude aldetik oso ezberdinak badira ere, modu berean orientatuta daude, osagaiak zuzenki proportzionalak baitira beren artean. Ondorioz, kosinu antzekotasunak erlazio hau antzeman dezake arazorik gabe; gainera, kasu honetan emaitza 1 da, hitzak edo esaldiak zentzu bera dutela adieraziz.

  Ez da berdina gertatzen, ordea, lehen aipatutako euklidear distantziarekin. Bi bektoreen arteko distantzia 7.48 unitatekoa izango litzateke metrika hau erabilita, oso lan erraza izanik antzeko distantziara baino orientazio ezberdinean dagoen beste c bektore bat aurkitzea; honela, c bektorearen esanahi semantikoak ez luke zerikusirik edukiko beste biekin, baina sistemak erlazionatu egingo lituzke distantzia berdinera daudelako. Horrexegatik, metrika honek hainbat arazo ditu ataza mota honekin lan egiterako orduan, kosinu antzekotasunaren bidez erraz eta eraginkorki gainditzen direnak.

2.4.2 Ataza

Gure proiektuaren helburua esaldiak konparatzea denez, garrantzitsua da SemEval [6] esaldi-errepresentazioen ebaluaziorako *workshop* serie ezagunaren barruan dauden STS (*Semantic Textual Similarity*) atazak aztertzea. [Cer et al., 2017] artikuluan azaltzen den bezala, 2012 eta 2017 bitartean urtero argitaraturiko erronka ezberdinak dira, helburua esaldien arteko antzekotasun maila neurtzea dutenak; horretarako, sistemak 0 eta 5 arteko puntuazio bat eman behar dio (2.5 irudiko jarraibideak kontuan hartuz) esaldi-pare bakoitzari, eta puntuazio horien eta giza epaiaren arteko Pearson [7] koefizientearen bidez egingo da ebaluazioa.

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

2.5 Irudia: STS ebaluatzeko erabilitako jarraibideak (adibide batzuez lagunduta), 0 eta 5 arteko balioak emanez bi esaldien antzekotasun mailaren arabera.

Iturria: https://www.ixa.eus/sites/default/files/dokumentuak/8880/SemEval2017_STS_June27.pdf

Ataza hauen bidez, gainera, mota orotariko ereduak eta teknikak probatzeko ingurune bat sortu nahi da, esaldien antzekotasunaren arloan artearen egoera zehazteko ere balioko duena.

Horretaz gain, STS atazetan ebaluatutako sistemek ondoren oso emaitza onak lortu dituzte

[6] https://semeval.github.io/
[7] https://eu.wikipedia.org/wiki/Korrelazio-koefiziente

aplikazio mota ezberdinetan (esate baterako, itzulpen automatikoan edo galdera-erantzun sistemetan), esaldien zentzua eta haien arteko erlazioak ulertzea beharrezkoa baita edozein HP lanetarako.

Bide batez, azpimarratu behar da STS atazek egindako ekarpena handia izan dela, aurretik ez baitzegoen baliabide gehiegirik era honetako problemak ebaluatzeko; gainera, sortutako datu-multzoak oso aberatsak dira, domeinu ezberdin askotatik erauziak baitira, atazei konplexutasun handiagoa emanaz eta ikertzaileentzako are erronka interesgarriago bat bihurtuaz.

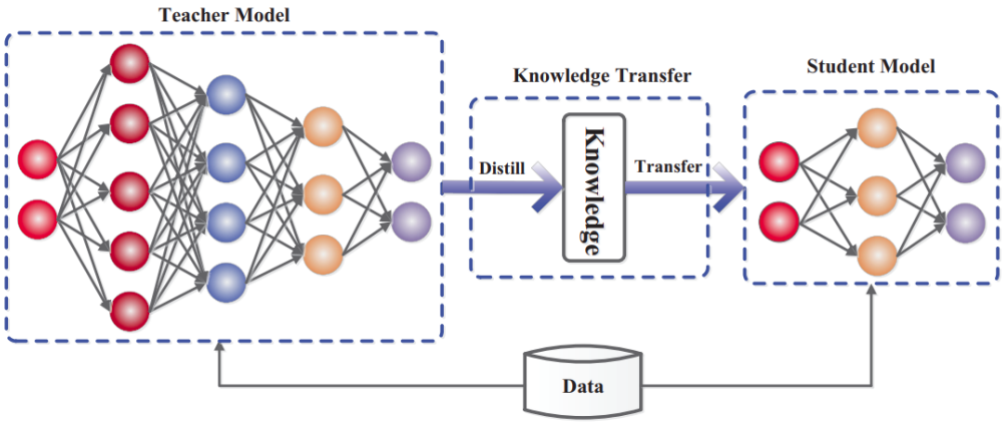
Azkenik, 2012 eta 2017 bitartean urtero plazaratutako datu-multzoez gain, STS *benchmark* [8] corpusa ere argitaratu da; hau osatzeko, ingelesezko adibide jakin batzuk aukeratu dira (edozein urtetakoak), argazki-oinetako testuetatik, albisteen titularretatik eta foro batzuetatik erauziak izan direnak. Sarean apurtxo bat arakatuz gero, erraz ikus daiteke corpus honek indar handia hartu duela antzekotasun semantikoaren arloan, artikulu garrantzitsu askotan erabili izan baita eredu edo teknika berriak probatzeko.

2.5 Distilazioa

Lehenago aipatutako esaldi-errepresentazio eredu eleanitzen alternatiba gisa, hauek lortutako errendimendua hobetzeko teknika ezberdinak erabiltzeko aukera dago (batez ere baliabide gutxiko hizkuntzentzako); ondorioz, proiektu honen helburua **distilazioa** oinarri gisa hartu eta euskaraz emaitza lehiakorrak lortuko dituen sistema bat garatzea izan da. Hori horrela, atal honetan distilazioaren inguruko azalpen sakon bat emango da, proiektua garatzerako orduan garrantzi handia izan duen artikulu bat aztertzeaz gain.

Honela, distilazioaren inguruko kontzeptuak barneratzen hasteko, [Gou et al., 2020] artikuluan egindako azalpenak erabiliko dira; bertan, teknika zertan datzan argi erakusten da eta hainbat sailkapen egiten dira distilazio moten eta moduen arabera. Hasteko, distilazioa era labur batean definitzen du: eredu indartsu batetik (irakaslea) beste hain ez garatu batera (ikaslea) ezagutza transferitzean datzan metodoa. Beste hitz batzuetan esanda, ikaslea irakaslea imitatzen saiatuko da, jakinda azken honek sortutako errepresentazioen kalitatea oso altua izango dela sakonki entrenatua izan baita.

[8] https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

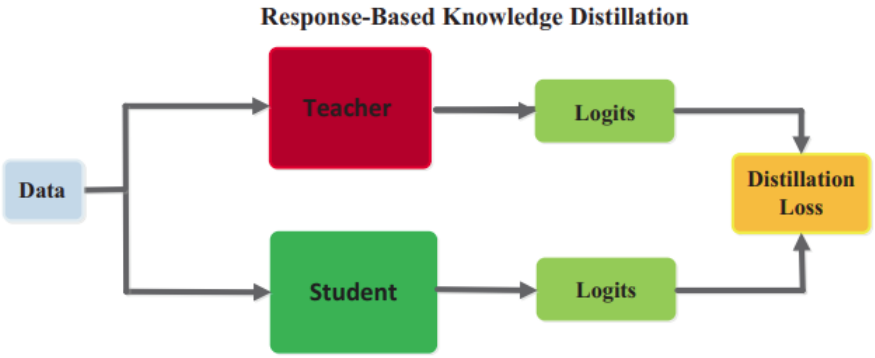


2.6 Irudia: Distilazio teknikaren irudikapen grafikoa. Bertan, irakasleak bere ezagutza transferitzen dio ikasleari distilazio bidez; eredu bakoitzaren arkitekturari begiratuta, irakaslearena garatuagoa dela ikus daiteke, geruza kopuruak adierazten duten bezala.

Iturria: https://analyticsindiamag.com/a-beginners-guide-to-knowledge-distillation-in-deep-learning/

Horretaz gain, artikulu honetan hiru distilazio mota nagusi bereizten dira, ikaslea imitatzen saiatzen den irakaslearen atalaren arabera:

- **Erantzunean oinarritutako distilazioa:** 2.7 irudian ikus daitekeen bezala, ikaslea irakaslearen iragarpena (azken geruzako irteera) imitatzen saiatzen da; hau da, errepresentazio bektorial bera lortzea da jomuga. Mota ezberdinen artean, ohikoena hau da.
- **Ezaugarrietan oinarritutako distilazioa:** Sakon entrenatutako irakasle batek informazio interesgarria du tarteko geruzetan ere, ezaugarri jakin batzuen artean diskriminatzen ikasiz; hain zuzen ere, ezaugarri espezifiko horiek ondoren ikaslea entrenatzeko erabil daitezke. Honela, helburua ikasleak irakaslearen ezaugarrien aktibazio berberak ikastea da.
- **Erlazioetan oinarritutako distilazioa:** Aurreko bi kasuetan geruza jakin batzuetako irteera imitatzen saiatzen da ikaslea; kasu honetan, ordea, geruza ezberdinen edo datu-multzoen arteko erlazioak aztertzen dira.



2.7 Irudia: Erantzunean oinarritutako distilazioa grafikoki azalduta; bertan, irakasleak zein ikasleak egindako iragarpenen (*logitak*) arteko galera-funtzioa kalkulatzen da, hura minimizatzeko helburuz; hau da, lortutako errepresentazioak ahalik eta antzekoenak izatea saiatzen da.

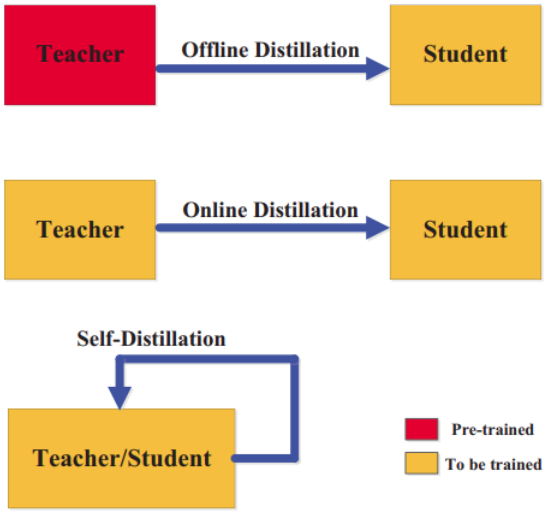
Iturria: https://analyticsindiamag.com/a-beginners-guide-to-knowledge-distillation-in-deep-learning/

Azkenik, artikulu honekin amaitzeko, distilazio moduaren araberako sailkapen bat egiten da, hiru mota nagusi bereiziz (2.8 irudian ikus daitekeen bezala).

- ***Offline* distilazioa:** Kasu honetan, ikaslea gidatuko duen irakaslea aurre-entrenatuta dago; hau da, ikaslea soilik entrenatuko da atazarako datu-multzoan. Ohikoena da, eta gainera, erabiltzeko oso erraza izateaz gain, irakasleak edonorentzat eskuragarri daude sarean.

- ***Online* distilazioa:** Oraingoan, irakasle eta ikasle ereduak aldi berean eguneratzen dira, ezagutza distilatzeko prozesu osoa entrenagarria bihurtuz. Nahiz eta *offline* distilazioa baino garestiagoa den, kalitate altuagoko errepresentazioak eraikitzeko gai da irakaslea; ondorioz, batez ere gaitasun eta errendimendu handiko irakasle eredurik ez denean, *online* distilazioa aukera ona izan daiteke.

- **Autodistilazioa**: Distilazio modu honetan eredu berbera erabiltzen da irakasle eta ikaslerako. Bertan, geruza sakonenetako informazioa azalenekoetara distilatzen da; horretaz gain, hasierako epoketako ezagutza ere bukaerakoetara pasa daiteke ikaslea entrenatzeko helburuz.



2.8 Irudia: Hiru distilazio modu nagusien errepresentazio grafikoak. *Offline* eran aurre-entrenatutako irakasle bat erabiltzen da, *online* moduan irakasleak ikaslearekin batera ikasten duen bitartean; azkenik, autodistilazioan, eredu bakarra erabiltzen da, bere buruaz elikatzen dena.

Iturria: https://analyticsindiamag.com/a-beginners-guide-to-knowledge-distillation-in-deep-learning/

2.5.1 Distilazioa esaldi-errepresentazio eleanitzak lortzeko

Behin distilazioko kontzeptu nagusiak ulertuta, proiektuaren abiapuntu eta oinarri izan den artikulua azalduko da: *Esaldi-errepresentazio elebakarrak eleanitz bihurtuz ezagutza distilazioa erabiliz* [Reimers and Gurevych, 2020]. Horretaz gain, egileek erabilitako inplementazioa [9] ere eskuragarri dago, oso baliagarria izan dena eta kodea hutsetik idazten hasi beharra ekidin duena.

Izenburuak iradokitzen duen bezala, artikuluak eredu elebakar indartsu batetik (irakaslea) abiatuta beste eredu eleanitz eraginkor bat (ikaslea) nola sortu erakusten du, distilazio teknikan oinarrituta. Honela, ideia nagusia irakasleak baliabide handiko hizkuntzan (ingelesez) errepresentazioak sortzea da, eta ondoren ikasleak haiek imitatuko ditu baliabide gutxiko hizkuntzara itzulitako esaldia espazio bektorial berberean mapeatzen saiatuz; ondorioz, erantzunean oinarritutako distilazioa burutzen dela adieraz dezakegu. Gainera, emaitzei erreparatuta, ikus daiteke abantaila asko dituen teknika bat dela:

[9] https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/multilingual/make_multilingual.py

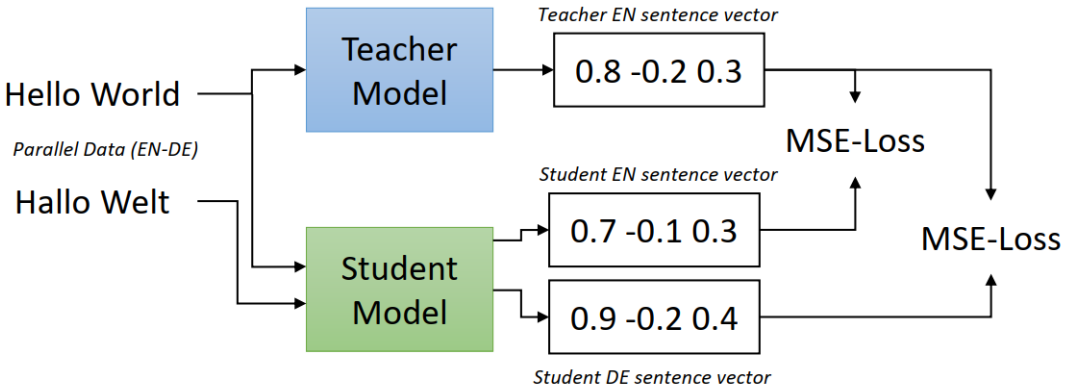
- Edozein hizkuntzetara orokortzeko gaitasuna dauka; horretaz gain, entrenamendurako oso esaldi-pare gutxi edukita ere, emaitza lehiakorrak lor daitezke. Ondorioz, oso aukera ona da baliabide gutxiko hizkuntzentzat.
- Espazio bektorialaren propietateak konfiguratzeko gai da, ataza bakoitzak dituen ezaugarrien arabera.
- *Hardware* aldetik ere entrenamendurako bete beharreko eskakizunak irisgarriak dira, baliabide indartsuenen beharrik gabe.

Entrenamenduaren inguruan xehetasun gehiago ematearren, alde matematiko batetik aztertuko da, ilustrazio baten bidez grafikoki erakusteaz gain (2.9 irudia); honela, hurrengo lerroetan azaldutako kontzeptuak argi eduki behar dira entrenamendua ongi ulertu ahal izateko.

- $M \rightarrow$ *Teacher Model* (irakaslea), elebakarra, ingelesezko errepresentazioak sortuko dituena.
- $M' \rightarrow$ *Student Model* (ikaslea), eleanitza, irakasleak sortutako errepresentazioak imitatzen saiatuko dena.
- $[(s_1, t_1), ..., (s_n, t_n)] \rightarrow$ Esaldi-pare multzo bat, non s_i ingelesezko esaldia den eta t_i baliabide gutxiko hizkuntzara itzulpena.
- **Helburua**:
  - $M'(s_i) \approx M(s_i) \rightarrow$ Ikasleak sortutako ingelesezko esaldiaren errepresentazioa eta irakaslearena (ingelesezkoa ere) ahalik eta antzekoenak izatea.
  - $M'(t_i) \approx M(t_i) \rightarrow$ Ikasleak sortutako baliabide gutxiko hizkuntzaren esaldiaren errepresentazioa eta irakaslearena (ingelesezkoa) ahalik eta antzekoenak izatea.

Honela, *batch* tamaina jakin bat zehaztuta (B), helburua batezbesteko errore koadratikoa minimizatzea da, hurrengo formulan adierazten den bezala:

$$\frac{1}{|B|} \sum_{j \in B} \left[(M(s_j) - M'(s_j))^2 + (M(s_j) - M'(t_j))^2 \right] \tag{2.7}$$



2.9 Irudia: Distilazio bidezko entrenamenduaren errepresentazio grafikoa. *Teacher Modelak* (irakasleak) ingelesezko esaldia soilik prozesatzen du, *Student Modelak* (ikasleak) bi hizkuntzak lantzen dituen bitartean; honela, ikasleak lortutako bi errepresentazioak irakaslearenarekin alderatzen dira (independenteki), helburua batezbesteko errore koadratikoa minimizatzea izanik.

Iturria: https://arxiv.org/pdf/2004.09813.pdf

Horretaz gain, esperimentuen atalera pasa aurretik, erabilitako datu-multzo ezberdinak aipatzen dira; gehienak, OPUS [10] [Tiedemann, 2012] izeneko webgunetik ateratakoak dira, non hizkuntza-pare askotarako datu paraleloak eskaintzen diren.

Esperimentuei dagokienez, autoreek eredu konbinazio jakin bat erabili dute batik bat froga gehienak burutzeko: biak ala biak aurretik azaldu eta sakonki aztertu ditugu, irakasle gisa ingelesezko SBERT eta ikasle gisa XLM-R erabiltzen baitira. Horretaz gain, *offline* distilazioa egiten da, irakaslea aurre-entrenatuta baitago (ez zaizkio berezkoak baino datu gehiago ematen). Egindako probei dagokienez, gainera, 3 multzo nagusitan sailkatu dira:

- **STS eleanitza eta hizkuntza-artekoa:**

  Helburua bi esaldi emanda beraien arteko antzekotasuna ebaluatzea da, 0-5 arteko puntuazioa emanaz. Testuinguru elebakarrean zein elebidunean egin dituzte probak.

- **Esaldi paraleloen erauzketa:**

  Helburua bi corpus ezberdinetan (bakoitza hizkuntza batekoa) paraleloak diren esaldiak bilatzea da.

[10] https://opus.nlpl.eu/

- **Antzekotasun bilaketa oso baliabide gutxiko hizkuntzetan**:

  Ingelesa eta beste hizkuntza txiki batzuetarako (Swahilia edo Tatarra, adibidez) esaldi paralelo batzuk hartuta (1000 hizkuntza bakoitzerako, zehazki), esaldi bakoitzarentzako antzekoena bilatzean datza, bi noranzkoetan.

Lortutako emaitzen kalitatea neurtze aldera, beste artearen egoerako eredu batzuek (LASER [Artetxe and Schwenk, 2018], mUSE [Yang et al., 2019] eta LaBSE [Feng et al., 2020]) zein distilatu gabeko ereduek lortutakoarekin alderatu dira. Labur azaltzeko, emaitzak distilatutako ereduak distilatu gabekoak baino eraginkorragoak direla erakusten dute, hobekuntza nabariak lortuz beti ere. Gainontzeko artearen egoerako ereduekin alderatuta, nagusiki emaitza hobeak lortzen dira, esaldi paraleloen erauzketako atazean izan ezik; bertan, LASER eta LaBSEk eskuratzen dituzte emaitza onenak, itzulpen guztiz zehatzak lokalizatzeko prestatuta baitaude. Hala ere, itzulpenak zertxobait aldatzen direnean arazo nabariak dituzte, distilatutako ereduek testuinguru guztietan emaitza lehiakorrak lortzen dituzten bitartean. Ondorioz, emaitza hauek erakusten dute distilazioa oso aukera egokia dela, bai erraztasun eta bai eraginkortasun aldetik, esaldien errepresentazioarekin lotutako ataza gehientsuenak ebazteko.

3. KAPITULUA

Metodologia

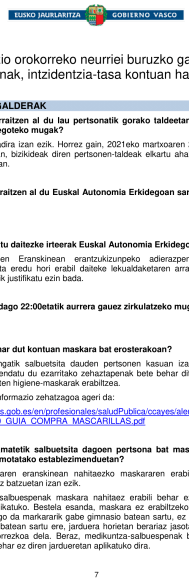
Atal honen helburua proiektua garatzeko eman diren urratsak azaltzea izango da, une bakoitzean hartutako erabaki garrantzitsuenak erakutsiz. Lehenik eta behin, lanaren hasieran definitutako esperimentazio ingurunea aurkeztuko da, erabilitako hizkuntzak, ereduak, datu-multzoak eta ebaluazio metrikak zehaztuz. Behin horiek ongi finkatuta, bukaerako emaitzen baliozkotasuna frogatzeko balioko duen oinarri-lerroa zein den eta nola lortu den adieraziko da; eta, azkenik, distilazio teknika burutzeko eman beharreko pausuak azalduko dira.

3.1 Esperimentazio ingurunea

Edozein esperimentaziorekin hasi aurretik, ideiak ondo finkatzeko helburuz, proiektuan garrantzi handia izan duten hainbat erabaki hartu dira.

3.1.1 Hizkuntzak

Hasteko, proiektuan zein hizkuntzetan lan egingo den erabaki da. Lehenengo asmoa soilik ingelesa-euskara hizkuntza parea jorratzea zen, lanaren jomuga euskarazko sistema eraginkor sortzea izan baita; hala ere, Osakidetzak gaztelaniazko bertsioa ere eskaintzen duela jakinda (ikus 3.1 irudia), aukera hori ere lantzea erabaki zen, baliabide gutxi eta askoko hizkuntzen arteko aldeak aztertzea ahalbidetuz. Ondorioz, esperimentazio berberak burutu dira bai ingelesa-euskararako eta bai ingelesa-gaztelaniarako.



EUSKO JAURLARITZA GOBIERNO VASCO

II. Aplikazio orokorreko neurriei buruzko galderak eta erantzunak, intzidentzia-tasa kontuan hartu gabe

OINARRIZKO GALDERAK

21. Indarrean jarraitzen al du lau pertsonatik gorako taldeetan espazio publiko eta pribatuetan egoteko mugak?

Bai, bizikideak badira izan ezik. Horrez gain, 2021eko martxoaren 26tik aprilaren 9ra bitarteko egunetan, bizikideak diren pertsonen-taldeak elkartu ahal izango dira soilik espazio pribatuetan.

22. Indarrean jarraitzen al du Euskal Autonomia Erkidegoan sartzeko eta irteteko muga?

Bai.

23. Nola justifikatu daitezke irteerak Euskal Autonomia Erkidegotik?

Dokumentu honen Eranskinean erantzukizunpeko adierazpenaren eredu bat eskaintzen da, eta eredu hori erabil daiteke lekualdaketaren arrazoia egiaztatzeko, beste modu batetik justifikatu ezin bada.

24. Indarrean al dago 22:00etatik aurrera gauez zirkulatzeko muga?

Bai.

25. Zer hartu behar dut kontuan maskara bat erosterakoan?

Arrazoi medikoengatik salbuetsita dauden pertsonen kasuan izan ezik, Kontsumo Ministerioak gomendatu du ezarritako zehaztapenak bete behar dituzten eta etiketan adierazi behar duten higiene-maskarak erabiltzea.

Esteka honetan informazio zehatzagoa ageri da:

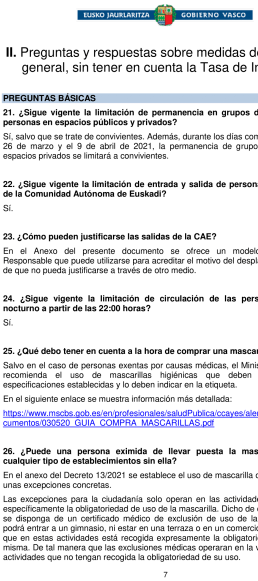
https://www.mscbs.gob.es/en/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/030520_GUIA_COMPRA_MASCARILLAS.pdf

26. Maskara eramatetik salbuetsita dagoen pertsona bat maskara gabe sar al daiteke edozein motatako establezimenduetan?

13/2021 Dekretuaren eranskinean nahitaezko maskararen erabilera ezartzen da, salbuespen zehatz batzuetan izan ezik.

Herritarrentzako salbuespenak maskara nahitaez erabili behar ez den jardueretan baino ez dira aplikatuko. Bestela esanda, maskara ez erabiltzeko mediku-ziurtagiria eduki arren, ezingo da markarik gabe gimnasio batean sartu, ez eta terraza batean egon edo saltoki batean sartu ere, jarduera horietan bezalaxe jasota baitago maskara erabiltzea derrigorrezkoa dela. Beraz, medikuntza-salbuespenak bide publikoan eta nahitaez erabili behar ez diren jardueretan aplikatuko dira.

7



EUSKO JAURLARITZA GOBIERNO VASCO

II. Preguntas y respuestas sobre medidas de aplicación general, sin tener en cuenta la Tasa de Incidencia

PREGUNTAS BÁSICAS

21. ¿Sigue vigente la limitación de permanencia en grupos de más de cuatro personas en espacios públicos y privados?

Sí, salvo que se trate de convivientes. Además, durante los días comprendidos entre el 26 de marzo y el 9 de abril de 2021, la permanencia de grupos de personas en espacios privados se limitará a convivientes.

22. ¿Sigue vigente la limitación de entrada y salida de personas en el territorio de la Comunidad Autónoma de Euskadi?

Sí.

23. ¿Cómo pueden justificarse las salidas de la CAE?

En el Anexo del presente documento se ofrece un modelo de Declaración Responsable que puede utilizarse para acreditar el motivo del desplazamiento en caso de que no pueda justificarse a través de otro medio.

24. ¿Sigue vigente la limitación de circulación de las personas en horario nocturno a partir de las 22:00 horas?

Sí.

25. ¿Qué debo tener en cuenta a la hora de comprar una mascarilla?

Salvo en el caso de personas exentas por causas médicas, el Ministerio de Consumo recomienda el uso de mascarillas higiénicas que deben cumplir con las especificaciones establecidas y lo deben indicar en la etiqueta.

En el siguiente enlace se muestra información más detallada:

https://www.mscbs.gob.es/en/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/030520_GUIA_COMPRA_MASCARILLAS.pdf

26. ¿Puede una persona eximida de llevar puesta la mascarilla entrar en cualquier tipo de establecimientos sin ella?

En el anexo del Decreto 13/2021 se establece el uso de mascarilla obligatorio salvo en unas excepciones concretas.

Las excepciones para la ciudadanía solo operan en las actividades que no recojan específicamente la obligatoriedad de uso de la mascarilla. Dicho de otra forma, aunque se disponga de un certificado médico de exclusión de uso de la mascarilla, no se podrá entrar a un gimnasio, ni estar en una terraza o en un comercio sin mascarilla, ya que en estas actividades está recogida expresamente la obligatoriedad de uso de la misma. De tal manera que las exclusiones médicas operaran en la vía pública y en las actividades que no tengan recogida la obligatoriedad de su uso.

7

(a) Euskarazko MEGak. **(b)** Gaztelaniazko MEGak.

3.1 Irudia: Osakidetzak eskainitako MEGen adibide batzuk, bai euskara eta bai gaztelaniarako; ikus daitekeen bezala, bietan dagoen informazioa berbera da, galdera zein erantzun berdinak edukiz.

3.1.2 Ereduak

Behin zein hizkuntza erabiliko ziren erabakita, sarean atzigarri dauden eredu guztietatik batzuk aukeratu ziren, proiektuaren ezaugarrietara hoberen egokitzen direnak. Gainera, irakasle eredu bakar batekin lan egitea erabaki arren, bi ikasle eredu ezberdin frogatzea zehaztu zen, emaitzak nola aldatzen diren ikuste aldera.

Irakaslea

Irakasle funtzioa betetzeko, 2.2 atalean azaldutako esaldi-errepresentazio eredu elebakar baten beharra dago; eredu hauek, lehenago aipatu bezala, ikaragarrizko datu kopuru handian birdoituta daude eta kalitate oso altuko errepresentazioak sortzeko gai dira.

Zorionez, eredu horietako asko kode irekikoak dira, sarean edonorentzat eskuragarri egonik. Hori gertatzen da zehazki SBERT.net [1] web-orrialdean, non, beste hainbat edukiren artean, 2.2.2 atalean azaldutako hainbat SBERT eredu eskaintzen diren; honela, beren

[1] https://www.sbert.net/

arteko desberdintasun nagusia birfintzeko erabili den datu-multzoa da, eta ondorioz, zer atazetan erabiltzeko prestatuta dauden.

Horretaz gain, eredua aukeratzerako orduan, bi ezaugarri eduki dira kontuan: alde batetik, errendimendu ona edukitzea, eta beste alde batetik, proiektura ondo egokitzen den datu-multzo batean (antzekotasuna jorratzen duena) birdoituta egotea. Azken honen harira, ereduak aukeratzearekin batera garapenerako datu-multzoa ere aukeratu zen (Quora corpusa); ondorioz, irakaslea bertan birdoituta egotea ere komeni dela erabaki zen, ahalik eta kalitate handieneko errepresentazioak sor ditzan.

Hau guztia kontutan hartuta, eskuragarri dauden eredu posible guztietatik **all-mpnet-base-v2** aukeratzea erabaki zen, BERT eta XLEnt ereduetan oinarritzen dena; gehiago zehaztuta, bakoitzaren aurre-entrenamendu teknika bat erabiltzen du, lehenengotik MLM eta bigarrenetik PLM (*Permuted Language Modeling*), berauen abantailaz baliatuz eta mugak minimizatuz. Hasteko, 3.2 irudian erakusten den bezala, batezbesteko errendimendu onena duena da, azkarra eta (memoria aldetik) arina izateaz gain. Ereduaren ezaugarrietara sartu baino lehen, entrenamendu datuei buruzko azalpen txiki bat emango da: eredua bilioi inguru (zehazki 1.170.060.424) esaldi-paretan entrenatuta dago, 32 corpus ezberdin erabiliz; horietako bat, guk nahi bezala, Quora da, eta antzeko formatua duten beste corpus batzuk ere aurki daitezke.

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
gtr-t5-xxl	70.73	55.76	63.25	50	9230 MB
gtr-t5-xl	69.88	55.88	62.88	230	2370 MB
sentence-t5-xxl	70.88	54.40	62.64	50	9230 MB
gtr-t5-large	69.90	54.85	62.38	800	640 MB
all-mpnet-base-v1	69.98	54.69	62.34	2800	420 MB
multi-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
multi-qa-mpnet-base-cos-v1	66.29	57.46	61.88	2800	420 MB
all-roberta-large-v1	70.23	53.05	61.64	800	1360 MB
sentence-t5-xl	69.23	51.19	60.21	230	2370 MB
all-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v1	68.83	50.78	59.80	7500	120 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB

3.2 Irudia: Irakasle gisa jardun dezaketen SBERT ezberdinak, bakoitzaren inguruko hainbat xehetasun emanez; proiektu honetarako, beste zenbait ezaugarriren artean, batezbesteko errendimendua hartu da kontuan eredua aukeratzerakoan.

Iturria: https://www.sbert.net/docs/pretrained_models.html

Aipatutakoaz gain, ezaugarri garrantzitsuenetako bat birdoitze prozesuan erabilitako kontrastezko ikaskuntza helburua da, non esaldi jakin bat emanda sistemak bere bikotea bilatu behar duen ausazko multzo zabal batean; gainera, gure proiektuan ideia berbera erabiltzen dugu, beraz hau ere lagungarria izan daiteke ahalik eta errendimendu onena lortzeko. Horretaz gain, antzekotasuna kalkulatzeko kosinu antzekotasuna erabiltzen da, eta, azkenik, entropia gurutzatuko galeraren bidez alderatzen dira iragarpenak eta egiazko balioak.

Beste hainbat ezaugarri erakustekotan, aipatu behar da 384 hitz baino gehiagoko esaldiak ezin dituela prozesatu, kopuru horretan izanik muga; halaber, esaldi horiek berak 768 dimentsioko espazio bektorial trinko batera mapeatuko dira, informazioaren berreskuratze, *clustering* eta esaldien antzekotasun atazetarako prestatua egongo dena (gehienbat). Azkenik, dimentsio jakin horretako errepresentazioak lortzeko erabiliko den teknika batezbesteko *poolinga* izango da.

Ikaslea

Aipatu bezala, ikasle funtzioa bete behar duen ereduak eleanitza izan behar du, eta gutxienez, euskara, gaztelania eta ingelesarekin lan egiteko gai izan behar du. Ondorioz, 2.1.3 atalean azaldutako **IXAmBERT** eta **XLM-R** ereduak aukeratu dira.

Bi ereduen inguruan jakin beharreko informazio gehiena jada emanda badago ere, aipatu behar da biek 768 dimentsioko errepresentazioak sortzen dituztela, batezbesteko *poolinga* izanik haiek lortzeko teknika. Azkenik, 128 azpi-token prozesatu ahal izango dituzte esaldiko, asko jota.

3.1.3 Datu-multzoak

Hurrengo pausua proiektuan zehar erabili diren corpus ezberdinak azaltzea izango da, bakoitza nondik eta nola lortu den adieraziz. Hiru azpimultzo nagusitan banatzen dira, erabili diren prozesuaren arabera sailkatuz: hasteko, entrenamendurako erabilitako esaldipareak erakusten dira (OPUS webgunetik erauzitakoak), asko eta mota guztietakoak; ondoren, garapenerako eta testerako itzulpen automatiko bidez sortutako datu-multzo txiki bat aurkezten da, Quorako galdera-pareen corpusean oinarritutakoa; eta, azkenik, domeinuz kanpoko testerako erabili den *COVID-19* gaixotasunaren inguruko (eta inkesta bidez jasotako) galderen datu-multzoa azaltzen da.

OPUS (entrenamendua)

Hasteko, distilazio bidezko ikasketan entrenatzeko erabiliko den datu-multzoa zehaztu behar da; horretarako, ingelesa-euskara zein ingelesa-gaztelania hizkuntza konbinazioetarako esaldi-pare paraleloen beharra dago. Hauek eskuratzeko hainbat iturri badaude ere, oso ezaguna eta erabilia den web-orrialde bat existitzen da, hizkuntza askotako esaldi paraleloen corpusak doan eskaintzen dituena: OPUS. Bertan, hizkuntzaren prozesamendurako beste hainbat tresna eskaintzeaz gain, sarean (hizkuntza ezberdinetan) publikoki atzigarri dagoen informazioa hartuz, prozesatuz eta lerrokatuz eraikitako datu-multzoak daude; honela, domeinu ugaritako corpusak eskuragarri uzten dituzte, hainbat formatu ezberdinetan eskainiz (kasu honetan, *Moses* interesatzen zaigu, lerrokatutako testu lauaz osatuta baitago).

Ingelesa-euskara kasuan, hasierako ideia eskuragarri dagoen informazioa guztia erabiltzea izan zen, hauek izanik hizkuntza-pare honetarako dauden corpus guztiak:

- **WikiMatrix:** [119.479 esaldi-pare] Wikipediatik erauzitako informazioa.
- **Wikimedia:** [18.878 esaldi-pare] Wikipediako itzulpenak Wikimedia fundazioa eta haren itzulpen automatikoko sistemaren eskutik.
- **CCMatrix:** [7.778.871 esaldi-pare] Sareko *crawl*-etatik erauzitako esaldiak, datu meatzaritza teknika [2] ezagun baten bidez.
- **EhuHac:** [585.210 esaldi-pare] Hizkuntzen Arte Corpusa [3].
- **GNOME:** [652.298 esaldi-pare] GNOMEko lokalizazio-fitxategiez osatutako corpus paraleloa.
- **XLEnt:** [800.630 esaldi-pare] CCAligned, CCMatrix eta WikiMatrix corpusetatik erauzitako datu-multzoa.
- **KDE4:** [100.160 esaldi-pare] KDE4ko lokalizazio-fitxategiez osatutako corpus paraleloa.
- **Bible-uedin:** [15.893 esaldi-pare] Bibliako pasarteetako esaldiak jasotzen dituen corpus eleanitza.

[2] https://github.com/facebookresearch/LASER/tree/master/tasks/CCMatrix
[3] https://www.ehu.eus/ehg/hac/

- **QED:** [16.913 esaldi-pare] Hezkuntza arloko hainbat bideoetako azpitituluekin sortutako corpusa.
- **TED2020:** [10.400 esaldi-pare] [Reimers and Gurevych, 2020] artikuluan deskribatuta, TED eta TED-X hitzaldietako transkripzioak jasotzen dituen datu-multzo eleanitza.
- **Tatoeba:** [2.061 esaldi-pare] Tatoeba [4] datubasetik erauzitako esaldiak.
- **ELRC_2922:** [316 esaldi-pare] Osasunari eta, bereziki, *COVID-19* gaixotasunari buruzko esaldiak, Wikipediatik erauzitakoak.
- **OpenSubtitles:** [805.780 esaldi-pare] Pelikulen azpitituluetatik ateratako esaldiz osatutako corpus eleanitza.
- **Ubuntu:** [79.474 esaldi-pare] Ubuntuko lokalizazio-fitxategiez osatutako corpus paraleloa.

4 kapituluan azalduko den bezala, esaldi-pare guzti hauek erabiltzea ez da bideragarria, sistemak entrenatzeko behar duen denbora handiegia baita; ondorioz, erdibideko puntu bat bilatu behar exekuzio denbora eta entrenamenduaren kalitatearen artean. Hori horrela, corpus bakoitzeko asko jota 500.000 esaldi-pare hartzea erabaki zen (ikus 3.1 taula); modu honetan, exekuzio-denbora asko jaisteaz gain, domeinu guztietako esaldiak mantendu eta corpusak gehiago parekatzea lortzen da (handienei garrantzia txikituz). Azkenean, garbiketa egin ondoren, entrenamendu prozesua burutzeko 2.683.574 esaldi-pare ezberdin erabili ditugu.

Ingelesa-gaztelania hizkuntza pareari dagokionez, entrenamendua ingelesa-euskararen ahalik eta antzekoena izateko helburuz, bertan erabilitako corpus berberak eskuratzeko ahalegina egin da; hala ere, ez da posible izan, EhuHac datu-multzoa ez baita existitzen hizkuntza konbinazio honetarako. Horretaz gain, lortutako esaldi kopuru totala ere aurrekoa baino txikiagoa zenez, honako corpusak gehitu dira 2.683.574 esaldi-pare horietara ailegatzeko:

- **DGT:** [585.210 esaldi-pare] JRCk (*Joint Research Centre*, Europako Batzordearen ikerkuntza talde bat) argitaratutako itzulpen-memorien [5] bilduma bat.

[4] https://tatoeba.org/eu/
[5] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

- **JRC-Acquis:** [709.496 esaldi-pare] Europear Batasuneko legegintza-testu bilduma bat da, 50. hamarkadatik gaur egun arte idatzitako testu jakin batzuk biltzen dituena.
- **SciELO:** [57.198 esaldi-pare] SciELO [6] (sareko) liburutegi zientifiko elektronikotik hartutako hainbat artikulutatik erauzitako esaldi-pareak.

Datu-multzoa	**Hizkuntza**	
	Euskara	*Gaztelania*
WikiMatrix	119.479	119.479
Wikimedia	18.878	18.878
CCMatrix	500.000	500.000
EhuHac	500.000	-
GNOME	500.000	13.965
XLEnt	500.000	500.000
KDE4	100.160	100.160
Bible-uedin	15.893	15.893
QED	16.913	16.913
TED2020	10.400	10.400
Tatoeba	2.061	2.061
ELRC_2922	316	316
OpenSubtitles	500.000	500.000
Ubuntu	79.474	8.311
DGT	-	500.000
JRC-Acquis	-	500.000
SciELO	-	57.198
Guztira	2.863.574	2.863.574

3.1 Taula: OPUS webguneko corpus ezberdinetatik erauzitako esaldi kopurua, euskara zein gaztelaniarako; datu hauek, praktikan, distilazio teknikaren bidezko ikaskuntzarako erabiliko dira.

Quora (garapena / testa)

Hurrengo datu-multzoaren bidez garapena zein testa burutuko dira, antzekotasun semantikoaren atazarako baliogarriak izanik. Honela, ahalik eta orokortze maila handieneko ereduak lortu nahi dira, ondoren domeinuz kanpoko testean errendimendu ona eskuratzeko helburuz; hori dela eta, datu-multzo honen gainean emaitzarik onenak lortzen dituzten ereduak izango dira hurrengo atalean erabiliko direnak.

[6] https://scielo.org/es/

Proiektuaren betebeharretarako prestatuta dagoen euskarazko corpus bat bilatzea oso zaila denez (hau da, antzekoak diren esaldi-parez osatuta dagoena), bat sortzeko erabakia hartu zen, nahiz eta hutsetik ez hasi; horretarako, HPko atazetan maiz erabilia den Quora [7] datu-multzoa oinarri hartu eta itzulpen automatikoaren bidez euskarazko esaldiak lortu dira. Gainera, prozesu berbera jarraitu da gaztelaniako corpusa sortzeko, bi hizkuntzetan (itzulitako) esaldi berak edukiz.

Quora corpusari dagokionez (entrenamenduko zatia dago soilik eskuragarri), izen bereko web-orriaren edukian oinarrituta dago, ezagutza partekatzeko helburuarekin sortu zena erabiltzaileei galderak egin zein erantzuteko aukera emanaz; gainera, iragazkien erabileraren eta moderatzaileen lanaren ondorioz, bertako edukia antzeko plataforma batzuena (adibidez, *Yahoo! Respuestas*) baino serio eta formalagoa da. Honela, corpusa erabiltzaileek proposatutako galderak erabiliz eraikita dago, 404.290 esaldi-pare ezberdin lortuaz; horretaz gain, pare bakoitza aldagai boolear batekin etiketatuta dago, esaldiak esanahi bera duten edo ez adierazten duena (ikus 3.3 irudia).

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

3.3 Irudia: Jatorrizko Quora corpusaren lau adibide; bertan, esaldi-pare bakoitza *is_duplicate* aldagai boolearraz lagunduta dago, biek esanahi bera duten (1) edo ez (0) adierazten duena.

Iturria: https://paperswithcode.com/dataset/quora-question-pairs

Gure corpusa eraikitzeko, esanahi bera duten adibideak soilik hartu dira kontuan, eta horietatik ausaz aukeratutako 1500 esaldi-pareko azpimultzo bat sortu da; honela, 1000 garapenerako erabili dira, eta gainerako 500 adibideak, testerako.

Behin erabiliko diren esaldi-pareak zehaztuta, Google itzultzailea erabili da (GoogleTrans [8] paketearen bidez) esaldiak bai euskarara eta bai gaztelaniara itzultzeko. Hauek aztertuta, ikus daiteke kalitate onekoak direla (azpiko adibidean erakusten den bezala), ehuneko handi batean emandako itzulpenak ez baitute zuzenpenen beharrik; hala ere, eskuzko (eta

[7] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs
[8] https://pypi.org/project/googletrans/

ondorioz, subjektibo) berrikuspenetan ez sartzearren, itzultzaileak lortutako emaitzei ez zaizkie aldaketarik gehitu.

- What are the best ways to improve your intelligence?
  - Zeintzuk dira zure adimena hobetzeko modurik onenak?
- How can I improve my intelligence?
  - Nola hobe dezaket nire adimena?

Azkenerako, euskara zein gaztelaniazko parekatutako eta esanahi berbereko esaldiak lortu dira, modu honetan izendatu direnak:

- **quora_eu_dev**: Euskaraz, garapenean erabiltzeko 500 esaldi-pare.
- **quora_eu_test**: Euskaraz, testean erabiltzeko 1000 esaldi-pare.
- **quora_es_dev**: Gaztelaniaz, garapenean erabiltzeko 500 esaldi-pare.
- **quora_es_test**: Gaztelaniaz, testean erabiltzeko 1000 esaldi-pare.

COVID-19 (domeinuz kanpoko testa)

Lehenago aipatu bezala, proiektuaren helburua herritarrok *COVID-19* gaixotasunaren inguruko dudak argitzeko sistema bat eraikitzea izan da, Osakidetzako MEGetan oinarrituta. Duda edo galdera hauek lortzeko modurik erosoena, ezinbestean, pertsona ezberdinen artean inkesta bat zabaltzea da; hala ere, zorionez lan hau aurretik burutua zuten IXA [9] taldeko ikerlariek, Eusko Jaurlaritzak proposatutako proiektu baten atal gisa. Bertan, kontzeptu-proba bat aurrera eramateko helburuz, ebaluaziorako balio izango zuen datumultzo txiki bat jasotzeko gai izan ziren, bai euskaraz eta bai gaztelaniaz.

Datu hauek lortzeko, ikerlariek garatutako sistemak erabiltzaileak sartutako galdera bakoitzeko 5 galdera antzekoenak itzultzen ditu (esaldi-errepresentazio eredu baten bidez), proiektu honetan erabiliko den MEG berdineko 100 galdera ezberdinak kontuan hartuz. Ondoren, erabiltzaileak erantzun hauek ebaluatzeko aukera du, horietako bakoitzak bere duda argitzeko balio duen ala ez adieraziz; honela, hauek izan dira hizkuntza bakoitzerako bildutako datuak:

[9] http://ixa.si.ehu.es/

- **Euskara:**
  - Bildutako galdera kopurua: 47
  - Bildutako feedback kopurua: 197 guztira, 49 positibo (%24,87), 148 negatibo (%75,13).
- **Gaztelania:**
  - Bildutako galdera kopurua: 114
  - Bildutako feedback kopurua: 214 guztira, 79 positibo (%36,92), 135 negatibo (%63,08).

Proiektu honetarako, erabiltzaileek emandako feedbacketan oinarritu da ebaluazioa, giza ezagutza kontuan hartzeko helburuz; era honetan, erabiltzaileek feedback positiboa eman dioten erantzunak soilik hartu dira kontuan, gure sistemaren jomuga galdera hori(ek) iragartzea izanik.

Ondorioz, corpusetako datu kopurua nabarmenki murrizten da: Euskaraz 21 galdera eta 49 feedback ezberdin geratu dira, gaztelaniaz 42 eta 79 mantendu diren bitartean, hurrenez hurren. Ikus daitekeen bezala, hasierako feedback guztietatik multzo txiki bat soilik geratu da erabilgarri; beraz, ebaluatzerako orduan komenigarria da kontuan hartzea atazaren zailtasuna, hein handi batean, erantzun zuzen posibleen araberakoa dela, corpus honetan galdera gehienek erantzun zuzen bakar bat edukiz.

Hala ere, galdera bakoitzak duen konplexutasun mailak zehazten du nagusiki iragarpenaren zailtasuna, alde handiak egonik adibide ezberdinen artean; honela, galdera errazenetatik hasita (adibidez, *Kalean erre al daiteke?*) zailenetara igarotzen da, azken hauek hainbat faktoreren menpe egonda: gramatika akatsak, zehaztasun falta, informazio gehigarriaren beharra, MEGean antzeko galderarik ez egotea... Argiago ikusteko helburuz, hona hemen hainbat adibide konplexu, sistemarentzat problematikoak izango liratekeenak:

- Joan naiteke kirola egitera?
- Zelako egoeran dago Bermeo?
- Nire dentista, gorrian dagoen beste herri batean dago. Joan naiteke han dudan hitzordura?

Azkenik, baliteke testerako datu-multzoak txiki samarra ematea, baina eskuragarri dauden baliabideak kontuan edukita oso corpus interesgarri zein baliotsua sortu da; gainera, tamaina hori izatearen ondorioz, erroreen-analisi sakon bat egitea posible da, kasu bakoitza independenteki aztertuz.

3.1.4 Ebaluazio metrikak

Emaitzen ebaluazioari dagokionez, behin esaldi bakoitzarentzako iragarpen onena(k) lortuta, hauen kalitatea neurtu behar da egiazko emaitzekin alderatuz; horregatik, atal honetan proiektuan erabili diren metrika ezberdinak azalduko dira. Honela, ataza honetarako eraginkorrenak diren metriketako batzuk IRE [10] (*Information Retrieval Evaluator*) paketean aurki daitezke, SBERT.net web-orrialdearen barnean. Bertan, bost metrika ezberdin eskaintzen dira, baina ulergarritasuna eta eraginkortasuna kontuan hartuta, iragarpenen kalitatea neurtzeko aukeratu direnak *Zehaztasuna* eta *MRR* izan dira:

- ***Zehaztasuna*:** Sistemak zuzen iragarri dituen adibideen (esaldien) portzentaia kalkulatzen duen teknika; balio hori lortzeko, honako zatiketa burutu behar da:

$$Zehaztasuna = \frac{\text{Zuzen iragarritako esaldi kopurua}}{\text{Iragarritako esaldi kopuru totala}} \tag{3.1}$$

- ***MRR*:** *Mean Reciprocal Rank* sistemak erantzun-zerrenda bat itzultzen duenean eta zerrenda horretan iragarpen zuzenen ordenari zein posizioari garrantzia eman nahi zaionean erabiltzen da; hau da, lehen posizioetan erantzun zuzena lortzea saritu nahi da, azken posizioetan asmatzea (edo, are gehiago, ez asmatzea) zigortu nahi den bitartean. Adibidez, esaldi bakoitzeko hiru iragarpen hartzen baditugu kontuan, esaldiarekiko antzekotasun mailaren arabera ordenatuta daudelarik, metrika honek zigor bat aplikatuko du iragarpenaren posizioa kontuan hartuta (geroz eta atzerago, geroz eta handiagoa); honela, lehen posizioan asmatuz gero balioa 1 izango da, bigarrenean $\frac{1}{2}$ eta hirugarrenean $\frac{1}{3}$ (ez bada asmatzen, 0 izango da). Honakoa kontuan hartuta, *MRR* metrikak esaldi guztien batezbestekoa hartzen du kontuan, formula hau jarraituaz (non $rank_i$ balioak i galderarako lehen erantzun zuzena zer posiziotan iragarri den adierazten duen eta Q aldagaiak, ordea, galdera kopurua):

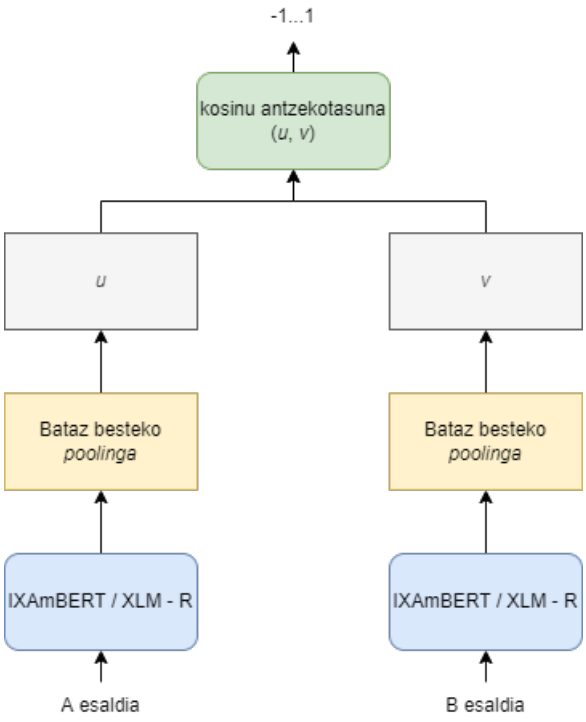
[10] https://www.sbert.net/docs/package_reference/evaluation.html#sentence_transformers.evaluation.InformationRetrievalEvaluator

$$\mathrm{MRR} = \frac{1}{\mathrm{Q}} \sum_{i=1}^{\mathrm{Q}} \frac{1}{\mathrm{rank_i}} \tag{3.2}$$

Horretaz gain, proiektu honetako bi metrikentzako bi aukera ezberdin probatu dira: alde batetik, iragarpen bakarra kontuan hartuta (*zehaztasuna@1* eta *MRR@1*), eta beste alde batetik, 5 iragarpen ezberdin kontsideratuta (*zehaztasuna@5* eta *MRR@5*). Modu honetan, lehen posizioan asmatu ez den kasuetan, iragarpen egokia hurrengo gertuko posizioetan dagoen ala ez jakin ahal izango da, sistemaren baliozkotasuna frogatuz; honen harira, *zehaztasuna* interpretatzea erraza den arren, *MRR* metrikaren bidez lortutako balioak nahasgarriak izan daitezke. Hau hobeto ulertzeko, demagun adibide baten *MRR@5* balioa 0.86 dela; honela, haren alderantzizkoak (kasu honetan, 1.163) batezbestekoa zer bi posizioren artean dagoen adieraziko du, bakoitzetik zer gertutasun duen erakusteaz gain. Hots, aipatutako adibidean, batezbestekoa 1.163 distantzian dagoenez, 1 eta 2 posizioen artean eta lehenengotik oso gertu dagoela ondoriozta dezakegu (hau da, iragarpen gehienak aurreneko saiakeran asmatu dituela). Azkenik, aipatu beharra dago garapen prozesuan *zehaztasuna@1* erabili dela eredu onena aukeratzeko irizpide gisa; gainera, iragarpen bakarra erabiltzen den kasuetan, *zehaztasuna* eta *MRR* metrikek balio bera dute beti, azken honetan ez baitago zigorrik aplikatzeko aukerarik iragarpen bat baino gehiago ez bada kontuan hartzen.

3.2 Oinarri-lerroa

Edozein esperimentu motarekin hasi aurretik beharrezkoa da oinarri-lerro bat finkatzea, ondoren garatutako sistema ezberdinen kalitatea ebaluatu ahal izateko. Kasu honetan, distilazioak emandako hobekuntzak aztertze aldera, distilatu gabeko hitz-eredu eleanitzak (IXAmBERT eta XLM-R) testatuta lortutako emaitzetan ezarri da oinarri-lerroa.

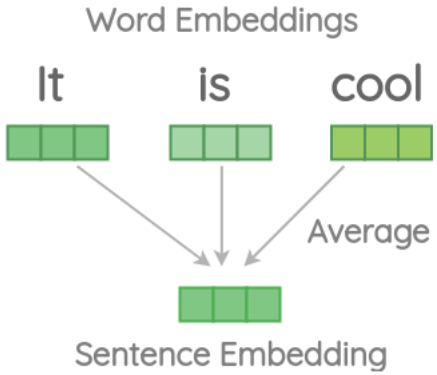


3.4 Irudia: Oinarri-lerroaren arkitektura grafikoki azalduta. Bertan, IXAmBERT edo XLM-R ereduetako batek bi esaldi prozesatzen ditu (*A* eta *B*), hitz bakoitzaren errepresentazio bektorial bat sortuz; ondoren, esaldi-bektore bana (*u* eta *v*) sortzen da hitz-bektore horien informazioan oinarrituta (batezbesteko *poolinga* aplikatuz), azkenik berauen arteko antzekotasun semantikoa ebaluatuta kosinu antzekotasunaren bidez.

Prozesu hau, 3.4 irudian ikus daitekeen bezala, oso intuitibo zein ulergarria da: bi esaldi (*A* eta *B*) hitz-errepresentazio ereduen bidez prozesatu eta hitz-bektoreak lortu ondoren, batezbesteko *poolingaren* bidez esaldi-bektore bana (*u* eta *v*) eraikitzen da hitz-bektoreen informazioan oinarrituta; azkenik, bi esaldien arteko antzekotasuna ebaluatzen da kosinu antzekotasuna erabilita. Prozesua sinplea bada ere, batezbesteko *poolingaren* atala berezi-ki aztertu beharra dago, hura baita hitz-errepresentazio eredu bat erabilita esaldiak beren osotasunean prozesatzea ahalbidetzen duen teknika.

Aipatu bezala, bai IXAmBERT eta bai XLM-R ereduek hitz mailan egiten dute lana; hau da, irteeran hitz (edo azpi-token, kasu batzuetan) bakoitzeko errepresentazio ezberdin bat lortzen dute. Honela, helburua luzera finko bateko esaldi-bektore bakar bat sortzea da, hitz-bektore independente bakoitzak duen informazioa jasotzeko gai dena; horretarako, hain zuzen ere, erabiltzen da *pooling* izeneko teknika. Hainbat modu ezberdin daude hau aurrera eramateko, baina bi aldakuntza ulerterraz eta eraginkor azpimarratu behar dira, oso erabiliak direnak ataza mota honetan: max *poolinga* eta batezbesteko *poolinga*. Ize-

nek iradokitzen duten bezala, lehenengoan bektorearen dimentsio bakoitzeko maximoa hartzen da, eta bigarrenean, batezbestekoa; proiektu honetarako, gainera, azkeneko hau aukeratu da, informazioa hoberen mantentzen duen aukera baita (max *poolingak* ezaugarri garrantzitsuenetan jartzen baitu arreta). Honela, 3.5 irudian ikus daitekeen bezala, tamaina bereko hiru bektore ezberdinetatik bakarra eraikitzen da, posizio berean dauden elementu guztien arteko batezbestekoa kalkulatuaz.



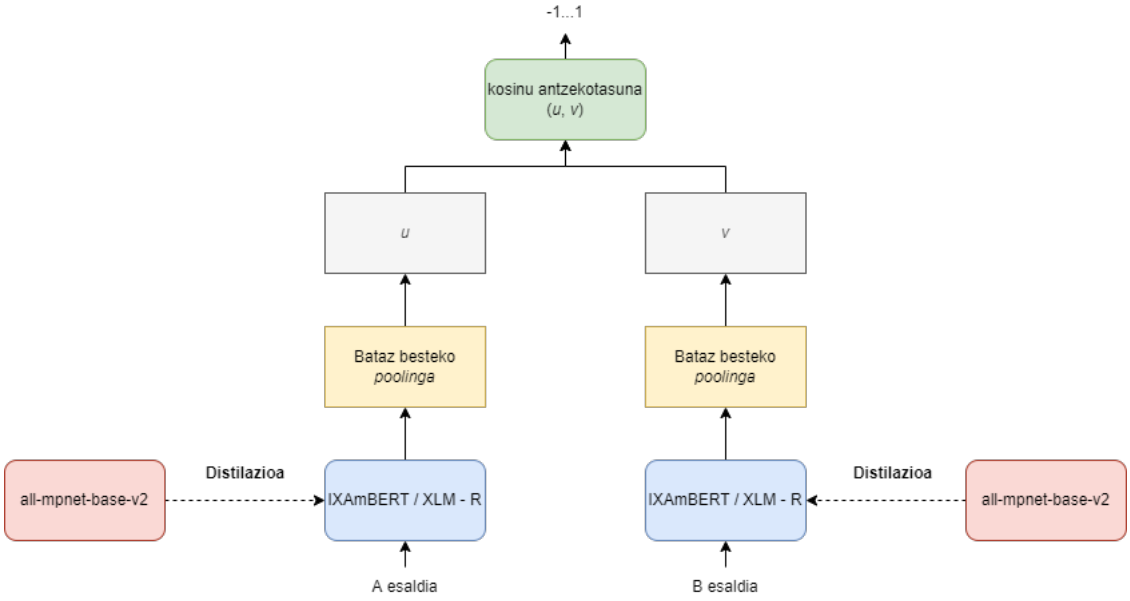
3.5 Irudia: Batezbesteko *poolinga* grafikoki errepresentatuta. Bertan, hitz-bektoreetako dimentsio bakoitzaren batezbestekoa kalkulatuaz esaldi-bektorea lortzen da, hitz-bektoreen luzera berdina duena.

Iturria: https://amitness.com/2020/06/universal-sentence-encoder/

Azkenik, behin bi esaldien errepresentazio bektoreak lortu direnean, beren arteko kosinu antzekotasuna ateratzea besterik ez litzake faltako, bi esaldien esanahiak zenbateraino diren antzekoak -1 (guztiz kontrakoa) eta 1 (berdinak) bitarteko balio numeriko baten bidez ebaluatuz.

3.3 Distilazioa

3.3 atalean aipatu bezala, teknika honen helburua eredu elebakar garatu (irakaslea, all-mpnet-base-v2) baten ezagutza eredu eleanitz ez hain ahaltsu batera (ikaslea, IXAmBERT eta XLM-R) pasatzea da. Metodo honen inguruko azalpen tekniko guztiak jada emanda daudenez, 3.6 irudian grafikoki erakusten da gure proiektuan nola erabiltzen den, oinarrilerroan azaldutako prozesu berbera jarraituz, baina kasu honetan ikasle ereduek distilazio bidezko ezagutza jasoaz irakasletik.



3.6 Irudia: Distilazioaren arkitektura grafikoki azalduta. Bertan, lehenik eta behin, IXAmBERT edo XLM-R ereduetako batek distilazio bidezko ezagutza jasotzen du all-mpnet-base-v2 irakaslearen eskutik; ondoren, bi esaldi prozesatzen ditu, hitz bakoitzaren errepresentazio bektorial bat sortuz. Horren ostean, esaldi-bektore bana (u eta v) sortzen da hitz-bektore horien informazioan oinarrituta (batezbesteko *poolinga* aplikatuz), azkenik berauen arteko antzekotasun semantikoa ebaluatuta kosinu antzekotasunaren bidez.

Distilazio prozesu horren inguruko xehetasun batzuk ematearren, hasieratik lehenetsita zeuden hainbat parametro azalduko dira. Hasteko, galera-funtzioa batezbesteko errore koadratikoa (*BbEK*) izango zela erabaki zen, iragarpenen eta benetako balioen arteko aldea neurtzeko balio duena; honela, helburua ahalik eta balio txikiena lortzea da, formula hau minimizatuz (y egiazko balioa, $\hat{y}$ iragarpena eta n esaldi-pare kopurua izanik):

$$\text{BbEK} = \frac{1}{\text{n}} \sum_{\text{i}=1}^{\text{n}} (y_{\text{i}} - \hat{y}_{\text{i}})^2 \tag{3.3}$$

Horretaz gain, *batcharen* tamaina ere zehaztu zen, 32 izanik aukeratutako balioa. Gainera, optimizatzaileari dagokionez, proiektu honetarako aukeraketa *AdamW* izan da, *Adam* originalaren aldakuntza bat, pisuaren gainbeheraren inplementazio garatuago bat eskaintzen duena; honela, metodo hau erabilita, *overfittinga* murriztea lortzen da kasurik gehienetan. Optimizatzailearen parametroen artean, *learning ratea* da garrantzitsuenetako bat, iterazio bakoitzean pisuak zer neurritan eguneratzen diren zehazten duena; honi, ordea, ez zaio balio jakin bat esleitu, esperimentazioan aukera ezberdinak frogatu nahi izan baiti-

ra (emaitzak nola aldatzen diren aztertzeko). Berdina egin dugu, gainera, beste parametro orokor batekin: *epoch* kopurua, datuak sisteman zehar zenbat aldiz (zikloak) igaroko diren adierazten duena.

4. KAPITULUA

Esperimentuak

Behin proiektua aurrera eramateko jakin beharreko kontzeptuak aztertuta eta jarraitutako metodologiaren inguruko azalpenak emanda, lortutako emaitzei erreparatzeko unea da. Lehen atal honetan, garapen prozesuan eskuratutako emaitzak aztertzen dira, ondoren domeinuz kanpoko testean (*COVID-19aren* ingurukoan) erabiliko diren ereduak aukeratzeko balio dutenak. Horretarako, 3 atalean aipatutako bi aukera ezberdinak burutuko dira; hau da, oinarri-lerroari eta distilazioari dagokienak. Horretaz gain, esperimentu gehigarri bezala, *zeroshot* teknika ere frogatuko da, hizkuntza pare batean entrenatuta dagoen eredua itsuan beste hizkuntzan testatuz; honela, ingelesa-euskara hizkuntza pareko datuekin entrenatutako eredua gaztelaniako esaldien bidez ebaluatuko da (eta berdina alderantziz). Datu-multzoei dagokienez, aurreko kapituluan aipatu bezala, bai garapenerako eta bai testerako erabilitako corpusak ingelesezko Quoraren itzulitako bertsioak izango dira, euskara zein gaztelaniara; gainera, bi hizkuntzetarako (itzulitako) esaldi berberak erabiltzen direnez, euskaraz eta gaztelaniaz lortutako emaitzen arteko konparazioak burutzeko aukera egongo da.

Honela, hurrengo taulen bidez garapen fasean lortutako emaitzak plazaratuko dira, eredu onenak aukeratzeko balio izan dutenak; horretaz gain, helburu nagusia distilazio teknikak dituen onurak aztertzea izango da, oinarri-lerroarekiko dituen hobekuntzei erreparatuz.

Azkenik, hurrengo lerroetan esperimentuetan zehar erabilitako ereduei (eta beren izendapenei) buruzko informazio gehiago ematen da:

- **IXAmBERT / XLM-R:** Distilatu gabekoa.

- **IXAmBERT_dist_en_eu / XLM-R_dist_en_eu:** Distilazio bidezko ezagutza jasota ingelesa-euskara hizkuntza-parean; garapenerako Quora_eu_dev corpusa erabili da.
- **IXAmBERT_dist_en_es / XLM-R_dist_en_es:** Distilazio bidezko ezagutza jasota ingelesa-gaztelera hizkuntza-parean; garapenerako Quora_es_dev corpusa erabili da.

4.1 Oinarri-lerroa

3 kapituluan azaldu bezala, lehenik eta behin distilazio bidezko ezagutzarik gabeko IXAmBERT eta XLM-R ereduak testeatu dira, oinarri-lerroa finkatuz.

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	
Quora_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.73
		Zehaztasuna@5	0.91
		MRR@5	0.8018
	XLM-R	*Zehaztasuna@1*	0.43
		Zehaztasuna@5	0.572
		MRR@5	0.4833
Quora_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.662
		Zehaztasuna@5	0.812
		MRR@5	0.7213
	XLM-R	*Zehaztasuna@1*	0.574
		Zehaztasuna@5	0.724
		MRR@5	0.6329

4.1 Taula: Oinarri-lerroari dagozkion emaitzak, euskarazko zein gaztelaniazko Quora corpusa hizkuntza bereko IXAmBERT eta XLM-R ereduekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute.

4.1 taula aztertuz gero, hainbat ondorio interesgarri atera daitezke. Lehena, eta garrantzitsuena, emaitza onenak IXAmBERTek lortzen dituela da, bai euskaraz eta bai gaztelaniaz XLM-R ereduak baino balio altuagoak lortuz; honakoa, lehenengoaren kasuan hizkuntza hauek oso ondo ordezkatuta daudelako gertatzen da, azkenekoaren kasuan beste hizkuntza askorekin nahastuta dauden bitartean.

Honela, *zehaztasuna@1* metrika kontuan hartuta, IXAmBERTek lortzen du emaitzarik onena, 0.73ko puntuazioa eskuratuz euskarazko testuinguruan; gaztelaniako kasuan, or-

dea, baxuagoa da, 0.662koa izanik. Aipatu bezala, XLM-R ereduak okerrago funtzionatzen du, eta nahiz eta gaztelaniaz aldea hain handia ez den (0.574), euskaraz 0.43ko emaitza soilik lortzen da. Azkenik, *zehaztasuna@5* metrikari erreparatuz, *zehaztasuna@1* balioekin alderatuz 0.14 eta 0.18 bitarteko hobekuntzak lortzen dira, IXAmBERTen kasuan oso emaitza onak lortuz (euskaraz %90etik gora). *MRR@5* balioei begiratuta, kasu gehienetan lehenengo eta bigarren iragarpenen artean ematen da bataz-bestean erantzun zuzena, XLM-R ereduak euskaraz izan ezik, bigarren eta hirugarrenaren artean eskuratzen baitu erantzun zuzena bataz-bestean.

4.2 Distilazioa

Behin oinarri-lerroa zehaztuta eta lortutako emaitzak aztertuta, distilazio teknika erabiltzeak zenbateko onurak ekar ditzakeen ebaluatu da. Horretarako, 3.3 atalean azaldu bezala, irakasle gisa jarduten duen ingelesezko eredu elebakar batek (all-mpnet-base-v2) bere ezagutza pasatzen dio beste eredu eleanitz bati (IXAmBERT edo XLM-R).

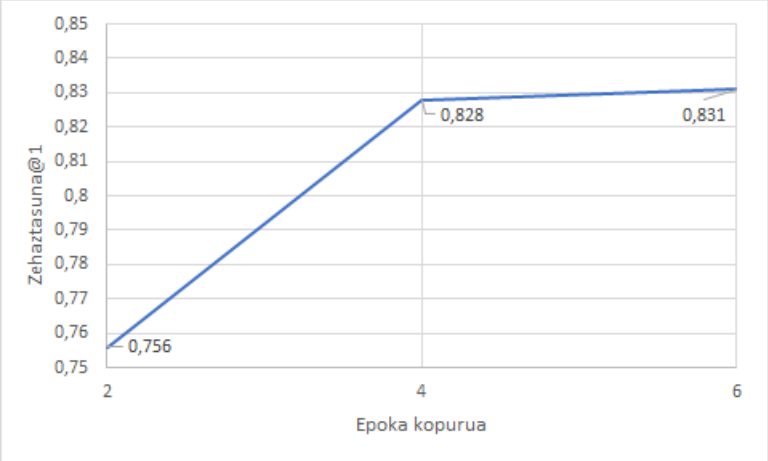
Hori horrela, atal honetako proba definitiboak (tauletan erakusten direnak) egin baino lehen, hainbat esperimentu burutu dira parametro jakin batzuk finkatzeko, entrenamendua ahalik eta eraginkorrena izateko helburuz; hala nola, erabilitako datu kopurua edo parametroen batzuen balioak (epokak edo *learning ratea*) zehaztu dira. Honela, hurrengo lerroetan erabaki hauen inguruko azalpen gehiago ematen dira, azkenik emaitzak erakustearekin batera.

4.2.1 Datuak murriztu

3.1.3 atalean aipatu bezala, hasierako proba batean OPUSen euskaraz eskuragarri dauden datu guztiekin saiatu zen entrenamendua burutzen, baina epoka bakar baten exekuzio denbora 29 ordu eta 23 minutukoa izan zen, batere bideragarria ez den testuinguru bat eskainiz; ondorioz, datu-multzoen itxura aztertu ondoren (tamainari dagokionez alde handia dago haietako batzuen artean) erabaki zen corpus bakoitzetik asko jota 500.000 esaldi-pare hartzea, datu-multzo handienak murriztuz eta denen tamainak gehiago parekatzea lortuz. Honela, exekuzio denbora bideragarri bat erdietsi da, 6 ordu eta 21 minutu behar izanik epoka bat burutzeko. Azkenik, lehenago azaldu bezala, gaztelaniaz ere datu kopuru total berbera erabili da (2.683.574 esaldi-pare), eta, ahal izan den heinean, corpus berak erabiltzeko saiakera ere egin da,

4.2.2 Epoka kopurua

Hurrengo pausua epoka kopurua zehaztea izan da; horretarako, IXAmBERT eredua oinarri hartuta eta euskarazko Quoran testatzeko helburuz, 2, 4 eta 6 epokekin probak egin dira, haien arteko aldeak ikusi eta epoka luzez entrenatzea pena merezi duen erabakitzeko. Denborari dagokionez, epoka bakoitza exekutatzeko 6 eta 6.5 ordu bitartean behar dira, 6 epoka exekutatzeko zehazki 37 ordu eta 38 minutu ematen direlarik; garrantzitsuena, hala ere, lortutako hobekuntzetan dator, 4.1 grafikoan ikus daitekeen bezala: *Zehaztasuna@1* metrikari erreparatuz, 2 epoka erabiliz 0.756ko balioa lortzen da, 4 epokaz exekutatuz 0.828ra hobetzea lortzen den bitartean; hortik aurrera, ordea, 6 epokekin (eta ondorioz, gehiagorekin ere) erdietsitako hobekuntza ia igartezina da, 0.003 puntukoa besterik ez baita. Hau ikusita, nahiz eta 4 epoka exekutatzeko egun bat baino gehiago behar den, lortzen den hobekuntza ikusita pena merezi du denbora hori entrenatzeko erabiltzeak; horretaz gain, epoka kopuru handiagoa zehazteak ez dauka zentzu handirik, hortik aurrera ez baita ia hobekuntzarik eskuratzen (baliabideak alperrik erabiliz).



4.1 Irudia: *Zehaztasuna@1* metrikaren eboluzioa epoka kopuruaren arabera; ikus daitekeenez, 2 eta 4 epoka bitartean hobekuntza handia da, baina puntu horretatik aurrera ez da (ia) hobetzerik lortzen.

Iturria: https://www.ixa.eus/sites/default/files/dokumentuak/8880/SemEval2017_STS_June27.pdf

4.2.3 Learning rate

Behin datu eta epoka kopuruak zehaztuta, definitu beharreko azken parametroa *learning ratea* izan da; horretarako, $7e-5$, $2e-5$ eta $7e-6$ balioak aukeratu dira, jomuga parametroaren balio optimoa lortzea izanik.

Honela, aipatutako testuinguru ezberdinetan burutu dira exekuzioak, emaitzak 4.2 taulan erakutsiz; distilazioari dagozkion balioez gain, urdinez margotuta oinarri-lerroari dagozkion emaitzak gehitu dira (*zehaztasuna@1* metrikakoak), informazio guztia taula berean jasota edukitzearren. Horretaz gain, entrenamenduari buruzko xehetasun gehiago emate aldera, exekuzio bakoitzean behar izan den memoria zein denborei buruzko informazioa adierazten da eranskinetako A.1 taulan.

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	***Learning rate***		
			2e-5	*7e-5*	*7e-6*
Quora_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.73		
Quora_eu_test	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.806	0.828	0.814
		Zehaztasuna@5	0.956	0.972	0.944
		MRR@5	0.8664	0.8897	0.8673
Quora_eu_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.43		
Quora_eu_test	*XLM-R_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.766	0.836	0.616
		Zehaztasuna@5	0.92	0.98	0.826
		MRR@5	0.8302	0.8928	0.6958
Quora_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.662		
Quora_es_test	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.814	0.822	0.8
		Zehaztasuna@5	0.968	0.962	0.956
		MRR@5	0.8763	0.8817	0.8644
Quora_es_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.574		
Quora_es_test	*XLM-R_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.804	0.84	0.7
		Zehaztasuna@5	0.956	0.972	0.86
		MRR@5	0.8654	0.8939	0.7651

4.2 Taula: Distilazio teknikari dagozkion emaitzak, euskarazko zein gaztelaniazko Quora corpusa hizkuntza bereko IXAmBERT eta XLM-R ereduekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute eta azpimarratuta azaltzen diren balioek eredu bakoitzerako *learning rate* onena zein den erakusten dute. Azkenik, urdinez margotutako emaitzak oinarri-lerrokoei dagozkie.

Emaitzak aztertuz gero, lehenik eta behin aipatu behar dena distilazio teknikaren eraginkortasuna da, kasu guztietan oinarri-lerroan lortutako emaitzak hobetzeko gai izan baita; hau aztertzeko, *zehaztasuna@1* metrikari erreparatuko diogu lehenik, esanguratsuena zein erabilgarriena baita proiektu honetan. Honela, *learning rate* ezberdinei erreparatuta,

ikus daiteke $7e-5$ balioarekin lortu direla emaitza onenak, 0.82 eta 0.84 bitarteko balioak lortuaz kasu guztietan; gainera, nahiz eta XLM-R ereduak oinarri-lerroan emaitza baxuenak lortzen dituen, distilazioaren bidez ezagutza jaso ondoren berak du errendimendu onena. Hala ere, balioak IXAmBERTekin lortutakoen oso antzekoak dira, biek ala biek esaldi guztien %16 inguruan ez baitute adibide zuzena iragartzerik izan. Azkenik, beste bi *learning rateak* aztertuta, eredu bakoitzean guztiz portaera ezberdina dutela ikus dezakegu: hasteko, IXAmBERTen kasuan espero genitzakeen emaitza batzuk lortu dira, onenetik gertu dauden balioak lortuz (puntu bat edo biko aldea egonik); guztiz desberdina da, ordea, XLM-R ereduaren kasua, alde nabarmenak baitaude *learning rate* batetik bestera, batez ere $7e-6$ kontuan hartuta. Beraz, eredu hau (testuinguru hauetan behintzat) oso egonkorra ez dela adieraz daiteke, ez baita oso ohikoa parametro honen doitzearen ondorioz horrelako ezberdintasunak lortzea.

Learning rate onenaren ($7e-5$, azpimarratuta daudenak) emaitzak kontuan hartuta, kasurik deigarriena XLM-R ereduarena da, batez ere euskarazko testuinguruan, 0.4tik gorako hobekuntza bat erdietsiz; beraz, distilazio bidezko ezagutzak bi hizkuntzetan eraginkorki jarduteko ahalmena eman dio ereduari. IXAmBERTen kasuan, hobekuntza hain nabaria ez bada ere, *zehastasuna@1* 0.82tik gorako balio batera eramatea lortu da bi hizkuntzetan.

Horretaz gain, *zehaztasuna@5* metrikari begiratuta, ikus dezakegu bi ereduak, kasurik onenetan, iragarpen guztiak asmatzetik gertu geratzen direla; horren adibide, XLM-R ereduak gaztelaniazko testuinguruan lortzen duen balioa, esaldien %2 soilik erratuz. Gainera, *MRR@5* metrikaren informazioa jarraituz, iragarpenen posizioa ona dela ondoriozta daiteke, batezbestekoa ia kasu guztietan lehenengo postuan asmatzetik oso gertu baitago (gehienetan, 1.1 eta 1.2 balioen tartean); hala ere, salbuespena XLM-R ereduaren kasuan dago, $7e-6$ *learning ratearekin*, iragarpenen balioak baxuagoak izanik (1.3 eta 1.4 inguruko balioekin).

Laburbilduz, distilazioaren teknika oso erabilgarria dela esan daiteke, bi eredu zein hizkuntzetarako oinarri-lerroan baino balio altuagoak lortuz; gainera, hobekuntza hauek batez ere XLM-R ereduarentzako izan dira nabariak, oinarri-lerroa baxuagoa izanik, alde handienak lortu dituena izan baita. Horretaz gain, kontuan hartu behar da datu-multzoa itzulpen automatiko bidez sortutakoa dela (inongo eskuzko lanik gabekoa), eta hala ere baliagarria dirudi, erdietsitako emaitzetan ikus daitekeen bezala. Azkenik, *learning ratearen* eragina ere aztertu da, eta IXAmBERTen kasuan diferentzia aipagarririk ez badago ere, XLM-R ereduan espero baino bariantza handiagoko emaitzak lortu dira.

4.3 Zeroshot - Beste hizkuntzan itsuan testatuz

Azkenik, garapen ataleko probekin amaitzeko, esperimentu gehigarri bat erantsi da, aplikazio ugaritan ikus daitekeena duen erabilpen handiagatik; aipatzen ari garen teknika *zeroshot* da, non eredu bat entrenatu ez den hizkuntza jakin batean testatzen den; hau da, kasu honetan, ingelesa-euskara hizkuntza parean entrenatutako ereduak gaztelaniaz testatuko dira, eta ondoren prozesu bera burutuko da baina hizkuntzak trukatuz.

Esperimentu hau aurrera eramateko, distilazioko ataletik emaitza onenak lortu dituzten parametroak aukeratu dira, bai XLM-R eta bai IXAmBERTentzako. Honela, aipatutako prozesua jarraituz, lortutako emaitza guztiak 4.3 taulan erakusten dira; gainera, berriz ere urdinez margotuta ageri dira aurreko ataletan lortutako emaitzak (oinarri-lerrokoak zein distilaziokoak).

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	
Quora_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.73
	IXAmBERT_dist_en_eu	*Zehaztasuna@1 / MRR@1*	0.828
Quora_eu_test	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.796
		Zehaztasuna@5	0.942
		MRR@5	0.8576
Quora_eu_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.43
	XLM-R_dist_en_eu	*Zehaztasuna@1 / MRR@1*	0.836
Quora_eu_test	*XLM-R_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.592
		Zehaztasuna@5	0.764
		MRR@5	0.66
Quora_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.662
	IXAmBERT_dist_en_es	*Zehaztasuna@1 / MRR@1*	0.822
Quora_es_test	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.794
		Zehaztasuna@5	0.946
		MRR@5	0.8579
Quora_es_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.574
	XLM-R_dist_en_es	*Zehaztasuna@1 / MRR@1*	0.84
Quora_es_test	*XLM-R_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.808
		Zehaztasuna@5	0.944
		MRR@5	0.8632

4.3 Taula: *Zeroshot* teknikari dagozkion emaitzak, euskarazko zein gaztelaniazko Quora corpusa aurkako hizkuntzako IXAmBERT eta XLM-R eredu onenekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute; azkenik, urdinez margotutako emaitzak aurreko ataletako oinarri-lerrokoei zein distilaziokoei dagozkie.

Taula aztertuz gero, eta *zehaztasuna@1* metrikari erreparatuz, ikus dezakegu analizatzen diren lau kasuetatik hirutan emaitzak oso antzekoak direla, 0.8 inguruko balioak lortuz; hala ere, ingelesa-gaztelania hizkuntza parean entrenatutako XLM-R eredua euskaraz testatzean, emaitza baxuagoa lortzen da (0.6 baliora ere ailegatzen ez dena), testuinguru horretan lan egiteko balio ez duela erakutsiz. Hots, IXAmBERT baliogarria da *zeroshot* teknika bi noranzkoetan erabiltzeko, XLM-R ingelesa-euskaran entrenatu eta gaztelaniaz testatuta soilik erabil daitekeen bitartean. Horretaz gain, *zeroshot* teknikako emaitzak oinarri-lerrokoekin zein distilaziokoekin alderatuz, bien arteko balio bat lortzen da testuinguru guztietan, distilazioan lortutako baliotik gertu egonik kasu gehienetan; salbuespena, berriz ere, XLM-R_dist_en_es eredua Quora_eu_test corpusean ebaluatzean gertatzen da, oinarri-lerroko emaitzaren balio antzeko bat eskuratuz.

Antzekoa gertatzen da *zehaztasuna@5* metrikarekin ere, aipatutako azken aukera hau kenduta (0.764ko balioa lortzen duena) gainontzeko hiru kasuetan 0.94tik gorako emaitzak lortzen baitira. Azkenik, *MRR@5* emaitzei dagokienez, berriz ere, XLM-R_dist_en_es euskaraz ebaluatzen den testuinguruan ezik (1.51 balioa edukiz, 1 eta 2 posizioen erdian), beste hiru egoeretan batezbestekoa lehenengo posiziotik gertu dago (1.16 inguruan).

5. KAPITULUA

Domeinuz kanpoko emaitzak

Burututako esperimentu guztiak aztertu eta eredu onenak aukeratu ondoren, proiektuaren hasieratik jomuga izan den *COVID-19* gaixotasunaren inguruko testuinguruan hainbat ebaluazio eraman dira aurrera; horretarako, 3.1.3 atalean azaldutako azken datu-multzoa erabili da, IXA taldeko ikerlariek pandemiari buruz egindako lan batetik eratorria izan dena; aipatu bezala, corpusa eraikitzeko erabiltzaileek *COVID-19aren* inguruko galdera batzuk egin, sistema batek Osakidetzaren MEGetatik erantzun posible batzuk itzuli eta, azkenik, erabiltzaileek baliogarriak zirenak markatu zituzten. Ondorioz, gure ereduen helburua galdera horietako bakoitzean erabiltzaileek zuzentzat markatu zituzten erantzunetako bat iragartzea izango da; honela, galdera bakoitzak baliozko erantzun kopuru ezberdin bat eduki dezake, gutxienez bat eta gehienez bost izanik. Galdera kopuruari dagokionez, euskaraz 21 galdera ezberdin daude eskuragarri, guztira 49 feedback jaso dituztenak (2.3 bataz-bestean); gaztelaniaz, ordea, 42 galdera eta 79 feedback aurki daitezke (1.88 bataz-bestean). Ondorioz, garapenean ez bezala, galdera (kasu honetan ez dira itzulpenak) eta kopuru ezberdinak daude bi hizkuntzentzako, beraz, konparazioak egiterako orduan kontu handiz ibili behar da, hizkuntza bateko zailtasuna bestekoa baina handiagoa izango baita seguruenik.

Hau guztia kontuan hartuta, jarraian domeinuz kanpoko testa burutuko da, garapenean egin bezala oinarri-lerroa, eredu distilatuak eta *zeroshot* teknika ebaluatuz; gainera, kapituluaren amaieran erroreen analisia burutuko da, corpusak txikiak izanik gaizki iragarritako adibide guztiak aztertzea bideragarria baita.

5.1 Oinarri-lerroa

Aipatu bezala, lehenik eta behin oinarri-lerroko ereduak *COVID-19aren* inguruko corpusean ebaluatu dira; honela, 5.1 taulan erakusten dira eskuratutako emaitza guztiak.

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	
Covid_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.3986
		Zehaztasuna@5	0.8095
		MRR@5	0.5651
	XLM-R	*Zehaztasuna@1 / MRR@1*	0.1905
		Zehaztasuna@5	0.4762
		MRR@5	0.2857
Covid_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.1905
		Zehaztasuna@5	0.5476
		MRR@5	0.3317
	XLM-R	*Zehaztasuna@1 / MRR@1*	0.1429
		Zehaztasuna@5	0.2857
		MRR@5	0.1984

5.1 Taula: Oinarri-lerroari dagozkion emaitzak, euskarazko zein gaztelaniazko domeinuz kanpoko corpusa (*COVID-19aren* ingurukoa) hizkuntza bereko IXAmBERT eta XLM-R ereduekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute.

Zehaztasuna@1 balioei begiratuta, garapenean gertatu bezala, IXAmBERT da distilatu gabe errendimendu onena ematen duen eredua; hala ere, kasu honetan, euskaraz gaztelaniazko emaitza bikoiztea lortzen da, 0.4 eta 0.2 inguruko balioak erdietsiz, hurrenez hurren. XLM-R ereduaren kasuan, bi hizkuntzetarako balio baxuak lortzen ditu, bereziki gaztelaniazko testuinguruan, 0.15era ailegatu ere egin gabe. Horretaz gain, *zehaztasuna@5* metrikari erreparatuz, aldaketa handienak IXAmBERT ereduan eman dira, *zehaztasuna@1* balioekiko 0.41 eta 0.36 puntuko hobekuntzak lortuaz euskaraz eta gaztelaniaz, hain zuzen ere; XLM-R ereduaren kasuan, ordea, diferentziak ez dira hain esanguratsuak, batez ere gaztelaniaz, 0.14ko aldea egonik *zehaztasuna@1* metrikarekiko. Azkenik, *MRR@5* metrikak ematen duen informazioa aztertuta, orain arteko bide berbera jarraitzen da: IXAmBERT euskaraz ebaluatu den kasuan lortu da emaitzarik onena, batezbestekoa 1 eta 2 posizioen artean egonik; baliorik baxuena, ordea, XLM-R gaztelaniaz ebaluatzen denean ematen da, kasurik gehienetan 5 posizioan iragarriz hautagai zuzena (0.1984 izanik batezbestekoa). Beste bi aukeretan, IXAmBERT euskaraz eta XLM-R gaztelaniaz, 3

eta 4 posizioen artean lortzen dira erantzun zuzenak bataz-bestean (3 eta 3.5 balioekin, hurrenez hurren).

Honela, taula honetatik atera daitekeen ondorio garrantzitsuena, garapeneko oinarri-lerroarekin alderatuz, oraingo atazak (domeinuz kanpokoa izateagatik) duen zailtasuna da, lortutako emaitzetan islatzen den bezala.

5.2 Distilazioa

Orain, behin oinarri-lerroaren emaitzak ikusita, eredu distilatuekin lortutako balioak aztertuko dira, 5.2 taulan erakusten direnak; aipatu bezala, atal honetan erabiliko diren ereduak garapeneko fasean emaitza onenak lortu dituztenak dira. Azkenik, aurreko kapituluan egin bezala, urdinez margotu dira oinarri-lerroko emaitzak.

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	
Covid_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.3986
Covid_eu_test	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.4762
		Zehaztasuna@5	0.9048
		MRR@5	0.6563
Covid_eu_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.1905
Covid_eu_test	*XLM-R_dist_en_eu*	*Zehaztasuna@1*	0.4762
		Zehaztasuna@5	0.8571
		MRR@5	0.6444
Covid_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.1905
Covid_es_test	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.5476
		Zehaztasuna@5	0.7857
		MRR@5	0.6246
Covid_es_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.1429
Covid_es_test	*XLM-R_dist_en_es*	*Zehaztasuna@1*	0.5
		Zehaztasuna@5	0.9048
		MRR@5	0.6476

5.2 Taula: Distilazio teknikari dagozkion emaitzak, euskarazko zein gaztelaniazko domeinuz kanpoko corpusa (*COVID-19aren* ingurukoa) hizkuntza bereko IXAmBERT eta XLM-R ereduekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute; azkenik, urdinez margotutako emaitzak oinarri-lerrokoei dagozkie.

Lehenik eta behin, *zehaztasuna@1* metrikari erreparatuz, argi eta garbi ikus daiteke distilazioa erabiltzeak onura handiak ekarri dituela, batez ere oinarri-lerroan emaitzan ba-

xuenak lortu diren testuinguruetan. Honela, gaztelaniazko kasuan, 0.36 puntu hobetu dira balioak bi ereduentzako; euskaraz, ordea, XLM-R ereduak diferentzia nabarmena erdietsi badu ere (0.29), IXAmBERTek 0.08 puntuko hobekuntza baino ez du izan, oinarri-lerroan lortutako emaitza jada altua baitzen. Horretaz gain, nahiz eta balio altuena (0.55) IXAmBERTek gaztelaniaz lortu duen, gainontzeko hiru kasuetan ere antzeko emaitzak erdietsi dira, gutxi gorabehera esaldi kopuru totalaren erdia asmatuaz; honela, garapen fasearekin alderatuta, agerikoa da ereduen errendimendua jaitsi egin dela, baina guztiz ulerkorra da domeinuz kanpoko testuinguru batean dihardutela aintzat hartuta. Gainera, hurrengo atalean erroreen analisia burutzean, okertzat eman diren iragarpenetako batzuk zuzenak izan daitezkeela erakutsiko da, edo, are gehiago, zenbait galderentzako ez dela posible erantzun zuzenik bilatzea; ondorioz, behin iragarpenak berrikusita, emaitzak puntu batzuk hobetzea espero daiteke.

Horretaz gain, *zehaztasuna@5* metrika kontuan hartuta, 0.4 inguruko aldeak lortzen dira (IXAmBERTek gaztelaniaz apur bat gutxiago, 0.25en biran). Honela, nahiz eta lehenengo aukeran sistemek akats dezente egiten dituzten, hurrengo iragarpenetan galdera zuzena asmatzeko gai dira, atazaren konplexutasuna kontuan hartuta errendimendu on bat erakutsiz; hau da, lehen aukera zuzena ez den kasuetan ere, sistemak ongi bideratuta daude (eta ez dituzte ausazko iragarpenak egiten). *MRR@5* metrikari dagokionez, kasu guztietan antzeko balioak lortzen dira, batezbestekoa 1 eta 2 posizioen erdiko puntuaren inguruan egonik (1.6 inguruan); horregatik, nahiz eta garapenean lorturiko balioak hobeak diren, iragarpenak ongi kokatuta daudela esan daiteke, batez ere domeinuz kanpoko corpus bat dela kontuan hartuta.

Laburbilduz, IXAmBERT izan da errendimendu onena izan duen eredua, nahiz eta probatutako eredu guztiak balio beraren inguruan ibili diren; hori horrela, oinarri-lerroan emaitza baxuenak eman diren egoeretan erdietsi dira hobekuntza handienak, 0.36 puntuko aldea lortuaz kasurik onenean. Azkenik, lehenengo iragarpena okerra izan den egoeretan ere sistemek ondo funtzionatzen dutela erakutsi da, erantzun zuzena hurrengo bost hautagaien artean emateko gai izan baitira ia kasu guztietan.

5.3 Zeroshot - Beste hizkuntzan itsuan testatuz

Garapenean egin bezala, *zeroshot* teknika burutu da domeinuz kanpoko testean ere, praktikan asko erabiltzen den metodo bat baita. Honela, garapenean errendimendu onenak lortu dituzten IXAmBERT eta XLM-R ereduen bidez egin da ebaluazioa, 5.3 taulan erakutsiz

lorturiko emaitzak; horretaz gain, orain arte egin bezala, aurreko ataletako oinarri-lerroko eta distilazioko emaitzak kolore urdinez adierazita daude.

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	
Covid_eu_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.3986
	IXAmBERT_dist_en_eu	*Zehaztasuna@1 / MRR@1*	0.4762
Covid_eu_test	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.4286
		Zehaztasuna@5	0.7619
		MRR@5	0.5556
Covid_eu_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.1905
	XLM-R_dist_en_eu	*Zehaztasuna@1 / MRR@1*	0.4762
Covid_eu_test	*XLM-R_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.2857
		Zehaztasuna@5	0.4286
		MRR@5	0.3452
Covid_es_test	*IXAmBERT*	*Zehaztasuna@1 / MRR@1*	0.1905
	IXAmBERT_dist_en_es	*Zehaztasuna@1 / MRR@1*	0.5476
Covid_es_test	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.3095
		Zehaztasuna@5	0.6429
		MRR@5	0.4313
Covid_es_test	*XLM-R*	*Zehaztasuna@1 / MRR@1*	0.1429
	XLM-R_dist_en_es	*[Zehaztasuna@1 / MRR@1*	0.5
Covid_es_test	*XLM-R_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.381
		Zehaztasuna@5	0.4663
		MRR@5	0.2857

5.3 Taula: *Zeroshot* teknikari dagozkion emaitzak, euskarazko zein gaztelaniazko domeinuz kanpoko corpusa (*COVID-19aren* ingurukoa) aurkako hizkuntzako IXAmBERT eta XLM-R eredu onenekin konbinatuz. Gorriz azaltzen diren emaitzek hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute; azkenik, urdinez margotutako emaitzak aurreko ataletako oinarri-lerrokoei zein distilaziokoei dagozkie.

Hasteko, *zehaztasuna@1* metrikari erreparatuta, garapenean baino bariantza handiagoa dagoela ikus daiteke: euskarazko IXAmBERT gaztelaniaz testatzean, 0.31ko balioa lortzen da, testuinguru berdinean XLM-R ereduarekin balio hori 0.07 puntutan hobetzen den bitartean; gaztelaniazko eredua euskaraz testatzen den egoeratan, IXAmBERTek aukera guztietatik emaitza onena lortzen du (0.4286), eta XLM-Rk, ordea, baxuena (0.2857). Horretaz gain, oinarri-lerroko zein distilazioko balioak kontuan hartuta, haien artean daude *zeroshot* bidez lortutako emaitzak aztertutako aukera guztietan, baina kasu honetan (garapenean ez bezala) orokorrean gertuago daude oinarri-lerrotik.

Horretaz gain, *zehaztasuna@5* metrika aintzat hartuta, hobekuntza handiena euskaraz IXAmBERT_dist_en_es testatzean ematen da, 0.76ko balioa lortzera iritsiz (0.33 puntu-

ko hobekuntza); gainontzeko kasuetan, eskuratutako aldeak txikiagoak dira, 0.08 eta 0.14 bitarteko hobekuntzak lortuaz *zehaztasuna@1* balioekiko. Azkenik, *MRR@5* metrikak ematen duen informazioa aztertuta, ereduen artean diferentzia nabarmenak daudela ikus daiteke: alde batetik, euskaraz entrenatu eta gaztelaniaz testatzen den kasuetan, IXAmBERTek 2 eta 3 posizioen artean du batezbestekoa (2.etik gertuago, balioa 2.32 izanik) XLM-Rk 3 eta 4 posizioen erdiko puntuan (3.5) duen bitartean; alderantzizko kasuetan, IXAmBERT iragarpen zuzena 1 eta 2 posizioen artean emateko gai da kasu gehienetan (1.8), eta XLM-Rk, ordea, 2 eta 3 posizioen artean dauka batezbestekoa (3tik gertuago, 2.9 balioarekin).

5.4 Errore-analisia

Aipatu bezala, domeinuz kanpoko testean jasotako galderetako asko ulertzea eta interpretatzea ez da erraza sistementzako, izan akats gramatikalengatik, informazio faltagatik edo zehaztasun ezarengatik. Honela, hainbat kasutan lortutako iragarpena zuzentzat har daiteke, nahiz eta ebaluazioak errorea dela esan; edo, are gehiago, baliteke egindako galderak antzekorik ez edukitzea MEGetan, iragarpena okertzat hartzea errorea izanik. Ondorioz, errore-analisia burutu da egoera guzti hauek aztertzeko, honakoak izanik aipatu beharreko kasu garrantzitsuenak:

- **Galdera orokorregiak (zehaztasun eza):** Hainbat eta hainbat galdera daude kontzeptu oso orokorrei buruz dihardutenak, eta ondorioz, erantzun zuzen bat baino gehiago onartzen dituztenak; askotan, *Gold Standarreko* (GS) aukera egokia izaten da, baina galdera hauek modu ezberdin askotara erantzun daitezkeenez, sistemak itzulitakoa ere onargarria da:

  `Galdera:` `Joan naiteke senitartekoak bisitatzera?`

  `GSko aukerak:` `[1] Dibortziatua naiz. Dibortzio-epaiaren arabera, astean bitan egon naiteke seme-alabekin; baina ez dira bizi nire Autonomia Erkidego berdinean. Bisita hauek egitera joan naiteke?`

  `[2] Nire alaba Bartzelonatik Gasteizera doa eta altzariak garraiatzeko laguntza behar du. Laguntzera joan naiteke?`

Sistemaren erantzuna: Baimenduta al dago adinekoen egoitzetara joatea?

Ikus daitekeen bezala, GSko aukerak bezain baliogarria da sistemak itzulitakoa, baina okertzat hartzen da ebaluazioan; hau ekiditeko, galderak zehatzagoa behar luke, kasu honetan senitartekoei buruzko informazio gehiago emanaz (adina, bizilekua...).

- **Informazio gehigarriaren beharra:** Galdera mota hauen kasuan, haietan ematen den informazioa interpretatzeko ezagutza gehigarriak behar dira; normalean, datu geografikoekin lotutakoak dira, galdera erantzuteko leku bakoitza non dagoen jakin behar delarik. Honela, iragarpena ez litzake okertzat hartu beharko, sistemaren betebeharrez haratago baitago eskatzen dena:

  Galdera: ¿Urkullu, que vive en Durango, puede subir al Gorbea mañana?

  GSko aukera: ¿Puedo ir al monte del municipio colindante acompañado?

  Sistemaren erantzuna: Si vivo en Durango y mi peluquería habitual está en Bilbao. ¿Puedo desplazarme hasta allí?

  Kasu honetan, galdera erantzuteko Durango zein Gorbea zer diren eta non dauden jakin behar da; GSko hautagaian bezala, nahiz eta Gorbea mendia dela ondorioztatu, haren kokapena (Zeanuri) eta Durango mugakideak diren jakin behar da aukera guztiz baliozkoa izan dadin (errealitatean ez badira ere). Sistemak itzulitako erantzunean, ikusten da nahiko ongi bideratuta dagoela, eta gainera bere erantzuna egokia izan daiteke erabiltzailearen galdera argitzeko. Laburbilduz, galdera mota hauek duten konplexutasuna aintzat hartuta, iragarpena ontzat har daiteke.

- **MEGetan galdera antzekorik ez:** Errore garrantzitsu zein errepikatuenekin amaitzeko, MEGetan antzekorik ez duten galderak aurki daitezke. Batzuetan, nahiz eta guztiz antzekoak ez izan, MEGean aurki daiteke egindako galderaren nahikoa berdintsua denik; hala ere, beste kasu batzuetan, MEGeko 100 galdera posibleetatik batek ere ez dauka egindakoarekin zerikusirik. Azken hauetan, iragarpena zuzentzat eman daiteke, edo bestela, galdera baliogabetu.

  Galdera: Puedo beber alcohol en la calle?

  GSko aukera: ¿Está permitido el consumo en barra?

Sistemaren erantzuna: ¿Se puede comer o beber en el transporte público?

Adibide honetarako, MEGetan bilatuz gero, ez dago antzekoa kontsidera daitekeen galderarik. Adibidez, taberna barruetako legediari buruzko aukerak badaude, alkoholaren kontsumoaren inguruan dihardutenak; hala ere, kanpoaldeko mugak aipatzen diren galderetan, ez dago alkohola edateari buruzko informaziorik. Ondorioz, MEGetan galdera honi erantzuteko hautagairik ez dagoela adieraz daiteke, eta ondorioz, ez dagoela iragarpen zuzenik.

5.4.1 Datu-multzoaren berrikuspen erdi automatikoa

Errore-analisia egin eta egoera aztertu ondoren, datu-multzoa berrikusi da galdera bakoitzarentzat sistemak itzulitako emaitzak zuzentzat har daitezkeen edo ez begiratuz. Horrela, berrikusitako datu-multzo horrekin berriz ere sistema ebaluatu da, 5.4 taulan erakutsiz *zehaztasuna@1* metrikarako lortutako emaitzak (urdinez margotuta daude berrikusi gabeak); izendapenari dagokionez, Covid_eu_test_BER deitu zaio euskarazko datu-multzo berrikusiari eta Covid_es_test_BER, ordea, gaztelaniakoari).

Datu-multzoa	Ereduaren izena	Ebaluazio metrika	
Covid_eu_test	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.4762
Covid_eu_test_BER	*IXAmBERT_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.6667
Covid_eu_test	*XLM-R_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.4762
Covid_eu_test_BER	*XLM-R_dist_en_eu*	*Zehaztasuna@1 / MRR@1*	0.6667
Covid_es_test	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.5476
Covid_es_test_BER	*IXAmBERT_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.7143
Covid_es_test	*XLM-R_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.5
Covid_es_test_BER	*XLM-R_dist_en_es*	*Zehaztasuna@1 / MRR@1*	0.6429

5.4 Taula: Distilazioko emaitzen konparaketa datu-multzoa erdi automatikoki berrikusi gabe (urdinez margotuta) eta berrikusi ondoren (*BER* etiketa gehitu datu-multzoaren izenean); okertzat hartutako iragarpenak aztertu dira, egindako galderarentzat onargarria den edo ez ebaluatuz. Gorriz azaltzen diren emaitzak hizkuntza bakoitzerako lorturiko *zehaztasuna@1* onena adierazten dute.

Taulari erreparatuta, euskarazko kasuan bi ereduek hobekuntza bera lortzen dute, 0.19 puntuko (4 esaldiko) aldea erdietsiz berrikusi gabeko ereduekiko; gaztelaniazko kasuan, ordea, IXAmBERTek lortzen du hobekuntza handiena (0.17ko diferentzia), XLM-Rk 0.14 puntuko aldea eskuratzen duen bitartean.

6. KAPITULUA

Ondorioak eta etorkizuneko hobekuntzak

Txostena amaitzeko, azken kapitulu honetan proiektutik ateratako ondorio nagusiak azalduko dira, aukera edukiz gero etorkizunean burutu daitezkeen hainbat hobekuntza aztertzearekin batera.

Honela, ondorioei dagokienez, bi zati nagusitan banatuko da azalpena, lehenbizi proiektuaren inguruko ondorio tekniko batzuk emanez eta ondoren kontzeptu orokorragoak (pertsonalagoak) aztertuz; etorkizuneko hobekuntzen inguruan, ordea, denbora zein baliabide mugengatik aurrera eraman ezin izan diren hainbat puntu erakutsiko dira, proiektua gehiago garatzen lagunduko luketenak.

6.1 Proiektuaren inguruko ondorioak

Proiektuaren hasieran, *COVID-19*ari buruzko dudak argitzeko gai den sistema bat garatzea proposatu zitzaigun, MEG multzo batean galdera horren semantikoki antzekoena bilatuz; horretarako, euskaraz dauden baliabide-mugak kontuan hartuta, esaldi-errepresentazio eredu lehiakor bat sortu nahi izan da, distilazio teknika erabiliz helburu hori lortzeko. Horretaz gain, baliabide gehiagoko hizkuntza baten errendimendua ere ebaluatu nahi izan da, eta horregatik gaztelaniaz ere egin dira proba guztiak. Laburbilduz, helburu nagusia antzekotasun semantikoko ataza honetan distilazioa erabiltzeak duen eragina aztertzea izan da, kontzeptu berri guzti hauek ondo ulertuz eta barneratuz. Honela, behin proiektua amaitutzat emanda, hasieran orokorrean planteatuko helburu guztiak bete direla esan daiteke, lortutako ezagutzetan zein emaitzetan ikus daitekeen bezala.

Ikasitako kontzeptu berri guztiez gain, hainbat ekarpen egitea ere lortu da; hasteko, IXAmBERT eta XLM-R ereduen aldakuntza interesgarri batzuk sortu dira, distilazio bidez ezagutza jaso dutenak bi esaldiren arteko antzekotasuna ebaluatzeko (euskara zein gaztelaniarako) eta *COVID-19aren* inguruko ataza ebazteko eraginkorrak izan direnak. Horretaz gain, itzulpen automatiko bidez sortutako datu-multzoa (Quoran oinarritutakoa) benetan erabilgarria da, beste hainbat atazetan ere aprobetxatu daitekeelarik; gainera, nahiz eta proiektuan azpimultzo txiki bat soilik erabili den, esfortzu handirik gabe itzul daitezke jatorrizko corpuseko esaldi guztiak (sortutako kodea berrerabiliz).

Emaitzei dagokienez, helburu nagusia bete dela esan daiteke, distilazio teknikak hobekuntza nabariak ekarri baititu oinarri-lerroarekin alderatuta; are gehiago, batez ere Quorako testa kontuan hartuta, benetan emaitza lehiakorrak lortu dira. Hain zuzen ere, azken hau ingelesezko bertsiotik itzulpen automatiko bidez (eta eskuzko gainbegiraketarik gabe) sortutako corpusa da, eta bertan lortutako errendimendu onak beste antzeko datu-multzo bat itzulita ere emaitza onak lor daitezkeela pentsarazten du; hau bereziki interesgarria da euskararentzat, baliabide gutxiko hizkuntza baita eta askotan ez baita erraza ataza bakoitzerako beharrezko datu-multzoak aurkitzea. Emaitzekin bukatzeko, eta hizkuntzen arteko konparaketa bat eginaz, euskararako gaztelaniarako bezain emaitza onak lortu direla ikus daiteke, bi testuinguruetarako erabilgarriak diren ereduak eskuratuz.

Behin ondorio garrantzitsuenak aterata, proiektuan emandako pauso ezberdinen inguruko ondorioak azalduko dira, kronologikoki ordenatuta; honela, lehen urratsa proiektua aurrera eramateko beharrezko ezagutza teorikoa eskuratzea izan da, eskuragarri dauden baliabide ezberdinak (ereduak, teknikak...) ulertuz eta egokienak aukeratuz. Modu honetan, kontzeptu berri asko ikasteko aukera eduki da, batez ere hizkuntza eredu, antzekotasun semantiko zein distilazioaren inguruan.

Ondoren, lanaren nondik norakoak definitzeko balio izan duten hainbat erabaki hartu dira, gehienbat datu-multzo, eredu eta teknikekin lotutakoak. Honela, hainbat aukera ezberdin baloratu ondoren, proiekturako erabilgarrienak direnak hautatu dira, ahalik eta errendimendu onena lortzeko helburuz. Hala ere, praktikan jartzerakoan aldaketa txiki batzuk egin behar izan dira, gehienak baliabideen mugek hala behartuta (adibidez, datu-multzoak murriztu behar izatea).

Horretaz gain, entrenamendua definitzen duten parametroen balioak ere zehaztu behar izan dira, hainbat proba eginez haien eragina ebaluatzeko. Horretarako, hasieran testuinguru sinpleago batean burutu dira esperimentuak, parametroen balioak definitzeko balio izan dutenak; honela, azken probak egiterako orduan jada zehaztapen guztiak eginak zeu-

den, hauek exekutatzeko denborak oso handiak baitira (eta alferrikako saiakerak egitea ekidin behar da).

Hala ere, aipatu bezala, testatze definitiboak ez dira hasieran planteatutakoaren zehazki berdinak izan, datu kopuruak murriztu eta nahi baino epoka gutxiagoz egin baitira entrenamenduak, bestela ez baitira bideragarriak. Edonola ere, egindako probak ez lirateke posible izango IXA taldeak eskainitako baliabiderik gabe, esperimentu hauek ordenagailu pertsonal batean burutzea ezinezkoa baita. Azkenean, beharrezko parametroak egokitu ondoren, eredu definitiboak lortu dira, proposatutako ataza eraginkorki ebazteko gai izan direnak.

Laburbilduz, eskuratutako emaitza zein ekarpenetan oinarrituta, proiektuaren hasieran ezarritako helburuak bete direla esan daiteke, batez ere distilatutako ereduek lortutako errendimendua ikusita.

6.2 Ondorio orokorrak (pertsonalak)

Pertsonalki, proiektua oso aberasgarria izan da alde guztietan: aipatu bezala, gaiaren inguruko ezagutza teorikoa zein praktikoa bereganatu da, eta azken honetan espero ez ziren egoerei aurre egin behar izan zaie; ondorioz, arazo edo problema ezberdinak antzemateko gaitasuna ere garatu da, hauek konpontzeko ahalmena eduki delarik.

Proiektua hasi aurretik, hitz- zein esaldi-errepresentazio moten eta antzekotasun semantikoaren inguruko ezagutza ez zen handiegia, distilazioaren kontzeptua guztiz berria izan den bitartean. Horregatik, lehen urratsa hauen inguruko informazioa eskuratzea izan zen, analisi sakon bat eginez artikuluak zein beste hainbat baliabide baliatuz. Honela, orduko eta gaur egungo egoerak alderatuz, esan daiteke ikasitako kontzeptu kopurua oso handia izan dela.

Horretaz gain, lehenago erabili gabeko hainbat baliabide ezagutzeko aukera egon da, horien artean garrantzitsuena *Sentence Transformers* [1] *Pythoneko frameworka* da, hitz-, testu- eta irudi-errepresentazioak sortzeko garatua izan dena. Baliabide honetan eskaintza zabal bat aurki daiteke proiektukoa bezalako atazak ebazteko, eredu, metrika zein beste hainbat ezaugarri zehazteko aukera ugari emanez.

Era berean, masterrean zehar jasotako irakasgai ezberdinetan bereganatutako ezagutzak asko lagundu du proiektua aurrera eramaterako orduan, batez ere *Deep Learning* eta *NLP*

[1] https://www.sbert.net/

applications (II) bezalakoak kontuan hartuta; bertan, eredu zein teknika ezberdinei buruzko informazioa jaso zen, benetan baliagarria izanik lanean erabilitako kontzeptuen inguruan oinarri bat edukitzeko.

Txostenari dagokionez, zailena ideiak ongi argitu zein egituratzea izan da, ondoren kontzeptu guztiak zuzen azaltzeko helburuz; berebiziko garrantzia du informazioa garbi eta sinpleki helarazteak, txostena ahalik eta irakurterrazena izan dadin. Honela, ideiak behar bezala ordenatu behar dira idazten hasi aurretik, zeregin honetan ahalegin gehigarri bat eginez denbora-galtze asko ekidin daitezke eta.

Arlo pertsonalean sartuta, ekarpen garrantzitsuena sortzen joan diren ezusteko arazoei irtenbidea emateko gaitasuna eduki izana da, denbora gutxian ahalik eta erantzun onena garatzeko helburuz. Gainera, irtenbidea erraz aurki ezin daitekeen kasuak dira onuragarrienak, haietan bidez eskuratzen baitira ezagutza handienak.

Azkenik, ondorioen atala bukatzeko, eta proiektuaren analisi global bat eginez, alderdi guztietatik oso emankorra zein interesgarria izan dela adieraz daiteke; ezagutza berriak lortzeaz gain, garapen pertsonalerako ere balio izan du, honelako lan karga duen proiektu batek dituen ezaugarriengatik. Horretaz gain, landutako kontzeptu guztiekiko (distilazioa, antzekotasun semantikoa...) interesa pizteko balio izan du, aurrerantzean haien inguruan lan ezberdinak burutzeko aukera sortuz.

6.3 Etorkizuneko hobekuntzak

Txosteneko azken atal honetan denbora zein baliabide mugengatik burutu izan ezin diren atazak azalduko dira; hala ere, aipatu behar da proiektuaren hasieran ezarritako helburu gehienak betetzea lortu dela, eta ondoren aipatuko diren hobekuntzak proiektua are gehiago garatzeko balioko luketela. Hori horrela, hauek dira, aukera edukiz gero, eman daitezkeen hurrengo urratsak:

- **Eredu gehiago testatzea:** Proiektu honetarako erabilgarriak diren ikasle eredu asko ez badaude ere, irakasle ezberdinak erabiltzea aukera polita izango litzake, honela beren arteko ezberdintasunak aztertuz; gainera, lanerako aukeratutako irakaslea (all-mpnet-base-v2) oso datu kopuru handian entrenatuta dagoenez, interesgarria izan daiteke atazarako baliagarriak diren corpusetan bakarrik oinarrituta dagoen eredu bat probatzea (nahiz eta txikiagoa izan). Hala ere, erabilitako eredua da

SBERT.net webguneko ebaluazioaren arabera emaitza orokor onenak lortu dituena, beraz ez da erraza izango haren errendimendua hobetzea.

- **Itzultzaile ezberdinak probatzea:** Proiektuko ataza garrantzitsuenetako bat garapenean erabilitako datu-multzoaren sorrera izan da. Kasu honetan, ingelesezko Quora corpusa euskara zein gaztelaniara itzuli da *Google* enpresaren itzultzailea erabilita; hala ere, interesgarria litzake beste aukera batzuekin saiatzea, adibidez, aski ezagunak diren *DeepL* [2] edo *Elia* [3] erabilita. Honela, emaitzen arteko ezberdintasunak ebaluatuz itzultzaile bakoitzaren errendimendua aztertzea posible izango litzateke.

- **Domeinuz kanpoko test handiagoa:** Egiari zor, garbiketaren ondoren *COVID-19aren* inguruko corpusa txiki samarra geratu da (21 galdera euskaraz eta 42 gaztelaniaz); ondorioz, testatzea osoagoa izan dadin, datu-multzo hau handitzea komeni da, erabiltzaile gehiagori inkesta zabalduz galdera zein feedback gehiago lortzeko helburuz. Hala ere, beste aukera interesgarri bat proiektuan garatutako ereduetan oinarritzen den sistema baten bidez, erabiltzaileek zuzenean ebaluazioa burutzea litzateke, ereduen errendimendua epaituz kanpoko baliabideen beharrik gabe.

- **Erabiltzailearentzako interfaze bat sortu:** Behin sistema garatuta eta ondo dabilela ikusita, interesgarria litzateke erabiltzailearentzako interfaze eroso zein deigarri bat sortzea, galderak erraztasunez sartu eta erantzun posibleak pantailaratzen dituena. Horretaz gain, aplikazio honen inguruko beste hainbat hobekuntza ere pentsa daitezke; adibidez, esaldi bakoitzaren alboan bere puntuazioa (kosinu antzekotasun balioa) adieraztea, erabiltzaileak erantzunaren sinesgarritasun maila jakin dezan.

[2] https://www.deepl.com/es/translator
[3] https://elia.eus/itzultzailea

Eranskinak

A. ERANSKINA

Entrenamenduko exekuzio denborak

Datu-multzoa	**Ereduaren izena**	**Ebaluazio metrika**	***Learning rate***		
			2e-5	*7e-5*	*7e-6*
Quora_eu_test	*IXAmBERT_dist_en_eu*	*Memoria (GB)*	8.481	8.623	8.576
		Denbora	25 h 30 m	25 h 28 m	25 h 36 m
	XLM-R_dist_en_eu	*Memoria (GB)*	9.857	9.825	9.919
		Denbora	28 h 30 m	28 h 22 m	28 h 16 m
Quora_es_test	*IXAmBERT_dist_en_es*	*Memoria (GB)*	8.765	9.011	8.579
		Denbora	25 h 8 m	24 h 42 m	24 h 47 m
	XLM-R_dist_en_es	*Memoria (GB)*	9.925	9.967	9.883
		Denbora	27 h 17 m	26 h 52 m	27 h 14 m

A.1 Taula: Distilazio teknika burutzeko entrenamenduan erabilitako memoria zein exekuzio-denborak. Emaitzak aztertuz, IXAmBERT (memoriaz) arinagoa eta (denboraz) azkarragoa dela ikus daiteke; adibidez, exekuzio denborari erreparatuta, 25 ordu inguru behar ditu entrenamendua burutzeko, XLM-R ereduak 28 ordu inguruko lana hartzen duen bitartean.

Bibliografia

[Adhikari et al., 2019] Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Docbert: BERT for document classification. *CoRR*, abs/1904.08398.

[Agerri et al., 2020] Agerri, R., Vicente, I. S., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. *CoRR*, abs/2004.00033.

[Artetxe and Schwenk, 2018] Artetxe, M. and Schwenk, H. (2018). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

[Beltagy et al., 2019] Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

[Cer et al., 2018] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, abs/1803.11175.

[Cer et al., 2017] Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

[Conneau et al., 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Feng et al., 2020] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

[Gou et al., 2020] Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2020). Knowledge distillation: A survey. *CoRR*, abs/2006.05525.

[He et al., 2020] He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

[Iyyer et al., 2015] Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. *Association for Computational Linguistics*, P15-1162:1681–1691.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

[Lample and Conneau, 2019] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

[Lan et al., 2019] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

[Lee and Hsiang, 2019] Lee, J. and Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained BERT model. *CoRR*, abs/1906.02124.

[Lee et al., 2019] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR*, abs/1301.3781.

[Otegi et al., 2020] Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for Basque. *European Language Resources Association*, 2020.lrec-1.55:436–442.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. *Association for Computational Linguistics*, D14-1162:1532–1543.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

[Reimers and Gurevych, 2020] Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR*, abs/2004.09813.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *European Language Resources Association (ELRA)*, L12-1246:2214–2218.

[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. *CoRR*, abs/1706.03762.

[Yang et al., 2019] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Ábrego, G. H., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2019). Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307.

[Zhang et al., 2010] Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1:43–52.