

Gradu Amaierako Lana

Informatika Ingeniaritzako Gradua

Konputazioa

Euskarazko laburpen automatikoak egiteko sistema baten garapena eta ebaluazioa

Ekhi Azurmendi Arrue

Zuzendariak

Oier Lopez de Lacalle Lecuona

Mikel Iruskieta Quintian

2023.eko ekainaren 23

Esker onak

Lehenik eta behin eskerrak eman nahi dizkiet etxekoei eta lagunei, baina bereziki Marta, Uxer eta Asierri. Nire gogoetak, arazoak eta gorabeherak entzun eta jasan baitituzte azken hilabete hauetan.

Eskerrak eman nahi dizkiet lan honetan lagundu didan taldetxoari, ostegunero kafe ordua bileretan igarotzeagatik eta proiektua aurrera eramaten lagundu izanagatik. Eskerrak beraz, Oier Lopez de Lacalle, Mikel Iruskieta, Unai Atutxa, Ekain Arrieta, Itziar Aldabe eta Iñigo Lopezi. Bereziki eskerrak eman nahi dizkiot Unairi, ereduek sortutako testuak zorrotasun eta objektibotasunez ebaluatzeagatik.

Azkenik, eskerrak eman nahi dizkiot IXA ikerketa taldeari lan honetan emandako babesagatik eta makinak erabiltzen uzteagatik. Baita ere, taldeko kideek ideiak biltzen eta arazoak izan ditudanean konponbidea topatzen lagundu izanagatik.

Laburpena

Gradu Amaierako Lan honen helburua da euskarazko laburpen automatikoko sistema baten garapena egitea. Abstrakzio-laburpenen kasua aztertu da, laburpen hauek informazio garrantzitsuena mantentzen duten berridazketa laburrak dira, hau da, laburpeneko esaldiak ez dira testutik hartuak. Horretarako, eleanitzak diren kodetzaile-deskodemak motako HEak diren MT0 transformerrak. Ondoren, entrenatutako ereduak bi eratara ebaluatu dira: lehena, metrika automatikoa, ROUGE, erabilia egin da. Bigarrena, Unai Atutxak bere doktoretza tesian oinarritutako eskuzko ebaluazioa izan da.

Ereduak entrenatzeko, laburpenak egiteko ingeleseko 6 datu multzo itzuli dira NLLB itzultzaile automatikoa erabilia. Guztira 728 mila testu eta horien laburpeneko testu elebiduna sortu da. BLEU metrika erabilia itzulpenaren kalitatea neurtu da eta kalitate egokikoak direla esan daiteke.

Behin corpusa sortu ondoren, 6 laburpen-eredu entrenatu dira. Eredu bakoitza ikerketa galdera bati erantzuteko entrenatu da, eredurik onena entrenatzeko modua aurkitzeko. ROUGE metrika objektiboa erabiliz, ereduak haien artean alderatu eta entrenamenduko erabaki egokienak zein diren aukeratu dira. Entrenatu den laburpen-eredurik onena, MT0-large, 1200 milioi parametro dituen, oinarritzat hartuta sortu da. PEGASUSEkin ingelesezko abstrakzio-laburpenen artearen egoera den ereduarekin alderatu da ROUGE-TOK deitu zaio metrika berri batekin. Metrika berri hau tokenizatzailearen erabilera metrikaren kalkuluan sartuta inplementatu da. ROUGE-TOK erabiltzearen arrazoia ingelesezko eta euskarazko balioak egokiago alderatzea izan da. Entrenatu den eredurik onenak PEGASUSEk baina emaitza hobekien eman lortu ditu.

Azkenik, eskuzko ebaluazioa egin da eta laburpen-eredu onenak datu multzoak egoki laburtzen dituela ikusi da. Ikusi gabeko testu motetan berriz, *prompt*-aren eraginez ez dira kalitatezko laburpenak sortu. Hala ere, euskararen gramatika egokia dela ikusi da ereduak sortutako laburpenetan, nahiz eta sarrerako testuek kalitate txarra izan.

Gaien aurkibidea

Gaien aurkibidea	v
Irudien aurkibidea	viii
Taulen aurkibidea	x
Algoritmoen aurkibidea	xi
1 Sarrera	1
1.1 Motibazioa	1
1.2 Proiektuaren helburu orokorra	2
1.3 Dokumentuaren antolaketa	2
2 Proiektuaren Helburu Dokumentua	3
2.1 Lan-paketeak eta atazen identifikazioa	3
2.1.1 Kudeaketa adarra	3
2.1.2 Corpora osatzeko adarra	4
2.1.3 Entrenamendu adarra	5
2.1.4 Ikerketa adarra	5
2.2 Atazen denborak eta epeak	6
2.3 Arriskuen analisisa	6
3 Aurrekariak	11
3.1 Hizkuntza ereduak	11
3.1.1 Transformers	14
3.1.2 Kodetzaile-Deskodetzaileak	17
3.1.3 MT5: Hizkuntza eredu eleanitza	18
3.2 Laburpen automatikoko sistemak	19
3.2.1 PEGASUS: Artearen egoerako sistema	20
3.2.2 MT0: Ataza anitzeko eredu sortzailea	22
3.3 Automatikoki sortutako testuen ebaluazioa	23
3.3.1 N-grametan oinarritutako ebaluazio metrika automatikoak	23
3.3.2 Eskuzko ebaluazioa: Unai Atutxaren tesia	26
3.4 NLLB Itzulpen automatikoko sistema irekia	32
4 Ingurune Esperimentala	33
4.1 Datu Multzoak	33
4.2 Datu multzoen euskararako itzulpena	35

4.2.1	Algoritmoa eta oinarrizko ideia	35
4.2.2	Itzulpenen analisia	37
4.2.3	Corpusaren deskribapena	40
4.3	Ereduak eta entrenamenduak	42
4.3.1	Entrenamenduen azalpen orokorra	42
4.3.2	Entrenamenduko erabakiak	43
4.3.3	Entrenatutako modelook	45
5	Emaitzak	49
5.1	Ereduen galera funtzioaren azterketa	49
5.1.1	Promptaren eragina aztertzeko entrenamenduak	49
5.1.2	Optimizatzailearen eragina aztertzen	49
5.1.3	Entrenamenduak luzatzearen eragina	52
5.1.4	HE handiagoa erabiltzearen eragina	53
5.2	Metrika automatikoen bidezko ebaluazioa	53
5.2.1	Promptaren eragina aztertzeko entrenamenduak	54
5.2.2	Optimizatzailearen eragina aztertzen	55
5.2.3	Entrenamenduak luzatzearen eragina	56
5.2.4	HE handiagoa erabiltzearen eragina	56
5.2.5	Modeloen alderaketak PEGASUSekin	57
5.3	Eskuzko ebaluazioa	59
5.3.1	Wikihow-eko testu onenak	60
5.3.2	Wikihow-eko testu okerrenak	61
5.3.3	Elikagaien laburpenen ebaluaketa	62
6	Ondorioak eta etorkizuneko lanak	65
6.1	Itzulitako testuak eta corpora	65
6.2	Entrenatutako ereduak	65
6.3	Prompt-aren garrantzia	67
6.4	Mugak eta etorkizuneko lanak	68
6.4.1	Corpusaren garbiketak eta eskuzko kalitate ebaluazioa	68
6.4.2	Datu multzoen eragina entrenamenduetan	69
6.4.3	Datu multzo gehiago itzultzea	69
6.4.4	Prompt ezberdinak sortzea	69
	Eranskinak	71
A	Plangintza eguneratua	71
B	Sortutako laburpenen adibideak	73
B.1	Datu multzoak	73
B.1.1	BillSum	73
B.1.2	Wikihow	74
B.1.3	Reddit	75
B.1.4	Arxiv	75
B.1.5	Pubmed	76
B.1.6	AESLC	78
B.1.7	ROUGE balioak	78

B.2	Elikagaien testuak	79
B.2.1	Eli0	79
B.2.2	Eli1	79
B.2.3	Eli2	80
B.2.4	Eli3	81
Bibliografia		83

Irudien aurkibidea

2.1	Proiektuko LDE diagrama.	4
3.1	RNN sare neuronalak HE moduan erabilia.	12
3.2	LSTM baten diseinu irudia.	13
3.3	GRU baten diseinu irudia.	13
3.4	Atentzio mekanismoaren diseinu irudia.	14
3.5	<i>Transformers</i> arkitekturaren diseinu diagrama <i>Attention is all you need</i> artikulutik aterata.	15
3.6	<i>Embedding</i> multzoak kasu honetan "airports" hitza aukeratu da eta koloreztatuta agertzen dira entrenamenduan zehar gertuago kokatu diren hitzak. <i>Embedding Projector</i> -etik aterata.	16
3.7	BART modeloaren diseinua.	17
3.8	BART modeloaren aurreentrenamenduaren azalpena.	18
3.9	BART ereduaren metrikaren balioak ataza ezberdinetan.	18
3.10	BART modeloaren metrikaren balioak laburpen automatikoan.	18
3.11	MT5ek erabili dituen datuak hizkuntzekiko. MT5 modeloen artikulutik aterata.	19
3.12	PEGASUS modeloen GSG eta MHE entrenamenduen diagrama.	21
3.13	PEGASUS modeloen GSG eta MHE entrenamendu ezberdinen arteko konparaketa. GSGren kasuan esaldien %30 maskaratzeko dira	21
3.14	PEGASUS, artearen egoera eta <i>transformers</i> BASE modeloen arteko alderaketak. Transformers BASE deitutako modeloa, GSG aurreentrenamendurik izan ez duen modeloa da.	21
3.15	MT0 entrenatzeko erabilitako corpusa.	22
3.16	MT0 modeloak entrenatzeko erabilitako testuen kopurua hizkuntzekiko.	22
3.17	MT0 modeloak entrenatzeko erabili diren <i>prompt</i> -en adibideak. Baita testu berdinen P3, xP3 eta xP3mt corpusetako adibidea.	23
3.18	ROUGE-N metrikaren formula	24
3.19	ROUGE-F metrikaren formula	25
3.20	BLEU metrikaren formula	25
3.21	RST-zuhaitzaren deskribapenetik abiatuz laburpenak egiteko bidea [1]	26
3.22	Unai tesian erabilitako irizpideak. [1]	27
3.23	NLLB ereduak entrenatzeko corpusa sortzeko erabilitako datuen iturriak	32
4.1	Datu multzo bakoitzaren itzulpenen BLEU balioak.	38
4.2	"O, ez!" esalditxoaren batz besteko agerpena testu bakoitzeko.	38
4.3	"Ez dut ulertzen" esalditxoaren batz besteko agerpena testu bakoitzeko.	39
4.4	"Amaitu dut" esalditxoaren batz besteko agerpena testu bakoitzeko.	39
4.5	Testuetako token kopuruen diagramak.	41

4.6	Laburpenen token kopuruen diagramak.	41
4.7	PEGASUS-Large modeloa entrenatzeko hiperparametroak [2].	42
4.8	Entrenamenduetan erabilitako LR planifikatzaileak sortzen duen grafika.	44
4.9	FP16 doitasunaren azalpena.	45
4.10	FP32 doitasunaren azalpena.	45
4.11	Train 1 eta Train 1P entrenamenduetan erabilitako <i>scheduler</i> -a.	47
5.1	Train 1 eta Train 1P entrenamenduetako galera balioak.	50
5.2	Train 1 eta Train 1P garapeneko galera balioak.	50
5.3	Train 2Adam eta Train 2Adafactor entrenamenduko galera balioak.	50
5.4	Train 2Adam eta Train 2Adafactor garapeneko galera balioak.	51
5.5	Train 2Adafactor eta Train 2Adam entrenamenduen GPUaren memoria erabilera.	51
5.6	Train 2Adafactor eta Train 2Adam entrenamenduen RAM memoria erabilera.	52
5.7	Train 3 entrenamenduko galera balioak.	52
5.8	Train 3 entrenamenduko garapeneko zatiko galera balioak.	53
5.9	Train 4 entrenamenduko galera balioak Train 2Adafactorrekin alderatuz.	53
5.10	Train 4 entrenamenduko garapen zatiko galerak Train 2Adafactor-ekin alderatuz.	54
5.11	Train 1P eta Train 1 ereduak lortutako ROUGE-F1 balioak.	54
5.12	Train 2Adafactor modeloak lortutako ROUGE1 balioak.	55
5.13	Train 2AdamW modeloak lortutako ROUGE1 balioak.	55
5.14	Train 3Long modeloak lortutako ROUGE-F1 balioak.	56
5.15	Train 4Large modeloak lortutako ROUGE-F1 balioak.	56
5.16	Train 4Large modeloak lortutako ROUGE-TOK balioak.	59
5.17	Wikihow-eko testu hobereenen batz besteko irizpideen balioak.	60
5.18	Wikihow-eko testu okerrenen batz besteko irizpideen balioak.	61
5.19	Elikagaien testuetako batz besteko irizpideen balioak.	63

Taulen aurkibidea

2.1	Ataza bakoitzari eskainiko zaion denbora.	7
2.2	Ataza bakoitzaren denboraldiak.	8
3.1	ROUGE balioen adibideak.	25
3.2	Google-ek BLEU balioei zehaztutako esanahiak	25
3.3	Irizpide bakoitzeko noten esanahiak	31
3.4	Irizpide bakoitzeko zenbakizko notak.	31
4.1	Datu multzo bakoitzeko batz besteko esaldi kopurua.	37
4.2	Arxiv-eko 100 testu hartuta itzultzeko denborak. Luzera max. balioak, esaldiko gehienez egon daitezkeen token kopurua adierazten du.	37
4.3	Datu multzo bakoitzeko banaketen testu kopurua.	41
4.4	Entrenamendu bakoitzak behar izan duen denbora.	46
4.5	Erabilitako testu kopurua datu multzoka.	46
4.6	Train 1 eta Train 1Pren hiperparametroak.	46
4.7	Train 2Adam eta Train 2Adafactor entrenamenduetako hiperparametroak.	47
5.1	Modelo guztiak datu multzoka duten ROUGE balioa. Letra lodiz datu multzo bakoitzeko lortu diren balio hoberenak.	57
5.2	Train 4Large eta PEGASUS ereduaren ROUGE-F1 eta ROUGE-TOK balioak.	59
5.3	Wikihow datu multzoko ROUGE balio altuena duten testuen eskuzko ebaluazioko nota.	61
5.4	Wikihow datu multzoko ROUGE balio baxuena duten testuen eskuzko ebaluazioko nota.	62
5.5	Elikagaien testuetako token-ratio balioak.	62
5.6	Elikagaien testuen eskuzko ebaluazioko nota.	63
6.1	Elikagaien testuen laburpenak <i>prompt</i> ezberdinekin.	67
6.2	AESLCKo testu bat <i>prompt</i> ezberdinak erabiliz laburtuta.	68
A.1	Ataza bakoitzari benetan eskaini zaion denbora.	71
A.2	Ataza bakoitzaren denboraldiak.	72
B.1	Adibideetako testuen ROUGE balioak.	78

Algoritmoen zerrenda

1	Datu multzoen itzulpenerako algoritmoa	36
---	--	----

Sarrera

1.1 Motibazioa

Lengoaia naturalaren prozesamenduak (LNP) gorakada garrantzitsua izan du azken urteetan. Hainbat atazatan emaitza oso onak lortu dira eta aurrerapausu handiak eman dira. Horren arrazoietakoa bat egunero sortzen diren datu kopuru masiboak dira. Baina, honek informazio gainkarga sortzen du eta horri aurre egiteko, laburpen sistema automatikoak tresna bikainak izan daitezke, testutik garrantzitsuena dena erazteko.

Laburpen automatikoak sortzeko sistemen garapena ez da nahi berria, ikerlari asko lanean aritu dira urteetan zehar. Bi adar bereizten dira laburpenak sortzeko garaian. Lehena, estrakzio-laburpenak dira, hauek sarrerako testutik garrantzitsuenak diren esaldiak soilik bilduz osatzen dira. Hau da, ez da berridazketarik edota testu sorkuntzarik egiten. Bigarrena, abstrakzio-laburpenak dira, hauek testuen gai nagusiak biltzen dituzten laburpen berridatziak dira. Kasu hauetan berridazketak dira eta testu sorkuntza barne hartzen da.

LNPko hainbat teknika erabili izan dira bi laburpen motak sortzeko, baina, azken urteetan, *transformer* arkitekturan oinarritutako sistemen garapenekin, hobekuntza oso nabaria izan da. Aurrerapausoa testu sortzaileak diren eruedetatik dator gehien bat. Eredu hauek oso egokiak dira abstrakzio-laburpenak sortzeko, sarreran testua hartu eta irteeran sarrerarekin lotutako testu berri bat sortzen dutelako.

Nahiz eta ataza oso interesgarria izan, zoritxarrez, aurrerapen gehiena ingeleserako egin da. Hizkuntza eredu handiak eta domeinu ezberdinetan entrenatutakoak daude eskura ingeleserako. Adibidez, osasun arloko testuetan entrenatutako ereduak aurki ditzakegu ingeleserako, baina ez baliabide urriko hizkuntzetan. Datu multzoei dagokienez ere, ingelesari begira sortu dira gehienak. Gainera, estandarrak diren laburpenen ebaluaziorako metrika ezagunenak, ingeleseko gramatika kontuan hartuta diseinatu dira. Ondorioz, ingelesa ez den hizkuntzetan arazoak daude metrika hauekin.

Euskara bezalako baliabide gutxiko hizkuntza batean laburpen automatikorako sistema bat garatzea ez da erraza. Euskara dakiten eta testu sortzaileak diren hizkuntza ereduak hizkuntza anitzak izan ohi dira eta entrenamenduan ikusitako euskarazko testu kopurua ez da handia. Eskura dauden laburpenerako corpusak ere ingeleserako dira gehienak eta euskararako ez dago kalitatezko testu kopuru handirik.

Proiektu honetan euskararako abstrakzio-laburpenak sortzen dituen sistema bat garatu nahi da. Ingeleserako artearen egoera den laburpen automatikorako sistemari oinarrituta, euskararako bertsioa sortuz. Eskuragarri dauden baliabide urriak kontuan hartuta, aurrentrenatutako eredu eleanitzak hartu eta laburpenak sortzeko entrenatuko dira. Arestian aipaturiko murriztapenak eta zailtasunak kontuan hartuta, posible al da ingelesezko artearen egoeraren pareko laburpen automatikorako sistema bat garatzea?

1.2 Proiektuaren helburu orokorra

Proiektu honen helburua euskararako lehen abstrakzio-laburpen sistema sortzea da. Horretarako, ingeleserako artearen egoera den PEGASUS ereduak oinarritzat hartuko dira. Behin laburpen sistema sortuta, Unai Atutxaren tesian [1] garatutako irizpideekin ebaluatuko da.

Eskura dauden baliabide mugatuak direla medio, MT0 [3] aurrentrenatutako ereduak euskarazko laburpenak egiteko finduko dira. Eredu hauek testu sortzaileak dira eta euskara ere ikasi dute entrenamenduan. Gainera, ingeleserako laburpen automatikoa ere landu dute. Beraz, interesgarriak izan daitezke euskarazko laburpen testuekin *finetuning* eginda.

Euskarazko laburpen automatikorako corpus egokirik ez dagoenez eskura, ingelesetik itzuliko dira datu multzoak. Itzulpen automatikoa erabiliko da corpusa osatzeko eta ondoren, kalitate neurketa bat egingo zaie testuei, erabilgarriak diren ikusteko.

Corpusa osatu ondoren, laburpen-ereduak entrenatuko dira. Ataza honetarako egokienak diren optimizatzaileak, hiperparametroak edota *prompt-ak* ikertuko dira. Esperimentu ezberdinak eginez eta emaitzak alderatuta erabaki egokienak zein diren aukeratuko dira.

Sistemaren ebaluazioa bi modutara egingo da. Lehena, estandarrak diren metrika automatikoak erabiliz. ROUGE metrika estandarra eta erabiliena da horrelako sistemak alderatzeko. Bigarren, Unai Atutxak sistemaren laburpenak ebaluatuko ditu, bere doktore tesian garatutako irizpideak kontuan hartuta. Ebaluazio honek garrantzi handia izango du sortutako testuen kalitatea zein den zehazteko garaian.

Atazaren zailtasuna euskararako hizkuntza eredu eta testuen urritasuna eta baliabide falta kontuan hartuta, ikerketa moduan planteatu da proiektu hau. Hori dela eta, hasieratik ez da espero produktu edo sistema onik sortzea. Lanarekin lortu nahi dena LNPko ataza honetan euskararako lehen pausuak ematea da, ondoren, GrAL-eko txostenean egindako esperimenduak eta emaitzak ongi dokumentatuta utziko dira.

1.3 Dokumentuaren antolaketa

GrAL-ren memoria hainbat kapitulutan banatuta dago, lehena kapitulu hau sarrera izanik. Jarraian, proiektuaren helburu dokumentua dago, bertan proiektua aurrera eramateko plangintza agertzen da. Hirugarren kapituluan aurrekariak eta memoria ulertzeko behar diren oinarri teorikoak azaltzen dira. Laugarren kapituluan esperimenduekin zerikusia duten atalak daude: Corpusaren sorkuntza eta entrenamenduen deskribapenak. Bosgarren kapituluan lotutako emaitzak eta ebaluazioak azaltzen dira. Bertan, eskuzko ebaluazioa agertzen da. Seigarren kapituluan emaitzetatik ateratako ondorioak aipatuko dira. Azkenik, zazpigarren kapitulua eranskinen atala da, mugak, edukitako arazoak eta etorkizunerako lanak aipatuko dira, baita ere, plangintza eguneratua eta sortutako laburpenen adibideak.

Proiektuaren Helburu Dokumentua

Proiektu bateko plangintza oso garrantzitsua da, lana modu ordenatu eta eraginkor batean egiteko. GrAL-aren kasuan, hilabeteetako proiektu bat izanik, hasierako plangintza eta zehaztapenak ongi egiteak, oinarri sendo bat finkatzen du lan egiteko moduan. Planifikazio oker edo zehaztu gabe batek, proiektuaren helburuak betetzea oztopatu dezake, GrAL-a porrotera eramanez.

Plangintzarekin hasteko, lehenik proiektuaren adarrak, lan-paketeak eta hauek osatzen dituzten atazak finkatu eta hauen garapen denboraldiak zehaztuko dira. Garapen denboraldiekin lotuz, proiektuko mugarrak eta ataza bakoitzari eskainiko zaion denboraren estimazioa egingo da, betiere, GrAL-a burutzeko 300 ordu inguru estimatzen dela kontuan hartuta. Amaitzeko, proiektuan zehar gerta daitezkeen arriskuen analisisa egingo da, bakoitzaren eragina neurtuz eta ahal bada konpontzeko planen bat diseinatuta.

2.1 Lan-paketeak eta atazen identifikazioa

GrAL honetan 4 adar definitu dira. Lehena, kudeaketa adarrari dagokiona. Adar honek, proiektuko plangintza, jarraipena eta txostenaren garapena biltzen ditu. Bigarren, corpus adarra; kasu honetan corpora osatzeko eta aurreprozesatzeko atazak barne hartzen ditu. Hirugarren, entrenamendu eta ereduaren garapena. Corpora osatzearekin batera mardulena izango den adarra da, corpora erabilia eredu egokiak entrenatzea barne hartzen du pakete honek. Azkenik, ikerketa adarra, bertan bilerak, informazio bilketa eta ondorioak lortzeko hausnarketak biltzen dira. Azken honek garrantzi handia du, proiektuaren helburua ez baita eredu funtzional bat sortzea, ikertzea baizik.

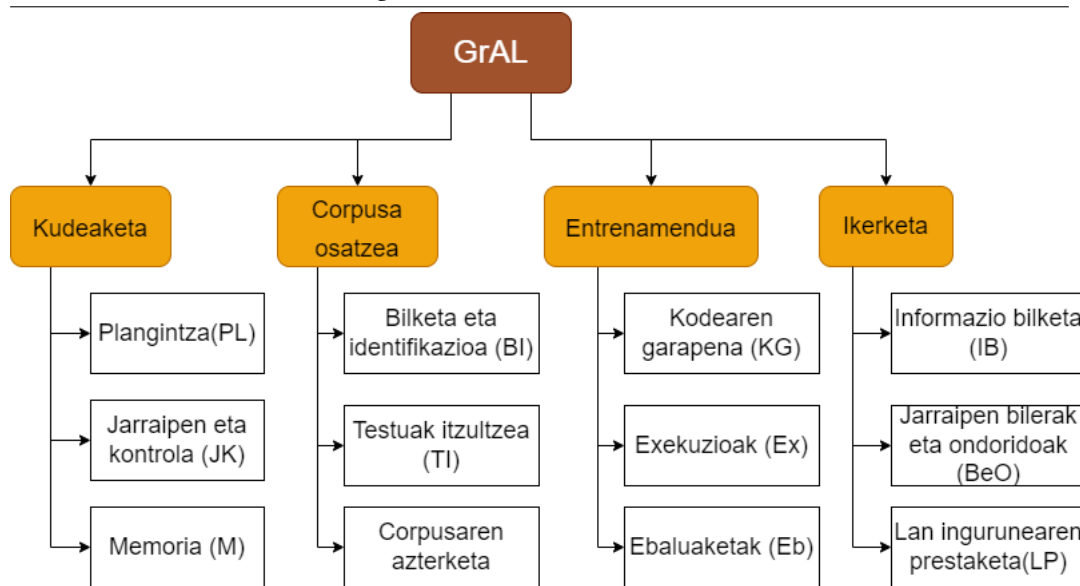
2.1 diagraman ikus daitezke adarrak eta lan-paketeak.

2.1.1 Kudeaketa adarra

Adar honetan, kudeaketarekin lotutako atazak biltzen dira. Proiektuaren plangintza burutzea, GrAL-eko memoria idaztea eta azkenik, proiektuaren jarraipena egitea.

Plangintza (PL): Hasierako plangintza egitea eta behar bada, plangintza egunean mantentzea.

Irudia 2.1 Proiektuko LDE diagrama.



- PL.1: Lan-paketeen eta haiek osatzen dituzten atazen identifikazioa.
- PL.2: Arriskuen analisia egitea eta arriskuei aurre egiteko planak burutzea.
- PL.3: Planifikazioa eta plangintza idaztea.
- PL.4: Plangintza eguneratzea beharrezkoa bada.

Jarraipen eta kontrola (JK): Proiektua aurrera doan heinean, egindako lanak, ondorioak eta erabakiak dokumentatzea eta emaitzak biltzea.

- JK.1: Atazei eskainitako orduen jarraipena egitea.
- JK.2: Arriskuen kudeaketa.

Memoria (M): GrAL-aren amaieran entregatu beharreko memoriaren garapena. Idaztea, erreparatzea eta amaierako defentsaren prestaketa barne hartzen ditu.

- M.1: Proiektuan lortutako ondorioak dokumentatzea.
- M.2: Txostena garatzea eta erreparatzea.
- M.3: GrAL-eko defentsa prestatzea.

2.1.2 Corpora osatzeko adarra

Adar honetan, entrenamenduan erabiliko den corpusarekin lotutako lan-pakete eta atazak daude. Lehen lan-paketea, datu multzo horien bilketa eta identifikazioa izanik. Bigarrena, testuak itzultzea eta azkenik corpora aztertzea.

Bilketa eta identifikazioa (BI): Euskarara itzuli behar diren datu multzoen identifikazioa.

- BI.1: Datu multzo posibleen zerrenda egitea.

- BI.2: Datu multzoen azterketa eta aukeraketa.

Testuak itzultzea (TI): Erabiliko diren datu multzoen itzulketa egitea.

- TI.1: Testuen itzulpenarako kodearen garapena.
- TI.2: Datu multzoak itzultzea.

Corpusaren azterketa (CA): Itzuli diren datu multzoen azterketa egitea.

- CA.1: Itzulpenaren kalitatea neurtzea.
- CA.2: Corpuseko testuak aztertzea.

2.1.3 Entrenamendu adarra

Adar honetan, bildutako corpora erabiliz modeloak entrenatzeko egin behar diren atazak biltzen ditu. Entrenamenduko kodearen garapena, entrenamenduen exekuzioak eta entrenatutako modelo ezberdinen ebaluazioa.

Kodearen garapena (KG): Entrenamenduak egiteko eta ebaluazioak egiteko erabiliko den kodearen garapena.

- KG.1: Entrenamenduko kodearen garapena.
- KG.2: Ebaluaziorako kodearen garapena.

Exekuzioak (Ex): Entrenamendua burutzeko exekuzioak eta entrenamenduen jarraipena.

- Ex.1: Entrenamenduen exekuzioak.
- Ex.2: Entrenamenduen jarraipena eta kontrola. Galera balioak, erroreak etab.

Ebaluaketak (Eb): Entrenatu diren ereduak ebaluatzea. Bai automatikoki, bai eskuzko ebaluazioa.

- Eb.1: Ebaluazio automatikoa.
- Eb.2: Eskuzko ebaluazioa.

2.1.4 Ikerketa adarra

Adar honetan, GrAL-aren ikerketa zatiari dagozkien atazak daude. Ikerketetan ohikoa den bezala, hasieran informazioa bilduko da, ondoren, GrAL-ean zehar bilerak egingo dira, ondorioak eta emaitzak komentatzeko. Exekuzio, proba eta itzulpen lan guztiak IXA ikerketa taldearen babesarekin egin dira, ondorioz, haien makinetan egin dira lan gehienak. IXAko makinetako lan ingurune prestaketarekin lotutako atazak ere adar honetan sartu dira.

Informazio bilketa (IB): Aurrekariak, artearen egoera eta erabil daitezkeen teknikei buruzko informazioa biltzea.

- IB.1: Aurrekariak eta artearen egoera aztertzea.

- IB.2: Erabiliko diren teknika eta ereduaren aukeraketa.

Jarraipen bilerak eta ondorioak (BeO): Lortutako emaitzak eta erabakiak hausnartze-ko bilerak egingo dira.

- BeO.1: Bilerak egitea.
- BeO.2: Bileretatik ateratako ondorioak dokumentatu eta erabakiak aplikatzea.

Lan ingurunearen prestaketa (LP): IXAko makinak erabiltzeko sarbidea eta behar diren lanerako tresnak prestatzea.

- LP.1: IXAko makinatarako sarbidea prestatzea.
- LP.2: Makinetako lanerako ingurunea prestatzea. Ingurune birtualak eta antzerakoak.

2.2 Atazen denborak eta epeak

2.1 taulan, ataza bakoitzari estimatzen zaion denborak agertzen da.

Lehen asteetan, plangintza eta informazio bilaketa egitea espero da, proiektuaren oinarriak ongi finkatzeko. Jarraian, datuak adarreko BI atazak egin eta TI.1eko kodearen garapena hasita espero da. Ondoren, TI.2 eta CA atazak burutzea espero da.

Testuak itzultzen ari diren bitartean, KG atazak burutuko dira, corpora prest dagoenean kodea ere eskuragarri eta akatsik gabe edukitzeko. KG atazak eginik eta corpora sortuta, Ex lan-paketearekin has daiteke.

Exekuzioak egin ahala, Eb.1 eta BeO lan-paketeak egingo dira, horrela, hurrengo exekuzioen nondik norakoak zehaztuko dira. Azkenik, Ex lan-paketea burututzak jota, ereduaren eskuzko ebaluazioa (Eb.2) egingo da. Proiektua, memoria lan-paketeko M.2 eta M.3 atazekin amaituko da.

Plangintza adarreko PL.4, JK eta M.1 atazak modu paraleloan egingo dira proiektu guztian zehar, beraz, ia amaiera arte ez dira guztiz burutuko.

2.2 gantt diagraman, ataza bakoitzaren garapen denboraldiak zehaztuta ikus daitezke. Proiektua ahal bada ekainaren erdialderako amaituta edukitzea espero da eta azken egunetan M.2 ataza amaituta edukitzea eta M.3 egiten hasita.

2.3 Arriskuen analisia

Proiektuan zehar arazoak izateko arriskua dago eta arazo horiek aurrez identifikatu eta eragina txikiagotzeko planak edukitzea garrantzitsua da. Hona hemen identifikatu diren arrisku posibleak eta aurre egiteko edo neurriak hartzeko planak:

Itzulpenen kalitatea baxua izatea: Gerta liteke, itzulitako testuen kalitatea oso baxua izatea. Honek arazo handia suposa dezake eta entrenamenduetan itzulpen kaxkarren eragina handiegia izatea. Aurretik neurri gogorak hartu daitezke, arrisku garrantzitsu bat baita. Jarraian hartuko diren neurriak:

Lan-paketea	Ataza	Denbora (ordutan)
Plangintza	PL.1	1
	PL.2	2
	PL.3	5
	PL.4	2
	Guztira	10
Jarraipen eta kontrola	JK.1	2
	JK.2	3
	Guztira	5
Memoria	M.1	5
	M.2	30
	M.3	5
	Guztira	40
Bilketa eta identifikazioa	BI.1	2
	BI.2	3
	Guztira	5
Testuak itzultzea	TI.1	5
	TI.2	50
	Guztira	55
Corpusaren azterketa	CA.1	2,5
	CA.2	2,5
	Guztira	5
Kodearen garapena	KG.1	7,5
	KG.2	2,5
	Guztira	10
Exekuzioak	Ex.1	100
	Ex.2	20
	Guztira	120
Ebaluaketak	Eb.1	15
	Eb.2	2,5
	Guztira	17,5
Informazio bilketa	IB.1	10
	IB.2	2,5
	Guztira	12,5
Jarraipen bilerak eta ondorioak	BeO.1	25
	BeO.2	5
	Guztira	30
Lan ingurunearen prestaketa	LP.1	1
	LP.2	1
	Guztira	2
Guztira		312

2.1 Taula: Ataza bakoitzari eskainiko zaion denbora.

- Itzulpen automatikoko artearen egoera zein den aztertzea, eredu eta teknika egokienak erabiltzeko.

2. PROIEKTUAREN HELBURU DOKUMENTUA

Ataza	urt.	ots.	mar.	apr.	mai.	eka.
PL.1	■					
PL.2	■					
PL.3	■					
PL.4						■
JK.1		■	■	■	■	■
JK.2		■	■	■	■	■
M.1	■	■	■	■	■	
M.2				■	■	
M.3						■
BI.1	■					
BI.2	■					
TI.1		■	■			
TI.2		■	■	■		
CA.1		■	■			
CA.2		■	■			
KG.1		■				
KG.2		■				
Ex.1			■	■	■	
Ex.2			■	■	■	
Eb.1			■	■	■	
Eb.2			■	■	■	
IB.1	■	■				
IB.2	■	■				
BeO.1	■	■	■	■	■	
BeO.2	■	■	■	■	■	
LP.1	■					
LP.2	■					

2.2 Taula: Ataza bakoitzaren denboraldiak.

- Itzuliko den datu multzo bakoitzaren laginak eskuz aztertzea, akats larriak edo itzulpen arazoak identifikatzeko.
- Ebaluazio automatikorako metrikak erabiltzea, itzulitako testuak ebaluatzeko.

Neurri sendoak izanda ere, arriskua ez da desagertzen eta kontuz tratatu beharreko kontua da. Arrisku honen maila **oso larria** da.

IXAko zerbitzariekin arazoak: Exekuzioak IXA taldeko makinetan egingo direnez, posible da zerbitzariekin arazoak edukitzea eta egun batzuetan erabiltezin geratzea. Arrisku honen aurrean ez dago zer egin handirik, baina, makinetatik kanpora egin daitezkeen exekuzioak Google Colab bezalako plataformetan exekutatzeko kodea prest edukitzea konponbide bat izan daiteke. Eragina **txikia edo ertaina** izan daiteke arazoaren iraupenaren denboraren arabera.

Corpusa edo kodea galtzea: Corpusa eta kodea IXAko zerbitzarietan egongo dira eta gerta liteke akats batzuen ondorioz ezabatu edo galtzea. Arazoa ekiditeko etxeko

makinan segurtasun kopiak egingo dira eta kodea GitHub plataformara igoko da maiz. Arrisku handiko gertaera izan daiteke, baina, neurriak hartuta, berreskurapena ordu pare bateko kontua izango da, ondorioz, arriskua **txikia** dela esan daiteke.

Kodearen dependentziekin arazoak izatea: Kodetzeko garaian, dependentziekin arazoak izatea oso ohikoa da, makina lokaletan lanean aritzean ez du arrisku handirik suposatzen. Hala ere, IXAko zerbitzarietan teknikariekin hitz egin behar da dependentziak instalatu edo eguneratzeko. Honek, hainbat egun har ditzake eta arazoa hainbatetan errepikatuz gero, denbora asko gal daiteke. Arrisku honen aurrean, kodea garatzean dependentziak ongi identifikatzea eta teknikariekin garaiz hitz egitea oso garrantzitsua izango da. Arrisku honen maila **txikia** da, neurri egokiak hartuz gero, egun batzuetan konpontzeko moduko arazoak izango direlako.

Aurrekariak

Kapitulu honetan txostena ulertzeko beharrezkoak diren oinarri teorikoak eta ereduak azalduko dira.

3.1 Hizkuntza ereduak

Azken urteetan Adimen Artifizialak gorakada izugarria izan du eta hainbat ataza burutzeko erabili dira Ikasketa Automatikoa izenez ezagutzen diren teknikak. Ataza horietako batzuk lengoia naturalaren prozesamenduak (LNP) barne hartzen ditu. Tresna oso erabilgarriak sortu dira LNPren baitan, itzulpen sistemak, txat botak edota sentimendu analisiak egiteko sistemak besteak beste. Teknologia hauen oinarrian hizkuntza ereduak (HE) izeneko eredu estatistikoetan oinarritutako modeloak daude.

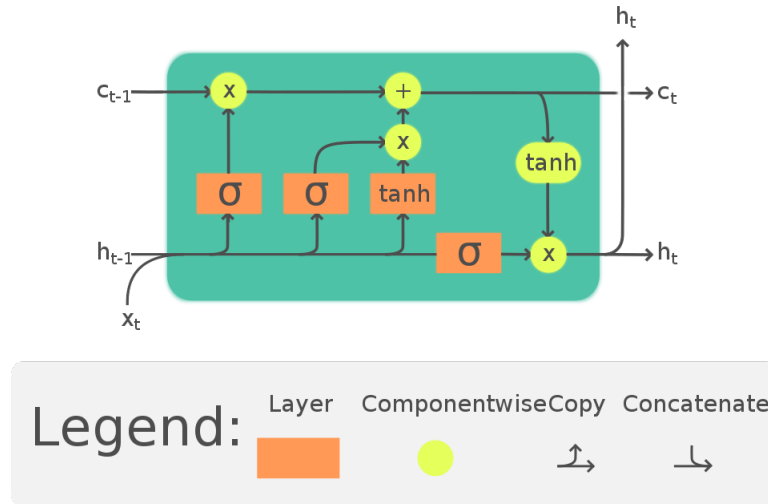
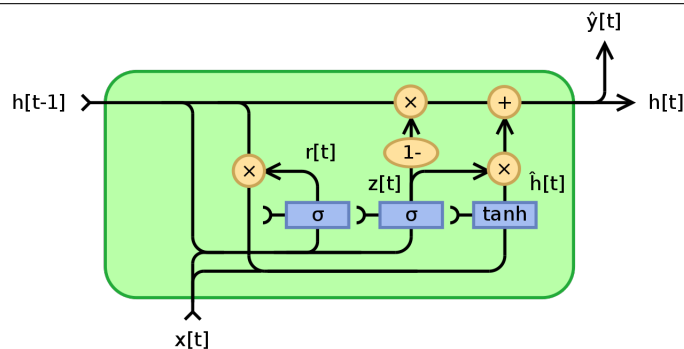
HEak testuen (hitz sekuentzien) gaineko probabilitatea kalkulatzeko dute. Probabilitate horrek sarrerako testua gizaki batek idatzia izateko duen probabilitatea adierazten du. Probabilitatearen katearen erregela bidez kalkulatzeko da [4]. $P(W_1 \dots W_n)$ sarrerako sekuentziaren probabilitatearen formula izanik, probabilitatearen formula honela hedatzen da:

$$P(W_{1:n}) = P(W_1)P(W_2|W_1)P(W_3|W_{1:2})\dots P(W_n|W_{1:n-1}) = \prod_{k=1}^n P(W_k|W_{1:k-1}) \quad (3.1)$$

HEak testuen probabilitateak kalkulatzeko gain, testu sortzaile moduan ere erabil daitezke, $P(W_i||W_1 \dots W_{i-1})$ moduko kalkuluak egiteko gai direnez, probabilitate gehien duen hurrengo hitza zein den asma dezake [4].

$$P(\text{zaude}|\text{kaixo, zer moduz}) \quad (3.2)$$

Honek HEen erabilera esparrua zabaltzen du eta LNP munduko ataza askotan eredu hauen presentzia bermatzen du.

Irudia 3.2 LSTM baten diseinu irudia.**Irudia 3.3** GRU baten diseinu irudia.

Nahiz eta GRUk parametro eta kalkulu gutxiago egin, ez du LSTMa ordezkatu eta atazaren arabera arkitektura bat edo bestea aukeratzen da.

LSTM eta GRU erabilia RNNek zeukaten arazoa txikitzea lortu zen eta HE hobeak sortzea lortu zen. Baina sekuentzia luzeetako arazoak ez ziren guztiz desagertu. Arazoa da sekuentzia osoa bektore batean kodetzen dela eta sarrera luzeak badira, oso zaila izanik informazio guztia ongi kodetzea. Konponbide gisa [9] artikuluan arreta mekanismoa aurkeztu zuten. Honakoa esaten da artikuluan arretaren beharra ulertzeko:

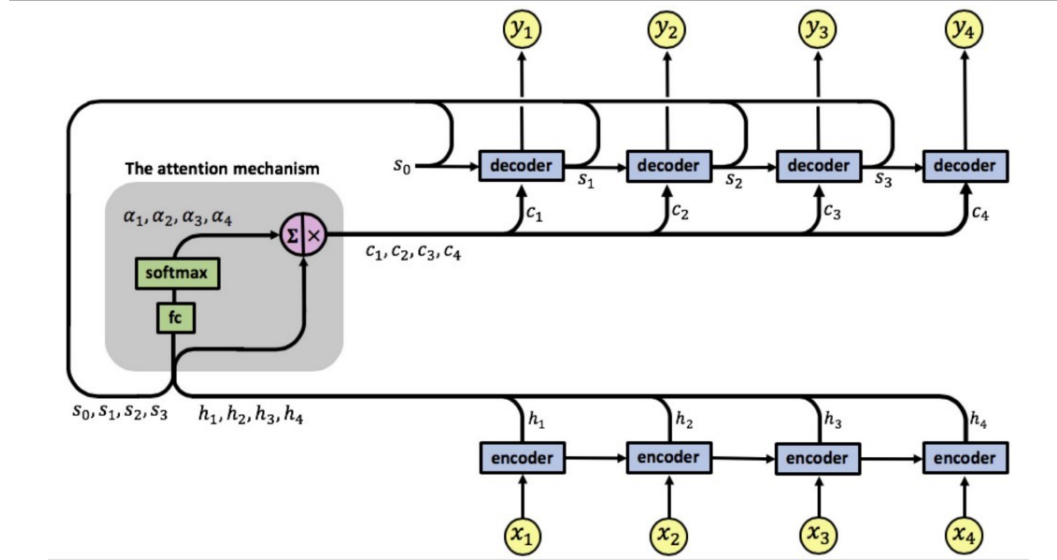
A potential issue with this encoder–decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus.

Arreta mekanismoak sarrerako hitz bakoitzaren kodeketa bektoreari eragiketa batzuen bitartez, garrantzia handiagoa edo txikiagoa ematen dio; ondoren, garrantziaren arabera kalkuluak egiteko. Sareak entrenamenduan zehar ikasiko du arreta ze sarrera motan jarri.

3. AURREKARIAK

3.4 irudian ikus daiteke arreta mekanismoaren implementazio diseinua. Ikusten den bezala, sarrerako hitz bakoitzarentzat arreta balio bana kalkulatzen da eta ondoren kodeketa bektoreekin nahastuz irteerak kalkulatzen dira.

Irudia 3.4 Atentzio mekanismoaren diseinu irudia.



3.1.1 Transformers

2017an *Attention is all you need* artikuluan *Transformers* arkitektura aurkeztu zuten [10]. Arkitektura hau arretan oinarritzen da eta sarrerako balio bakoitzari garrantzi bat esleitzen dio ondorengo kalkuluetarako. LNPra ekarriz Izendun Entitateen Identifikazioa (IEI) mota-tako atazetan lokailuek eta juntagailuek ez dute garrantzirik eta ongi entrenatutako eredu batek arreta bitartez pisu gutxiago esleituko die hitz mota hauei.

Arkitekturaren diseinua 3.5 irudian ikus daiteke. Sare neuronal hauek bi atal dituzte, bata kodetzaile izenez ezagutzen dena eta sarrerako informazioa kodetzeaz arduratzen dena. Besteak deskodetzaile izena hartzen du eta sarrerako informazioa eta kodetzailetik jasotakoa erabiliz, irteerako probabilitateak kalkulatzen ditu. Kodetzaile eta deskodetzaileak arreta mekanismoa ezberdin kalkulatzen dute [10].

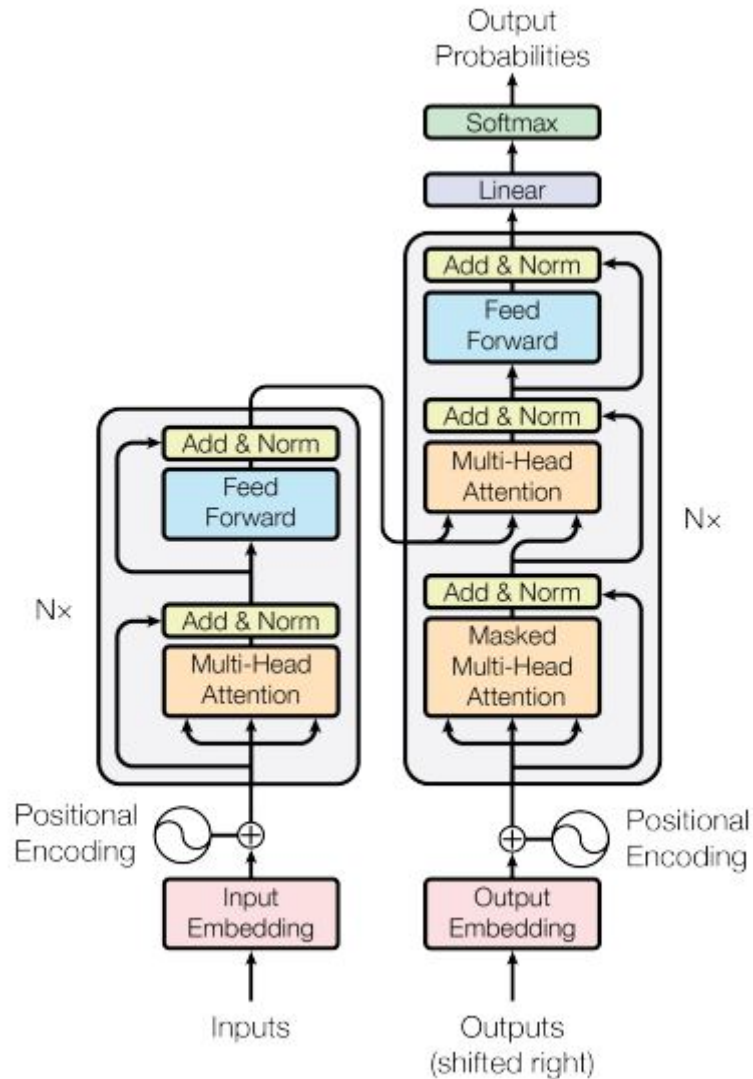
Kodetzaileak autoatentzia erabiltzen du, honek sarrerako sekuentziako kide guztiak elkarren artean konparatzen ditu eta irteerak eraldatzen ditu atentzio balioak erabiliz. Beraz, sarreran 512 hitz badaude, hitz bakoitzeko 512 arreta balio egongo dira, sarrera bakoitza sarrera guztiarekin alderatzen duelako, bere burua barne. Azkenik, arreta balio guztiak erabiliz, sarrerako hitz bakoitzeko irteerako balio bakarra lortzen da.

Deskodetzaileen kasuan arreta gurutzatua erabiltzen da. Autoatentzioaren antzeko funtzionamendua du, baina W_i hitzaren balioak kalkulatzeko $W_{1:i-1}$ hitzak erabiliko dira. Deskodetzaileek beraz ez dituzte hainbeste atentzio balio kalkulatzen, aurreko hitzen atentzioak bakarrik erabiltzen dituzte eta.

LNP munduan sarrerako datuak tokenak izaten dira. Tokenak HE batek hizkuntza prozesatzeko eta sortzeko erabiltzen dituen oinarritzko testu edo kode unitateak dira. HE gehienek tokenekin funtzionatzen dute, baina karakterekazko HEak ere eskuragarri daude.

Testuak tokenetan banantzeko tokenizazio izeneko prozesu bat erabiltzen da. Tokenizazioa gauzatzeko ez dago algoritmo orokorrik eta tokenizatzailerako bakoitzak daukan hiztegia erabiliz gauzatzen du tokenizazioa.

Irudia 3.5 *Transformers* arkitekturaren diseinu diagrama *Attention is all you need* artikulutik aterata.



3.5 irudian sarrerako datuak *Input Embedding* izeneko geruza batetik igarotzen dira, bertan, token bakoitza *embedding* batean bihurtzeko. *Embedding*-ak hainbat dimentsiotako bektoreak izaten dira, zati hau entrenamenduan zehar fintzen joaten da eta erlazioanatu-tako hitzak espazio bektorial horretan gertuago kokatzen ditu [11]. 3.6 irudian hainbat dimentsiotatik 3 dimentsiotara igarotako *embedding* multzoak irudikatu dira.

Embedding-ei kodetzaile eta deskodetzaile ataletara sartu aurretik, posizio-kodeketa izeneko balio batzuekin eragiketak egiten zaizkie. Posizio-kodeketa balioek tokenaren posizioari buruzko informazioa ematen die *embedding*-ei; hori egin ezean, hitzen ordenaren informazioa galduko litzateke. Adibidez, autoatentzioa kalkulatzeko garaian, hitzen ordenaren informazioa ez da mantentzen, baina posizio-kodeketa aplikatuta ordena ere

guztiak kontutan hartu ditzake, probabilitatearen katearen erregela inplementatuz. Hori bai, sarrerako hitz kopurua mugatuta dago RNNetan ez bezala. Baina, RNNekin baina emaitza hobekak lortu dira. [4]

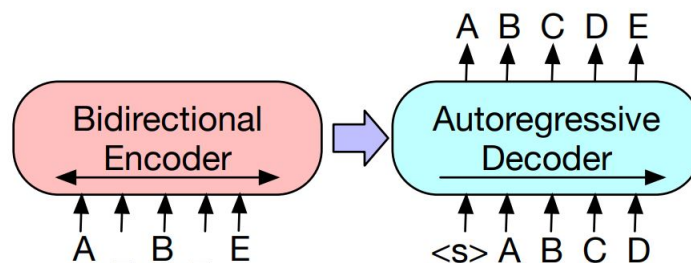
3.1.2 Kodetzaile-Deskodetzaileak

Transformer arkitekturretako bi zatiak, kodetzaile eta deskodetzaileak, era independentean erabili daitezke. Kodetzaileak testu edo tokenen sailkapenak egiteko erabili daitezke. Deskodetzaileak aldiz sarreraren galdera bat eginda irteeran galderaren erantzuna itzul dezake. Baina, LNPn ataza batzuetan bi zatiak erabiltzen dira, adibidez, itzultzaileak diren modeloak, lehenik, sarrerako testua kodetu eta ondoren itzulpena sortzen dute. Deskodetzaileak bakarrik diren modeloek aldiz, arretarekin arazoak erakutsi dituzte irteerako testua idatzi ahala. Eredu hauek sarrerari arreta balio txikiagoak ematen dizkiote eta ez da egokia itzulpen automatikoa bezalako atazetarako [15]. Horregatik, kodetzaile-deskodetzaile arkitekturak erabiltzen dira sarrerarekin menpekotasun handia duten atazetan.

Transformer mota hauek sarrera bi zatitatik jasotzen dute, ia berdina izanik, token berezi batzuk izan ezik. Kodetzailearen lana sarrerako informazioa modu egoki batean kodetu eta ulertzea da, ondoren, deskodetzaileari pasatzeko. Deskodetzaileak aldiz, sarrerako informazioa prozesatu, kodetzailearen irteera jaso eta irteerako tokenen probabilitatea itzultzen du [10]. Nahiz eta testu sorkuntzan oso erabiliak izan, kodetzaile eta deskodetzaile motako *transformer*-ek egiten dituzten atazetan ere erabili daitezke [16].

Kodetzaile-Deskodetzaile arkitekturan oinarritutako HE baten adibidea, BART (Lewis et al.) izenez ezagutzen den eredu da eta [17] artikuluan aurkezten dute. Eredu honen arkitektura [10] artikuluan aurkezten dutenaren berdina da aktibazio funtzioaren bat aldatuta, 3.7 irudian diseinu diagrama ikus daiteke. Eredu honek mota ezberdinetako aurreentrenamenduak izan ditu, 3.8 irudian aurreentrenamenduen azalpen labur bat ikus daiteke. Entrenatzeko erabilitako corpusa BERT ereduak entrenatzeko erabilitako berdinak izan ziren, ingelesezko 160GB inguruko corpus erraldoi bat.

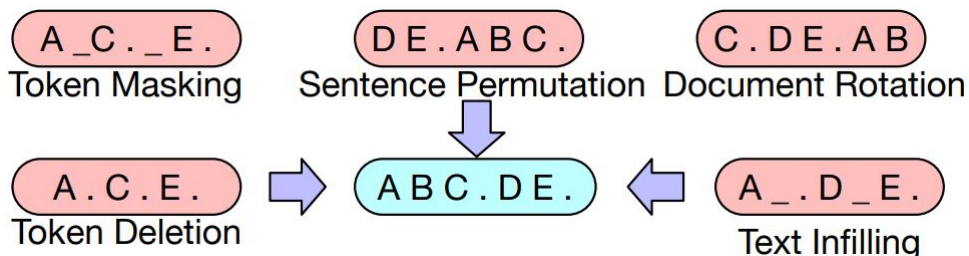
Irudia 3.7 BART modeloaren diseinua.



BART eredu hainbat ataza burutzeko entrenatu zuten. Adibidez, laburpen automatikoa edota galdera-erantzunetan. Lortutako metrikaren balioak ordura arteko artearen egoeraren (State of the Art edo SOTA ingelesez) emaitza antzekoak edo hobekak izan ziren. 3.9 eta 3.10 irudietan hainbat atazetan BART ereduaren metrikak ikus daitezke eta RoBERTa-ren antzeko balioak lortzen ditu ataza batzuetan, baina, laburpenetan ordura arteko artearen egoera baina hobekak.

3. AURREKARIAK

Irudia 3.8 BART modeloaren aurreentrenamenduaren azalpena.



Irudia 3.9 BART ereduaren metriken balioak ataza ezberdinetan.

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	89.0/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/ 94.6	86.5/89.4	90.2/90.2	96.4	92.2	94.7	92.4	86.6	90.9	68.0
BART	88.8/ 94.6	86.1/89.2	89.9/90.1	96.6	92.5	94.9	91.2	87.0	90.4	62.8

Irudia 3.10 BART modeloaren metriken balioak laburpen automatikoan.

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	44.16	21.28	40.90	45.14	22.27	37.25

3.1.3 MT5: Hizkuntza eredu eleanitza

MT5 modeloak kodetzaile-deskodemtzaile motako *transformer*-ak dira [18]. Sarrerako tokenak hartu eta irteeran ere tokenak itzultzen dituzte. MT5 ereduak T5 eredu elebarkarrean oinarritzen dira eta ikasketa eta helburu berdina dute. T5 ereduak ingeleserako bakarrik sortutako kodetzaile-deskodemtzaileak dira eta LNPko hainbat atazatan entrenatuta daude [16]. Entrenamendurako C4 (*Common Crawl's web crawl corpus*) corpusa erabili zuten.

MT5 modeloak eleaniztunak dira, hau da, hainbat hizkuntzatan entrenatuta daude, nahiz eta adibidez, ingelesez datu gehiago ikusi dituen Euskaraz baino. Guztira 101 hizkuntzatan entrenatu da, Euskara barne.

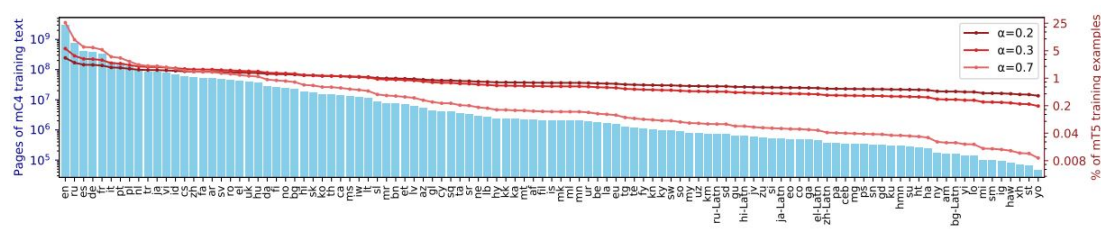
mC4 Corpusa erabilia entrenatuak izan dira, hau C4 corpusa handiagotuz sortu zuten, ingelesezko testuak bakarrik dituen corpusa. mC4-en kasuan 100 hizkuntzetako webeko orrialdeak gehitu zituzten, cld3¹ erabilia orrialdeetako hizkuntza zehaztuz. *Web scrape*-a eginda, guztira 107 hizkuntza bildu zituzten, baina azken 6ak baztertu zituzten hizkun-

¹<https://github.com/google/cld3>

tza batzuen aldaera batzuk zirelako (Errusiarra alfabeto zirilikoarekin eta latindarrarekin agertzen da eta ziriliko baztertu zen).

mC4-ean hizkuntzen agerpen kopurua oso ezberdina da, ingelesez datu gehiago daude euskararekin alderatuta. Horren aurrean, entrenamenduan ingelesez gehiago ikasiko luke euskaraz baino. Hau ekiditeko, $p(L) = |L|^\alpha$ probabilitatea erabiliz lagintzen dira testuak, non $P(L)$ L hizkuntzatik testu bat lagintzeko probabilitatea den eta $|L|$ hizkuntza horretako datu kopuruak. α balioa handiagotuz, benetako distribuziora gehiago gerturatzen da eta txikiagotuz uniformeago bihurtzen da laginketa. mC4 corpusaren datu kopurua eta α balioarekin lagintzeko distribuzioa 3.11 irudian ikus daiteke.

Irudia 3.11 MT5ek erabili dituen datuak hizkuntzekiko. MT5 modeloen artikulatik aterata.



MT5 modeloak ez daude laburpen automatikoa (LA) atazetan entrenatuta eta galdera-erantzun eta IEI atazetako emaitzak soilik aurkezten dituzte artikuluetan. Eredu hauen helburua eleaniztunak diren, aurreentrenatutako kodetzaile-deskodemak argitaratzea da, ondoren, erabiltzaileek *finetuning* egiteko behar duten atazetan.

3.2 Laburpen automatikoko sistemak

Automatic text summarization: A comprehensive survey artikuluan LA ataza burutzeko erabili diren metodo, algoritmo eta sistemen errepaso egiten da [19]. LA LNPko ataza konplexuenetako bat da eta urteetan ikertzaile asko saiatu dira testuak laburtzeko sistemak sortzen. Lehen, 1958an Luhn izeneko Alemaniar baten lana izan zen, non estatistikan oinarrituta artikulua eta aldizkarietako testuak laburtzeko [20] sistema bat diseinatu zuen.

Aurrekarien azalpenekin hasi aurretik, bi laburpen mota bereizi behar dira. Lehen, estrakzio-laburpenak, jatorrizko testutik esanguratsuenak diren esaldiak hartu eta laburpena osatzeko konkatatu egiten dira, beraz, testutik esaldiak kopiatzen dira eta testutik kanpoko hitzik ez da erabiltzen. Bigarrena, abstrakzio-laburpenak, hauek jatorrizko testuetako ideia nagusiak barne hartzen dituzten laburpenak dira, kasu honetan, ez dira esaldiak aukeratu, ideiak baizik eta idazten den testua originala izan behar da, hau da, esaldi berriak sortu behar dira, ideia nagusiak azaltzen dituztenak.

Hainbat algoritmo eta sistema definitu dira bi laburpen motak sortu ahal izateko. Azken urteetan, Ikasketa Automatikoa eta Ikasketa Sakona teknikak erabili izan dira. Estrakzio eta abstrakzio laburpen sistemen garapena era independentean landu dira, ataza ezberdinak direlako:

Estrakzio-laburpenak: Mota hauetako laburpenak egiteko, ML erabiliz, sailkapen arazo bat bezala aurre egin zaio. Testua esaldika banatzen da eta ondoren, esaldi bakoitza laburpenean egongo den edo ez erabakitzen da. Adibidez, [13] artikuluan, BERT *transformer* bati *finetuning* egiten diote estrakzio-laburpenak egiten ikas dezan.

Abstrakzio-laburpenak: Abtrakzioen kasuan, sekuentziatik sekuentziara (seq2seq) motako modeloak erabili izan dira. Hauek, sarreran testu bat hartu eta irteeran testuaren abstrakzio-laburpena itzultzen dute. Ataza konplexua da, sortzen den testuak ideia nagusiak bildu behar dituelako. Adibidez, [21] artikuluan, LSTMak erabiltzen dituzte testuen abstrakzioak egiteko. Azken urteetan, *transformer* arkitektura erabili da gehien bat mota honetako laburpenak egiteko, zehazki kodetzaile-deskodematzaile motako *transformer*-ak. [17] artikuluan aurkezten duten BART eredu abstrakzio-laburpenak egiteko entrenatutako kodetzaile-deskodematzaile bat da.

3.2.1 PEGASUS: Artearen egoerako sistema

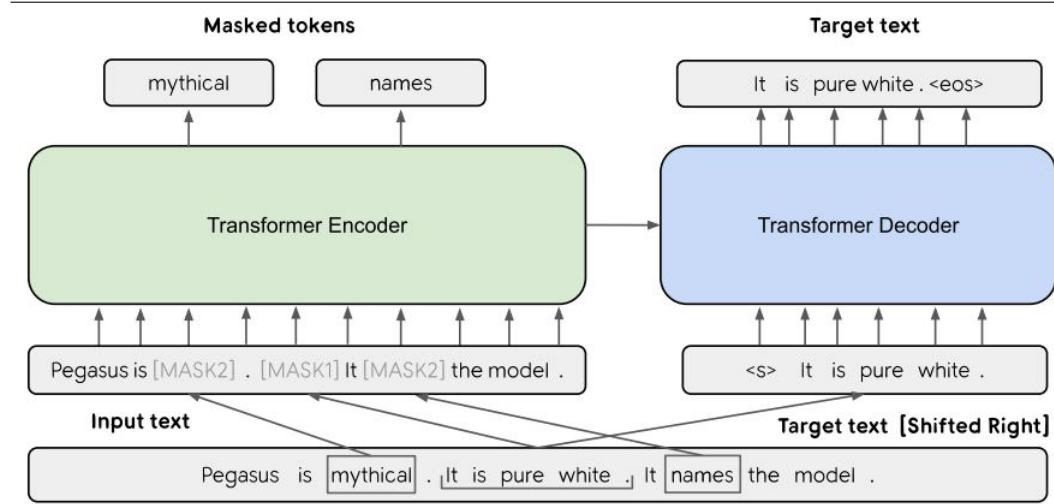
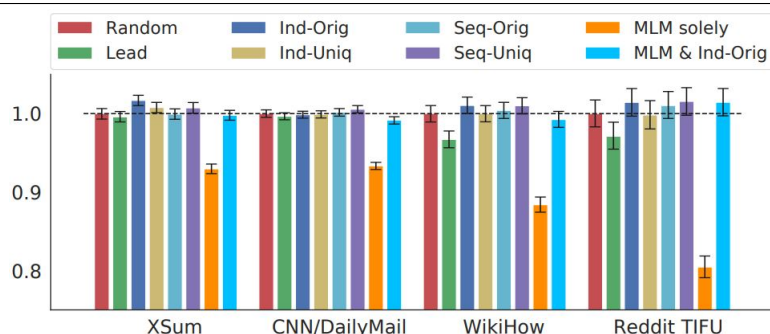
PEGASUS modeloak LA atazarako helburuarekin entrenatu dituzte [2]. Modelo hauek ere kodetzaile-deskodematzaile motakoak dira, beraz, irteeran sarrerako testuaren laburpen abstraktiboa itzultzen dute.

PEGASUSek T5 eredu arkitektura bera jarraitzen dute. Kasu honetan, ez dute aurreentrenatutako oinarritzko hizkuntza eredu bat aukeratu oinarri gisa, modeloak zerotik entrenatu dituzte. Ohiko Maskaratutako Hizkuntza Eredu (MHE edo MLM ingelesez) entrenamenduz gain, *Gap Sentences Generation* (GSG) deitu dioten moduan ere entrenatuta daude. GSGren kasuan, esaldiak maskaritzen dituzte eta maskaratutako esaldiak konkatenatuz, helburuko testua sortzen dute. GSG ikasketaren helburua LAren antzekoa den ataza zail bat ikastea da, maskaratutako esaldiak idaztea, ondoren LAan emaitza hobekak lortzeko ideiarekin.

3.12 irudian ikus daiteke entrenamenduko adibide bat, MASK1 token bereziarekin esaldi bat maskaratu dela adierazten da eta MASK2rekin MHE erabilita token horiek maskaratu direla. Kodetzaileak MASK2 tokenak sortu behar ditu eta Deskodematzaileak berriz MASK1 maskaratutako esaldiak. Maskaratuak izango diren esaldiak aukeratzeko hainbat modu daude: ausaz, testuaren hasierako esaldiak edota esaldi esanguratsuenak hartuz egin daiteke. Esaldi esanguratsuenak aukeratzeko ROUGE-F1 metrika erabili zuten eta testu bateko esaldien %30a maskaratu zuten. Probak egin ondoren, egokiena esaldi esanguratsuenak hartzea zela ikusi zuten eta GSG entrenamendua soilik aplikatuta, emaitza egokiak lortzen zituztela. MHE aplikatuta berriz, irabazia oso txikia da. Azken hau 3.13 irudian ikus daiteke, Ind etiketa dutenek, esaldiak aukeratzean ROUGE-F1 balioa independenteki kalkulatzen dela esan nahi du, Seq diren kasuan, esaldi multzoak hartzen dituzte metrika kalkulatzeko. Uniq eta Orig-en kasuan, Uniq-ek ROUGE-F1 kalkulatzeko berdinak diren N-gramak behin bakarrik hartzen ditu kontuan eta Orig-ek berriz agerpen kopuru guztiak.

Aurreentrenamendua egin ondoren, LA atazako hainbat datu multzo hartu zituzten. Batzuk artikulua, besteak korreo eta istorioak eta azkenik Wikihow-eko gomendioak ere erabili zituzten eta GSG aurreentrenamendurik gabeko *transformer* batekin, artearen egoera zenarekin eta PEGASUS-en modelo ezberdinak alderatu zituzten. 3.14 irudian argi ikusten da GSGaren eragina oso ona dela ataza honetarako, artearen egoera den modeloarekin alderatuta ere PEGASUS Large modeloak hobekak direla ikusten da. Bi Large agertzen dira irudian, aurreentrenamenduan erabilitako Corpusen arteko aldea ere ikusteko.

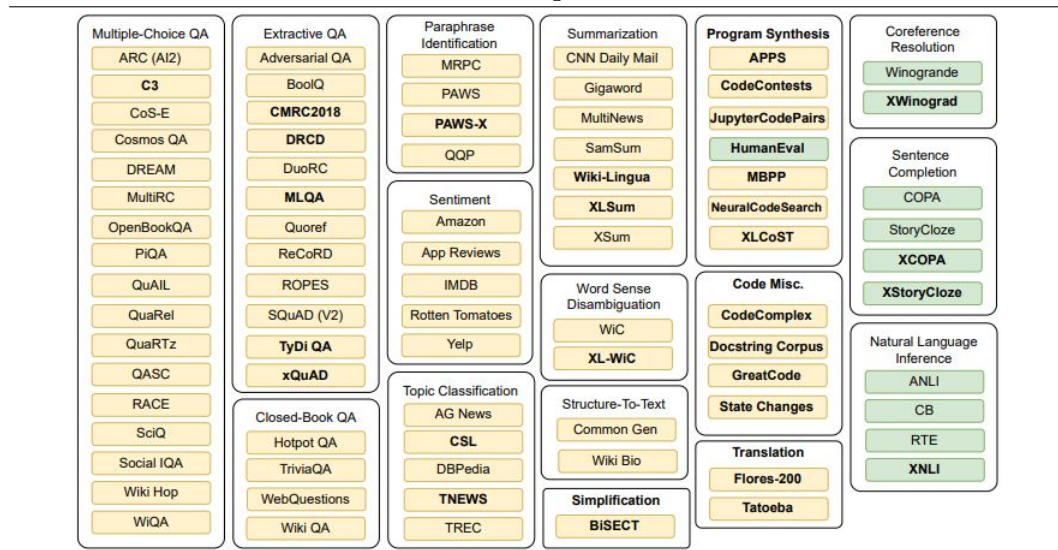
PEGASUS modeloak artearen egoera baino hobekak direla ikusten da entrenatu dituzten datu multzoetan eta GSG entrenamenduak eragin oso ona duela ikus daiteke. Modelo hauek ingelesez bakarrik daude entrenatuta, beraz, gainontzeko hizkuntzetan ezingo lirerateke erabili.

Irudia 3.12 PEGASUS modeloen GSG eta MHE entrenamenduen diagrama.**Irudia 3.13** PEGASUS modeloen GSG eta MHE entrenamendu ezberdinen arteko konparaketa. GSGren kasuan esaldien %30 maskaratzeko dira**Irudia 3.14** PEGASUS, artearen egoera eta *transformers* BASE modeloen arteko alderaketak. Transformers BASE deitutako modeloa, GSG aurreentrenamendurik izan ez duen modeloa da.

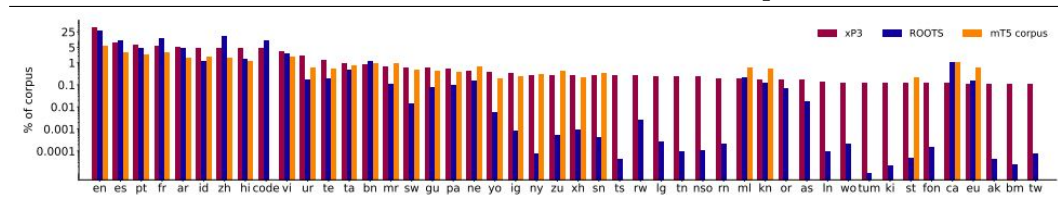
R1/R2/RL	Dataset size	Transformer _{BASE}	PEGASUS _{BASE}	Previous SOTA	PEGASUS _{LARGE} (C4)	PEGASUS _{LARGE} (HugeNews)
XSum	226k	30.83/10.83/24.41	39.79/16.58/31.70	45.14/22.27/37.25	45.20/22.06/36.99	47.21/24.56/39.25
CNN/DailyMail	311k	38.27/15.03/35.48	41.79/18.81/38.93	44.16/21.28/40.90	43.90/21.20/40.76	44.17/21.47/41.11
NEWSROOM	1212k	40.28/27.93/36.52	42.38/30.06/38.52	39.91/28.38/36.87	45.07/33.39/41.28	45.15/33.51/41.33
Multi-News	56k	34.36/5.42/15.75	42.24/13.27/21.44	43.47/14.89/17.41	46.74/17.95/24.26	47.52/18.72/24.91
Gigaword	3995k	35.70/16.75/32.83	36.91/17.66/34.08	39.14/19.92/36.57	38.75/19.96/36.14	39.12/19.86/36.24
WikiHow	168k	32.48/10.53/23.86	36.58/15.64/30.01	28.53/9.23/26.54	43.06/19.71/34.80	41.35/18.51/33.42
Reddit TIFU	42k	15.89/1.94/12.22	24.36/6.09/18.75	19.0/3.7/15.1	26.54/8.94/21.64	26.63/9.01/21.60
BIGPATENT	1341k	42.98/20.51/31.87	43.55/20.43/31.80	37.52/10.63/22.79	53.63/33.16/42.25	53.41/32.89/42.07
arXiv	215k	35.63/7.95/20.00	34.81/10.16/22.50	41.59/14.26/23.55	44.70/17.27/25.80	44.67/17.18/25.73
PubMed	133k	33.94/7.43/19.02	39.98/15.15/25.23	40.59/15.59/23.59	45.49/19.90/27.69	45.09/19.56/27.42
AESLC	18k	15.04/7.39/14.93	34.85/18.94/34.10	23.67/10.29/23.44	37.69/21.85/36.84	37.40/21.22/36.45
BillSum	24k	44.05/21.30/30.98	51.42/29.68/37.78	40.80/23.83/33.73	57.20/39.56/45.80	57.31/40.19/45.82

3. AURREKARIAK

Irudia 3.15 MT0 entrenatzeko erabilitako corpusa.



Irudia 3.16 MT0 modeloak entrenatzeko erabilitako testuen kopurua hizkuntzekiko.



3.2.2 MT0: Ataza anitzeko eredu sortzailea

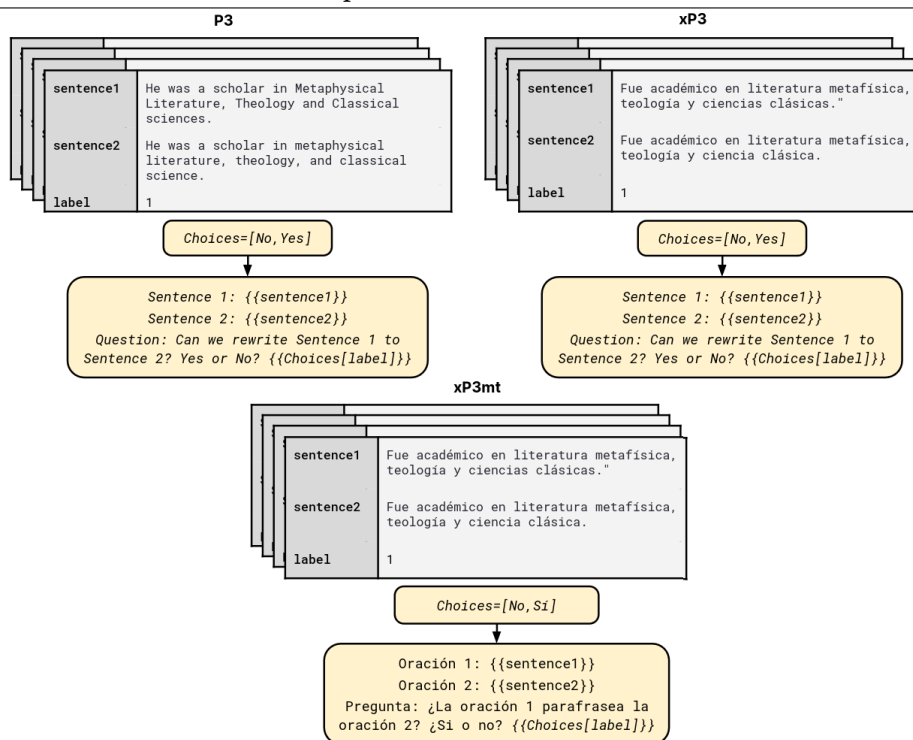
MT0 ereduak T0 HEetan oinarritzen dira, MT5ak T5ean oinarritzen diren bezala. T0 ereduak *prompt*-etan oinarrituta hainbat ataza egiteko entrenatuta daude galdera-erantzun atazetan edota LAN. Zehazki 12 atazatan entrenatu dira 62 datu multzo erabiliz [22]. GPT eta T5 ereduarekin alderatuta, emaitza hobeak lortu dituzte T0 ereduarekin.

MT0 modeloak MT5 modeloen *finetuning*-ak dira. Aurreentrenamendua aprobetxatuz, MT0ak hainbat atazatan entrenatu dituzte eta *prompt* ezberdinak sortu dituzte ataza eta testu motak kontuan hartuta, T0ren kasuan bezala [3].

Entrenamendurako hainbat corpus erabili dituzte, hainbat ataza barne hartuz. 3.15 irudian ikus daiteke erabili dituzten datu multzo guztiak, baina, berde kolorea dutenak ebaluaziorako utzi dituzte. xP3 izena eman diote sortu duten Corpus honi eta 46 hizkuntza ezberdin daude bertan, euskara bat izanik, 3.16 irudian ikus daitezke hizkuntza bakoitzeko datu kopuruak. GrAL honetarako garrantzitsua da LA atazak ere ikusi dituela azpimarratzea, nahiz eta ataza honetarako erabilitako datu multzo guztiak ingelesez edo euskara ez den hizkuntzetan egon.

xP3 corpusean ingelesezko *prompt*-ak bakarrik agertzen direnez, nahiz eta datuak beste hizkuntzetan ere egon. xP3mt izeneko Corpus berri bat sortu dute, automatikoki itzulitako *prompt*-ekin. xP3mt erabiliz modelo bat entrenatu dute eta ingelesezkoak ez diren hizkuntzetarako emaitza hobeak lortu dituzte. Ingelesaren kasuan, emaitza txarragoak lortu dira. Nahiz eta modelo oso interesgarria izan, eskuragarri bakarra jarri zuten MT0-xxl-mt

Irudia 3.17 MT0 modeloak entrenatzeko erabili diren *prompt*-en adibideak. Baita testu berdinaren P3, xP3 eta xP3mt corpusetako adibidea.



izenekoa eta 13 mila milioi (13B amerikar) parametro ditu, honek baliabide asko eskatzen ditu.

Eredu hauek *prompt* ezberdinak ikasi dituzte ataza bakoitzaren arabera. Galdera-erantzun atazarako *prompt*-en erabilera nolakoa den erakusten dute 3.17 irudian. Bertan, *prompt*-ak sarrerako testuan duen egitura ikus daiteke, baita, P3, xP3 eta xP3mt corpusetako testu beraren adibide bat. *Prompt*-ak konplexuak erabil daitezke irudikoak bezala, baina, sinpleagoak ere badaude, ataza eta testu motaren arabera *prompt* egokiak aukeratu behar dira.

LA atazari dagokionez, MLSUM datu multzoko metrikak bakarrik aurkezten dituzte artikuluan, gainera, gaztelaniako zatikoak bakarrik; beraz, ez dira PEGASUS ereduakin konparagarriak, ez baitira PEGASUSak datu multzo horretan ebaluatu. MT0an erabilitako *prompt* ezberdinak datu multzo bakoitzarentzat ikusgai daude artikuluan, hau interesgarria da GrAL honetarako testu mota ezberdinentzat *prompt* bereziak aukeratuta emaitza hobeak lortu daitezkeelako.

3.3 Automatikoki sortutako testuen ebaluazioa

3.3.1 N-grametan oinarritutako ebaluazio metrika automatikoak

Automatikoki modeloekin sortutako testuak erreferentziazko testuekin alderatzeko, N-grametan oinarritutako metrikak erabiltzen dira. Metrika hauek testu batean agertzen diren N-gramak bestean zenbat aldiz agertzen diren kontatzen dituzte [23]. Oso erabiliak

dira adimen artifizialekin sortutako testuak ebaluatzeko eta hainbat atazetan metrika estandarrak dira.

Kasu honetan N-grametan oinarritutako bi metrika landuko dira. Bata, ROUGE izeneko, laburpen automatikorako atazarako erabiltzen dena. Bestea berriz, BLEU, itzulpenen kalitatea neurtzeko erabiltzen dena.

3.3.1.1 ROUGE: Laburpenak automatikoki ebaluatzeko metrika

Abstrakzioak automatikoki ebaluatzeko 2004ean ROUGE metrika (Chin-Yew et al.) aurkeztu zuten [23]. Metrika honek, laburpen bat erreferentziazko laburpen batekin alderatzen du eta 0 eta 1 arteko balio bat ematen dio. Bi testuak zenbat eta antzekoagoak izan, orduan eta balio altuagoa ezarriko dio.

Metrika honek erreferentziazko testuan agertzen diren N-gramak, modeloak sortutako testuetan zenbat aldiz agertzen diren kontatzen ditu. Kalkulu honek estaldura izena hartzen du.

Hainbat ROUGE metrika ezberdin aurkeztu zituzten artikuluan, baina, ezagunenak honakoak dira:

ROUGE-N: ROUGE-N metrikak, n-gramak alderatzen ditu 3.18 irudiko formula erabiliz. Ikus daitekeenez, laburpenean eta erreferentziazko testuan dauden N-gramak kontatzen ditu eta adostasuna kalkulatzen du. Bi testuak berdinak izango balira, ROUGE-N balioa 1 izango litzateke. Ohikoa izaten da, Nren balioa handiagotuta ROUGE-N balio txikiagoak lortzea.

ROUGE-L: ROUGE-L-k ROUGE-N-ren antzeko funtzionamendua du, baina, kasu honetan N-grama luzeenak hartzen ditu kontuan eragiketak egiteko. Longest Common Subsequence (LCS) bitartez egiten dira kalkuluak, beraz, ez da N-gramen luzera finkatu behar.

Nahiz eta ROUGE balioa estaldura moduan kalkulatu, bi aldaera gehiago ere erabiltzen dira. Lehena, ROUGE doitasun balioa da, hau estaldura bezala kalkulatzen da, baina, kasu honetan sortutako testuan agertzen diren N-gramak erreferentziazko testuan duten agerpen kopurua hartzen da kontuan. Bigarren aldaera, aurreko bien arteko kalkulua da eta F-puntuazio izenez ezagutzen da. 3.19 irudian dago ROUGE-FN izenez ezagutzen den metrikaren formula. N balioa berriro ere N-gramen tamaina adierazteko erabiltzen da.

Irudia 3.18 ROUGE-N metrikaren formula

$$\begin{aligned} & \text{ROUGE-N} \\ &= \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \end{aligned}$$

ROUGE balioaren adibide gisa, 3.1 taulan ikus daitezke probak eta balioak.

Irudia 3.19 ROUGE-F metrikaren formula

$$2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Testua1	Testua2	ROUGE-F1
Kaixo, proba egiteko esaldia da	Egun on, probarako testua da	0,2
Kaixo, proba egiteko esaldia da	Egun on, proba egiteko esaldia da	0,72
Kaixo, proba egiteko esaldia da	Kaixo, proba egiteko esaldia da	1,0

3.1 Taula: ROUGE balioen adibideak.**3.3.1.2 BLEU: Itzulpenak automatikoki ebaluatzeko metrika**

Metrika hau itzulpenak automatikoki ebaluatzeko erabiltzen da, 2002an aurkeztu zuten (Kishore et al.) [24]. Metrika honek ere 0 eta 1 arteko balioak zehazten ditu, N-gramen kontaktak eginez.

Metrika honen formula 3.20 irudian ematen da. Bertan, *clip* kontzeptua agertzen da. *Clip* egiteak N-grama berdina hainbat aldiz agertzen denean behin bakarrik kontaktzea eragiten du. Adibidez, "Bai, egia da hori" esaldia eta "Bai bai bai" alderatzean, bigarrenengoaren lehen "Bai" hitza kontuan hartuko da, bestela BLEU balioa 1ekoa izango litzateke.

Irudia 3.20 BLEU metrikaren formula

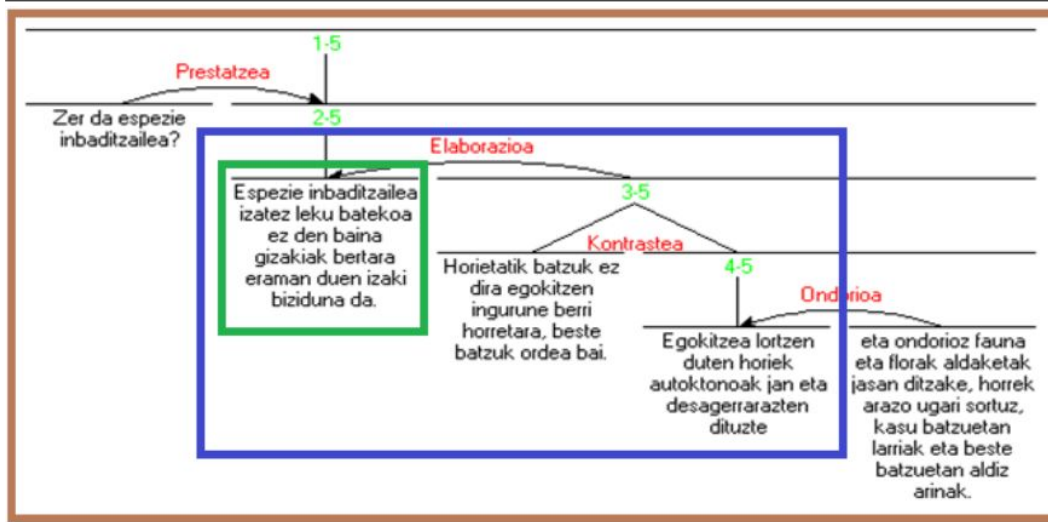
$$\frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-gram}' \in C'} Count(n\text{-gram}')}$$

BLEU kalkulatzeko N parametroari balio bat esleitu behar zaio, baina orokorrean erabiltzen dena 4-gramak dira. N balio ezberdinen ponderazioa ere erabil daiteke eta ondoren pisu ezberdinak esleituta BLEU balio bakarra kalkulatu.

Google-ek BLEU metrika ezberdinei esanahia jartzen die, 3.2 taulan ikus daitezke esanahiak.

BLEU balioak	Esanahia
< 0.1	Ia alferrikakoa
0.1 - 0.19	Esanahia atzemateko zaila
0.2 - 0.29	Esanahia argia da, baina akats gramatikal esanguratsuak ditu
0.3 - 0.4	Ulergarria
0.4 - 0.5	Kalitate handiko itzulpenak
0.5 - 0.6	Kalitate oso handiko itzulpenak eta egokiak
> 0.6	Giza kalitatea baino hobea

3.2 Taula: Google-ek BLEU balioei zehaztutako esanahiak

Irudia 3.21 RST-zuhaitzaren deskribapenetik abiatuz laburpenak egiteko bidea [1]

3.3.2 Eskuzko ebaluazioa: Unai Atutxaren tesia

Gaur egun, laburpenak egiteko gaitasuna izatea oso garrantzitsua da, ondorioz, hezkuntzan laburpen gaitasunak, sintesia, ideia zentralaren identifikazioa etab. lantzea ezinbestekoa litzateke [1]. Hala ere, ikasketa prozesuan laburpena lantzeko jarduerak oso murriztak eta jarraikortasunik gabeak izan ohi dira, ondorioz, ikasleen laburtzeko gaitasunak ez dira behar bezala lantzen eskoletan.

Irakaslearen aldetik laburpenen eskuzko ebaluazio egitea oso lan neketsua eta zaila izan ohi da. Lehenik, laburtze prozesua ebaluatu beharko litzatekeelako eta bigarren laburpen testua bera. Laburpenen ebaluazioa oso neketsua da, baina, prozesua ebaluatzea ikasle bakarrarentzat ere bereziki zaila da, are gehiago, atzeraelikadura egokia eman nahi bada.

Unai Atutxak bere tesian laburpen-gaitasunak lantzeko oinarri teoriko eta baliabide didaktikoak proposatzen ditu. Tresna hauen bitartez, irakasleari laburtze prozesuko ebaluazio karga arindu nahi da eta ikasleak modu egokituan atzeraelikadura jasotzeko baliabideak eskaini nahi dira.

Testuen egitura eta koherentzia deskribatzeko *Rhetorical Structure Theory* (RST) izenez ezagutzen den teoria erabili du Unai [25]. RST testu-antolaketaren teoria bat da, testu-zatien artean dauden erlazioak eta testuaren hierarkia deskribatzen dituena. 3.21 irudian RST-zuhaitzaren deskribapenetik abiatuz laburpenak egiteko bidea agertzen da.

Tesian 3.22 irudian ikusten diren 9 irizpideak erabiltzen dira laburpenak ebaluatzeko. GrAL honetako eskuzko ebaluazioa egiteko irizpide edota esanahi berdinak erabiltzea ez da egokia. Hori kontuan hartuz 8 irizpide aukeratu dira.

Euskararen kalitatea neurtzeko 5 irizpide erabiltzea erabaki da, HABEko azterketak ebaluatzeko erabiltzen direnak. Zehazki, egokitasuna, koherentzia, kohesioa, aberastasuna eta zuzentasuna dira. Irizpide hauekin, testuen zuzentasuna eta euskararen egokitasuna kontuan hartzen dira. Sortutako laburpen testuek kalitate minimo bat izan behar dute euskararen aldetik, bestela, nahiz eta laburpen irizpideak ongi bete, ez da testu ulergarria edota egokia izango.

Irudia 3.22 Unai tesian erabilitako irizpideak. [1]

Edukia (2,5 puntu)	Ideien hierarkia (2 puntu)
	Objektibitatea (0,5 puntu)
Moldea (4,75 puntu)	Erregistroa (0,25 puntu)
	Parafraasi-kantitatea (puntu 1)
	Parafraasi-kalitatea (puntu 1)
	Kohesioa (puntu 1)
	Koherentzia (1,5 puntu)
Zuzentasuna (2 puntu)	Ortografia + puntuazioa+ sintaxia (2 puntu)
Luzera (0,75 puntu)	Luzera (0,75 puntu)

3 irizpide gehitu dira HABEko azterketetan erabiltzen direnez gain. Hauen helburua laburpenen kalitatea neurtzea da bereziki. Zehazki, hierarkia, objektibitatea eta parafraasi kantitatea dira. Irizpide hauekin abstrakzio-laburpen kalitatea neurtu nahi da. Hizkuntza aldetik egokia izanda ere, laburpen egokitzat hartzeko, irizpide horiek bete behar ditu.

Jarraian, irizpide bakoitzaren definizioa eta nota bakoitzaren esanahia.

Egokitasuna: Egokitzea dokumentuaren edukiari (ez azaldu, ez kritikatu), erabiltzaileari (ideiak fidelki islatzea, baina hizkuntza ulergarria mantenduz), sistemari (nolabaiteko homogeneotasunari eustea, araudia betetzea).

Koherentzia: Diskurtso baten zatien arteko garrantzi-maila mantentzea. Testuaren ka-teatze semantikoa errespetatzea.

Kohesioa: Esaldi, paragrafo eta testu barneko loturen (testu-antolatzaileak, kohesio-formak eta puntuazio-markak) erabilera egokia.

Aberastasuna: Hizkuntza-baliabide ugari erabiltzen dira. Testu luze eta konplexuak onduz lexikoaren eta egituren aberastasuna erabiliz.

Zuzentasuna: Dokumentuaren eduki nagusia nabarmentzea; funtsezko informazioa eta informazio anekdotikoa bereiztea.

Hierarkia: Testuko unitate zentrala (UZ), ideia nagusiak, mantentzen dira laburpenean.

Objektibitatea: Laburpenean ez da idazlearen iritzia edo testuko gairik ez den informazioa ematen. Objektibotasuna mantentzen da.

Parafraasi kantitatea: Testuko ideiak berridatzita agertzen dira laburpenean, esaldiak kopiatu beharrean.

Irizpide bakoitzaren notak 3 balio har ditzake, "Gutxi(G)", "Nahiko(N)" eta "Ongi(O)". Nota bakoitzak esanahi ezberdin bat izango du zer ebaluatzen ari garen kontuan hartuta. Irizpideen notekin laburpenari zenbakizko balio bat emateko, irizpide bakoitzeko notei

3. AURREKARIAK

zenbakizko balioak esleitu zaizkie. GrAL honetarako ez da egokia irizpide guztiek pisu berdinak izatea. Giza ebaluatzailearen iritziz ebaluazio egokiagoa egiteko garrantzitsua-
goak dira adibidez hierarkia eta parafraasi kantitatea egokitasuna baino. Irizpideko noten
esanahiak 3.3 taulan ikus daitezke eta zenbakizko balioak 3.4 taulan. Zenbakizko balioak
gehituz azken nota kalkulatzen da.

Irizpidea	Gutxi	Nahiko	Ongi
Hierarkia	Ez dago UZrik	Testuko UZ bat behintzat badu	Testuko Unitate zentral (UZ) guztiak ditu
Objektibitatea	Idazlearen iritzia ematen duen edo testuko gaiari buruzkoa ez den informazioa 3 aldiz edo gehiagotan modu leunean ageri bada, edo behin gogor egin badu.	Idazlearen iritzia ematen duen edo testuko gaiari buruzkoa ez den informazioa birritan dago, eta biak leuna dira.	Asko jota behin baino ez du idazlearen iritzia ematen duen edo testuko gaiari buruzkoa ez den informazioa, eta berau leuna da.
Parafraasi kantitatea	Gutxiegi berridatzi du; jatorrizko testuaren oso oso antzekoa.	Gehiegi berridatzi du. Sarri, berridazketa oso behartuta egin du; zertan berridatzi ez diren pasarteak parafraseatuz.	Egokia; berridazketa orekatua da, jatorrizko testuaren ezaugarriek eskatzen duten bestekoa.

3.3. Automatikoki sortutako testuen ebaluazioa

Egokitasuna	Nekez asmatuko du gaiak, testuinguruak eta abarrak eskatzen duen erregistroarekin.	Gaia kultura orokorrekoa izanik, hots, haren munduezagutzatik eta interesetik gertuko bada, ez du arazorik adierazi nahi duena egoki adierazteko, nahiz eta ñabarduretan edo lexikozehaztasunean garrantzi txikiko okerrak egin. Gai orokorretan ez du aparteko arazorik testuinguruari dagokion erregistro mota baliatzeko.	Azalpen eta argudioetan, gaia kultura orokorrekoa izanik ere, maila jasoko hizkera erabiltzen ahaleginduko da, eta informazioa egoki adieraziko du, nori eta zertarako zuzentzen zaion eta gaiaren nolakoa kontuan hartuz. Lexikoa ere egoki eta zehatz erabiliko du, hitzen balio semantikoa oso ongi menperatzen baitu. Kasuan kasuko erregistroa ongi aukeratzen du.
Koherentzia	Diskurtsoaren egitura eta pasarteentolaketak ez da egokia. Ez dira garbi bereizten ideia nagusiak eta osagarriak, eta lehen irakurraldian ez da ongi ulertzen testua.	Oro har, egokia da diskurtsoaren egitura eta pasarteentolaketa, tartekahutsegiteren bat egiten badu ere. Ideia garrantzitsuak, osagarriak eta xehetasunak bereizten ditu, informaziojauziak.	Baliabide linguistikozin testualak zuzen erabiltzen dituzenez, testua erabat argia, ongi egituratua eta jarraipen onekoa da. Ideia nagusiak eta xehetasunak behar bezala bereizten ditu eta gaia egoki garatzen du, testu koherentea sortuz.

3. AURREKARIAK

Kohesioa	Nekez baliatzen ditu ondo esaldi, paragrafo eta testu barneko loturak. Testu-antolatzaileak, kohesio-formak eta puntuazio-markak ere ez ditu egoki erabiltzen. Puntuazioaren erabilera desegokiak zaildu egiten du ulergarritasuna.	Gehienetan, ondo baliatzen ditu esaldi, paragrafo eta testu barneko loturak: testu-antolatzaileak, kohesio-formak eta puntuazio-markak. Puntuazioa, oro har, zuzena da, tarteka akatsen bat edo beste egiten badu ere.	Ez du eragozpenik eraginkortasunez baliatzeko esaldi, paragrafo eta testu barneko loturak (testu-antolatzaileak, kohesio-formak eta puntuazio-markak). Puntuazioa zuzen darabil.
Aberastasuna	Eragozpena du mailari dagozkion testuak idazteko. Darabiltzan egiturek, ideiek, lexikoaren zenbateko eta nolakoak, besteren artean, hizkuntzaren ezagutza ona ez duela adierazten dute.	Maila honi dagozkion testuak egoki osatzen ditu, kontuan hartuz baliatzen dituen ideia, egitura edota lexikoaren zenbatekoak eta nolakoak. Gutxieneko baliabide linguistiko zein testualak erabiliz, gai da ideiak zehatz adierazteko.	Kultura orokorreko edozein gairen inguruan, hizkuntzabaliabide ugari erabiliz idazten du. Testu luze eta konplexuak on-tzen ditu, nabaria baita lexikoaren eta egituren aberastasuna.

3.3. Automatikoki sortutako testuen ebaluazioa

Zuzentasuna	Maila morfo-sintaktikoan, lexikoan eta ortografian egiten dituen akatsek agerian uzten dute ez dituela hainbat egitura eta forma menderatzen. Akats sistematiko horiek direla eta, ulergaitza gertatzen da idatzia.	Maila honi dagozkion testuak idaztean ez du alderdi morfo-sintaktikoan aparteko arazorik, nahiz eta zenbaitetan adierazi beharrekoaren konplexutasunagatik sintaxiaren arloko okerren bat egiten duen. Dena den, ez du akats sistematikorik egiten. Lexikoa zuzen erabiltzen du, baliatzen dituen sinonimo, birformulazio eta abarretan ikusi ahal izango denez. Puntuazioa eta ortografia zuzen erabiltzen ditu, tarteka oker tipografikoak egiten baditu ere.	Ez du inolako eragozpenik kultura orokorreko zeinahi gai zuzen idazteko. Morfosintaxiaren, lexikoaren, ortografiaren eta puntuazio-marken erabilera zuzena da, akatsik gabea
-------------	---	---	--

3.3 Taula: Irizpide bakoitzeko noten esanahiak

Irizpidea	Gutxi	Nahiko	Ongi
Hierarkia	0	1	2
Objektibitatea	0	0,25	0,5
Parafraisi kantitatea	0	0,5	1
Egokitasuna	0	0,125	0,25
Koherentzia	0	0,75	1,5
Kohesioa	0	0,5	0,5
Aberastasuna	0	0,5	1
Zuzentasuna	0	1	2

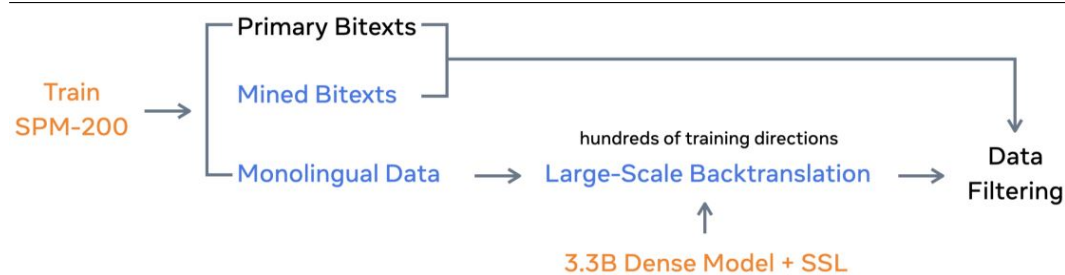
3.4 Taula: Irizpide bakoitzeko zenbakizko notak.

3.4 NLLB Itzulpen automatikoko sistema irekia

No Language Left Behind (NLLB) ereduak [26] kodetzaile-deskodematzaile motako *transformer*-ak dira eta itzulpen automatikorako entrenatuta daude. 200 hizkuntzetan entrenatuta daude, ondorioz, 200 hizkuntzen arteko itzulpenak egin ditzake hizkuntza horietako bat euskara izanik. 40.000 itzulpen bide posible onartzen ditu ereduak, token bereziak izanik jatorrizko eta helburuko hizkuntzak adierazteko.

Eredu hauek sortzeko Flores-200 datu multzoa sortu dute, 200 hizkuntza dituen corpusa da eta hainbat iturritako testuak ditu. Ereduak entrenatzeko 3 iturritako datu multzoak erabili dituzte. Lehena eskuz itzultitako testuak profil oso zehatzeko itzultzaile profesionalak erabili dira eta kalitate maila altuko testuak bakarrik erabili dituzte. Hizkuntza batetik ingeleserako itzulpenak bakarrik dira. Bigarren, *Mined Bitext* izeneko teknikak erabilia sortutako testu-laburpen pareak. LASER (Artetxe eta Schwenk, 2019) izenez ezagutzen den teknika erabili dute, baliabide gutxiko hizkuntzen arteko esaldiak elkar lotzeko [27]. Azkenik, elebakarrak diren testuak erabili dituzte aurreentrenamendu gisa, 192 hizkuntzetako testuak dira hauek. 3.23 irudian corpusa sortzeko erabili dituzten datuen iturriak ikusten dira.

Irudia 3.23 NLLB ereduak entrenatzeko corpusa sortzeko erabilitako datuen iturriak .



Ereduak ebaluatzeko automatikoki eta giza ebaluazio bitartez egin dute. Automatikoki egiteko ohikoak diren metrikak erabili dira, CHRF++ (Popović et al. 2015) esaterako [28]. Ebaluazio automatikoarekin X hizkuntzatik ingelesera itzultzeak emaitza hobeak ematen dituela ikusi dute, ingelesetik X hizkuntzera itzultzeak baino. *Google Translator*-ekin egindako alderaketetan, emaitza hobeak lortu dira, gehien bat baliabide gutxiko hizkuntzetan. Oso kontutan hartzekoak dira emaitzak, 40.000 itzulpen posibletatik ia 38.000 modeloak inoiz ikusi gabekoak baitira. Euskaratik ingeleserako itzulpenean 57,7ko CHRF++ balioa lortu da eta ingelesetik euskarara 48,5ekoa.

Ingurune Esperimentala

Kapitulu honetan, esperimentuekin zerikusia duten gaiak azaltzen dira.

4.1 Datu Multzoak

GrAL honetarako, datu multzo ezberdinak erabili dira. Helburua laburpen sistema bat egitea denez, laburpen abstraktibo atazarako eskuragarri dauden datuak hartu dira. MT0 modeloek laburpenak egiten apur bat ikasi dute, zehazki: CNN Daily Mail, Gigaword, MultiNews, SamSum, Wiki-Lingua, XLSum eta XSum datu multzoak. Gainera, *prompt* ezberdinak sortu dituzte testu mota ezberdinen laburpenak egiteko. Testu hauek, ingelesezkoak dira eta gainontzeko hizkuntzetan *language transfer* delakoaren bitartez ikasi dute laburtzen [3].

Lan honetan sortu nahi den modeloa euskarara zuzenduta dagoenez, euskarazko datuak behar dira. Kalitatezko euskarazko laburpenerako datu multzorik eskuragarri ez dagoenez, irtenbide gisa, PEGASUS ereduak entrenatzeko erabili, baina, MT0ak ikusi ez dituen testuak euskaratu eta horiek entrenamenduan erabiltzea da. 3,5 milioi testu-laburpen pare dira baldintza hori betetzen dutenak, euskaratzeko denbora eta baliabide faltagatik 700 mila inguru itzuli dira. Hona hemen, euskaratu diren datu multzoak:

arXiv, PubMed Artikulu zientifiko luzeak dira, hizkuntza tekniko eta jaso erabiltzen da testu hauetan. Honek arazoak ekar ditzake teknizismoak prozesatzeko garaian, MT0rentzat inoiz ikusi gabeko kontzeptuak izan daitezkeelako. Laburpen gisa, artikuluen *abstract* atala erabiltzen da. Guztira, 215 mila testu pare daude Arxiv-en eta 113 mila PubMed-en. Hona hemen Arxiv-eko testu eta laburpen pare bat.

Artikulua: additive models @xcite provide an important family of models for semiparametric regression or classification . some reasons for the success of additive models are their increased flexibility when compared to linear or generalized linear models and their increased interpretability when compared to fully nonparametric models . it is well - known that good estimators in additive models are in general less prone to the curse of high dimensionality than good estimators in fully nonparametric models ...

Abstract-a: additive models play an important role in semiparametric statistics . this paper gives learning rates for regularized kernel based methods for additive models . these learning rates compare favourably in particular in high dimensions to recent results on optimal learning rates for purely nonparametric regularized kernel based quantile regression using the gaussian radial basis function kernel , provided the assumption of an additive model is valid...

WikiHow Wikihow orrialdetik ateratako gomendioak biltzen dituen testuak dira. Hizkuntza arrunta eta edozeinek ulertzekoa erabiltzen da, orokorrean teknizismo eta kontzeptu teknikorik gabekoa. Testuak sortzeko, gomendio zerrendak testuetan elkartzen dira eta laburpenetarako, gomendioen laburpenak. Guztira 200 mila testu pare inguru daude. Jarraian gomendio testu eta laburpenen adibide bana.

Gomendio testua: If you're a photographer, keep all the necessary lens, cords, and batteries in the same quadrant of your home or studio. Paints should be kept with brushes, cleaner, and canvas, print supplies should be by the ink, etc. Make broader groups and areas for your supplies to make finding them easier, limiting your search to a much smaller area. Some ideas include: Essential supplies area – the things you use every day. Inspiration and reference area. Dedicated work area .Infrequent or secondary supplies area, tucked out of the way. This doesn't mean cleaning the entire studio, it just means keeping the area immediately around the desk, easel, pottery wheel, etc. clean each night..

Laburpena: Keep related supplies in the same area. Make an effort to clean a dedicated workspace after every session. Place loose supplies in large, clearly visible containers. Use clotheslines and clips to hang sketches, photos, and reference material...

Reddit TIFU Reddit-eko TIFU sub-reddit-eko argitalpenak biltzen dituen datu multzoa. Erabiltzaileek, beraien bizipenak kontatzen dituzte. Hizkuntza informala eta ulerterraza erabiltzen da. Hala ere, gainontzeko sub-Redditek baina testu kalitate hobea du. Bi motatako argitalpenak daude, TIFU-long eta TIFU-short izenez ezagutzen dira. TIFU-long-ak istorio luzeagoak dira eta laburpenak, pare bat esaldi dituzten ondorio edo ikaskizunak dira. TIFU-short berriz, istorio motzagoak dira eta laburpen gisa, izenburuak erabiltzen dira, esaldi bakarrekoak. Guztira 42 mila testu pare daude TIFU-long atalean eta 78 mila inguru TIFU-short-ean. Jarraian Short eta Long publikazio eta laburpen adibide pare bat.

Short publikazioa: I was on skype on my tablet as i went to the toilet iming a friend. i don't multitask very well, so i forgot one of the most important things to do before pooping. i think the best part was when i realised and told my mate who just freaked out because i was talking to him on the john!

Short laburpena: forgetting to pull my underwear down before i pooped.

Long publikazioa: this actually happened a couple of years ago. i grew up in germany where i went to a german secondary school that went from 5th to 13th grade (we still had 13 grades then, they have since changed that). my school was named after anne frank and we had a club that i was very active in from 9th grade on, which was dedicated to teaching incoming 5th graders about anne franks life, discrimination, anti-semitism, hitler, the third reich and that whole spiel. basically a

day where the students' classes are cancelled and instead we give them an interactive history and social studies class with lots of activities and games...

Long laburpena: confuse a 5th grade girl for a boy in front of half of her class. kids are mean. sorry sandra.

AESLC Enron Corporation izeneko enpresako langileen arteko korreoak dituen datu multzoa da. Hizkuntza estandarra erabiltzen da eta gehienak teknikariei zuzendutakoak eta informatikarekin zerikusia dutenak dira. Laburpen gisa mezuen "gaia" atala erabiltzen da. Guztira, 18 mila testu pare daude. Jarraian mezu eta "gaia"atalaren adibide bat.

Mezua: Your Internet Banking accounts are now setup again for accessing. The login id is still your main acct. # with the password being reset to the last six (6) digits of your ssn# (this is the ssn# or tin# of the primary account holder). You will then be directed to a screen that asks for information such as: Name Address City, State Zip+4 Phone # E-Mail Address After entering this information and clicking submit, you will then be instructed to change your password to anything that you desire. Always remember that your password is case sensitive...

Gaia: Internet Banking

BillSum AEBko kongresuko 1993-2018 urte tarteko lege proiektuak biltzen dituen datu multzoa. Hizkuntza formala eta teknikoa erabiltzen da. Laburpenak eskuz eginak dira, beraz, kalitatezkoak dira. Guztira 23 mila testu pare daude. Jarraian testu eta laburpen adibideak.

Testua: SECTION 1. LIABILITY OF BUSINESS ENTITIES PROVIDING USE OF FACILITIES TO NONPROFIT ORGANIZATIONS. (a) Definitions.–In this section: (1) Business entity.–The term "business entity" means a firm, corporation, association, partnership, consortium, joint venture, or other form of enterprise. (2) Facility.–The term "facility" means any real property, including any building, improvement, or appurtenance. (3) Gross negligence.–The term "gross negligence" means voluntary and conscious conduct by a person with knowledge (at the time of the conduct) that the conduct is likely to be harmful to the health or well-being of another person...

Laburpena: Shields a business entity from civil liability relating to any injury or death occurring at a facility of that entity in connection with a use of such facility by a nonprofit organization if: (1) the use occurs outside the scope of business of the business entity; (2) such injury or death occurs during a period that such facility is used by such organization...

4.2 Datu multzoen euskararako itzulpena

4.2.1 Algoritmoa eta oinarritzko ideia

Itzulpenak burutzeko NLLB ereduak erabili da. Ingelesetik euskarara itzultzen duen modeloa delako eta *Easy-Translate* kodearekin bateragarria delako, eskuragarri dauden baliabideak kontuan hartuta, zehazki, [nllb-200-1.3B](#) modeloa aukeratu da. Gainera FP16 doitasuna erabili da exekuzioak azkarrago egiteko.

Gaur egungo itzulpen sistema gehienek esaldika itzultzen ikasten dute, hau da, esaldi bat jasota bere itzulpena sortzen. Ondorioz, testu luze bat jasota itzulpen kalitatea txikiagoa izango da, ez baitaude token kopuru handi bat itzultzen ohituta.[29]

Itzulpen algoritmoaren oinarria esaldiak itzultzea da. Lehenik, testu bateko esaldiak banatzen dira, “.” karakterea banatzaile gisa erabiliz. Jarraian, esaldiak itzultzen dira eta azkenik.¹ Algoritmoan ikus daiteke itzultzeko erabili den algoritmoaren sasikodea. Itzuli funtzioa egikaritzeko, *Easy-translate*¹ paketea erabili da, baina aldaketa txiki batzuk egin zaizkio.

Algorithm 1 Datu multzoen itzulpenerako algoritmoa

```
ItzuliDatuMul =  $\emptyset$ 
for  $T \in$  DatuMul do
    Ts = T esaldietan banandu
    ITs = itzuli(Ts)
    IT = ITs esaldiak elkartu
    ItzuliDatuMul  $\leftarrow$  IT
end for
```

Algoritmoa bere horretan utzita AESLC bezalako datu multzo txikietarako egokia da. Baina, Arxiv eta Pubmed-en kasuan, aurreprozesaketak denbora gehiegi behar dute. Aurreprozesaketak azkartzeko paralelizatu egin da, hainbat hari erabiliz testuak zatitzeko. Arxiv datu multzoaren kasuan, testuak moztu dira, testuko esaldi guztien azken %20 eta %16 inguru itzuli gabe utziz. Entrenamenduan erabiliko den token kopurua 1024 ingurukoa izango denez, honek ez du inongo eraginik izango eta itzulpen prozesua azkartzeko du, azalpen gehiago 4.3.1 atalean.

Aurreprozesaketa optimizatuta ere, itzulpenean dago zama handiena. Demagun datu multzo guztietako testuen batz besteko esaldi kopurua k balioa dela. Beraz, testu bat itzultzeko, itzulpen modeloak k esaldi prozesatu behar ditu batz beste. 700 mila testu itzuli ditugu gutxi gorabehera, ondorioz, $k \cdot 700.000$ aldiz burutu du propagazioa ereduak. Esaldiak sortetan prozesatuta ere, *batch* deitzen diegun sortak erabilita ere, $k \cdot 700.000 / \text{bs}$ ordeneko denbora behar da testu guztiak itzultzeko, *bs batch*-aren tamaina (*batch size*) izanik. k balio aldakorra da datu multzo bakoitzeko, 4.1 taulan ikus daitezke datu multzoka bakoitzaren esaldi kopuruak. Batz besteko haztatu bat eginenez testu mota bakoitzeko kopuruak kontuan hartuta, k -ren balioa 29,9koa da. Honek, ereduak 20930000/bs sorta prozesatu dituela esan nahi du. Denbora eta baliabide aldetik kostu handia suposatzen dute esaldi kopuru hori prozesatzeak.

Post-prozesaketen inguruan berriz, ezer gutxi esateko. Paralelizatu gabe ere oso azkarra da gainontzeko bi atalekin alderatuz.

4.2 taulan exekuzio denbora ezberdinak ikus daitezke, kodean egin diren aldaketa bakoitzarekin.

Itzulpenak burutzeko erabili diren GPUak NVIDIA A100 80GB eta NVIDIA Tesla V100 32GB izan dira. Nahiz eta GPU oso ahaltsuak izan, denbora asko behar izan da 700.000 testu-laburpen pareak itzultzeko.

¹<https://github.com/ikergarcia1996/Easy-Translate>

Dataset	Bataz besteko esaldi kopurua
AESLC	4,57
Arxiv	40,85
BillSum	18,75
Pubmed	53,52
Reddit TIFU	14,3
Wikihow	16,42

4.1 Taula: Datu multzo bakoitzeko bataz besteko esaldi kopurua.

Aurrepro. denbora	Itzulpen denbora	Batch ta-maina	Luzera max.	CPU /GPU kop.	Hari baztertzea	Azken esaldien
8min	43min	32	128	1/1		%0
8min	25min	64	64	1/1		%20
8min	7,85min	128	32	1/1		%16
8min	2,86min	128	32	1/3		%16
8min	2min	256	32	1/3		%16
2,68min	2min	256	32	4/3		%16

4.2 Taula: Arxiv-eko 100 testu hartuta itzultzeko denborak. Luzera max. balioak, esaldiko gehienez egon daitezkeen token kopurua adierazten du.

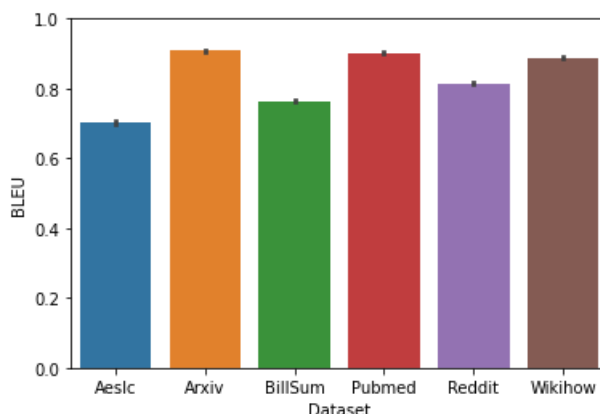
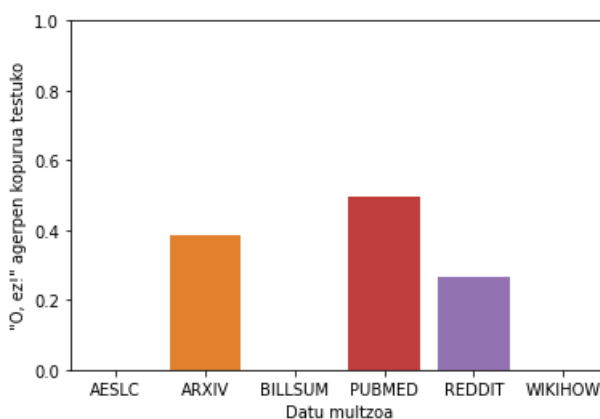
4.2.2 Itzulpenen analisia

Aukeratutako datu multzo guztiak itzultitakoan, 700 mila testu-laburpen pare inguru dituen Corpora sortu dugu.

Corpusa ingelesetik itzulia izan da, ondorioz, ez da euskarazko jatorrizko testuak bezain kalitate onekoa izango. Itzulpenen erruz egin den errorea estimatu nahian, BLEU metrika erabili da. BLEU metrika hizkuntza berdineko testuak ebaluatzeko erabiltzen da. Kasu honetan, ingelesezko eta euskarazko testuak daude, beraz, ebaluazioa zuzenean egitea ezinezkoa da. Kasu honetan, datu multzo bakoitzeko 2000 testu aukeratu, ingelesera berriz itzuli eta BLEU metrika kalkulatu da. BLEU hauen balioak 4.1 barra diagraman ikus daitezke. Errorearen balioa oso txikia da, BLEU balioak 1etik oso gertu daudelako. Azterketa honetatik bi emaitza atera dira:

- Euskararako itzulpenen kalitatea egokia da. Modeloa gai izan da euskaratu duen testua ingelesera itzultzeko, honek, testuak kalitate minimo bat eduki behar duela adierazten du, beraz, itzultitako testuek kalitate nahikoa dutela onartzen da.
- NLLB modeloa gai izan da euskaratzean egindako erroreak eta alborapenak desegiteko ingelesera itzultitakoan. Horregatik, testu batzuek metrikari 1eko balioa dute.

Orokorrean kalitatezko datuak ditugula onar daiteke, baina, AESLC datu multzoarekin arazoak daude. Datu multzo horretako BLEU balioak ikusita, gainontzeko datu multzoak baino itzulpen txarragoa izan duela ondoriozta daiteke. Itzultzerako orduan testu batzuk ez dira egoki itzuli eta akats arraroak dituzte. Honen interpretazioa ingelesetik itzultzerako

Irudia 4.1 Datu multzo bakoitzaren itzulpenen BLEU balioak.**Irudia 4.2** "O, ez!" esalditxoaren batz besteko agerpena testu bakoitzeko.

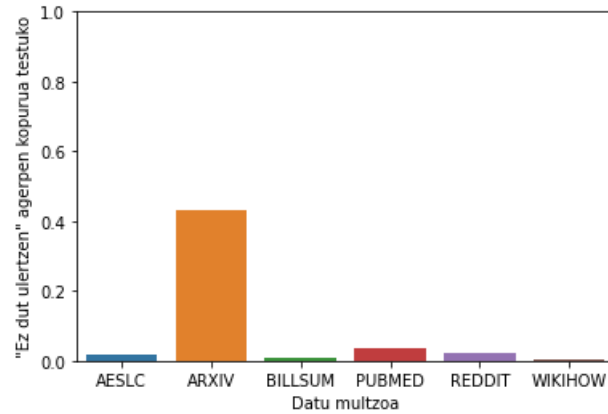
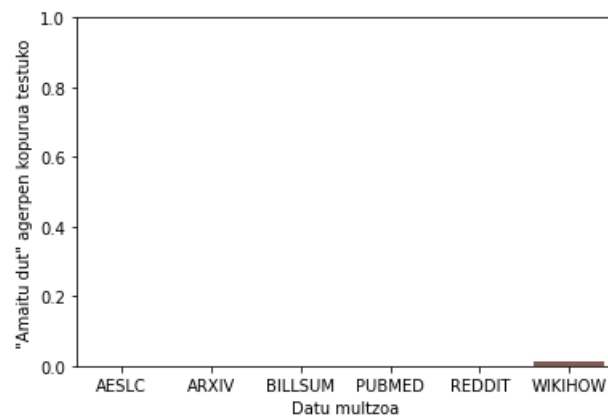
orduan NLLBk zailtasunak izan dituela da, AESLCko mezuak orokorrean teknikariei zuzendutako mezuak izanik, kontzeptu informatiko batzuk itzultzeko arazoak dituela suposatu da. Hala ere, AESLCko testuak mantentzea erabaki da, orokorrean ez baitira testu asko eta eragin negatibo oso txikia sortuko dutelako.

4.2.2.1 Itzulpenaren mugak

Itzulitako corpora aztertuz, NLLB modeloak sartu dituen akats batzuk aurkitu dira. Zehazki 3 esalditxo errepikatzen dira gehien itzulitako corpusean zehar:

Lehena, "O, ez!" da. Orokorrean ez dago arrastorik zer dela eta esalditxo hori gehitzen duen jakiteko. Esaldi amaieretan gehitzen du, baita amaierako puntuaren ondoren ere. Corpuseko testuko batz besteko agerpen kopurua datu multzoka 4.2 irudian ikus daitezke. Arxiv eta Pubmed testuetan agertzen da gehien, testu zientifikoetan alegia, baina, Redditeko istorioetan ere agertzen da.

Bigarrena, "Ez dut ulertzen" da. Esalditxo hau sarrerako testuan ezezagunak diren kontzeptu, hitz edota laburtzapenekin lotu dira. Reddit TIFUko kasuan, ohikoa da hizkuntza ez formaleko laburtzapenak agertzea eta NLLBk entrenamenduan horrelakorik ikusi ez badu, "Ez dut ulertzen" moduko testuak gehitzen ditu. 4.3 irudian testuko batz besteko

Irudia 4.3 "Ez dut ulertzen" esalditxoaren batz besteko agerpena testu bakoitzeko.**Irudia 4.4** "Amaitu dut" esalditxoaren batz besteko agerpena testu bakoitzeko.

agerpen kopurua datu multzoka ikus daiteke. Arxiv-eko testuetan agertzen da gehien, kasu honetan zentzu gehiago izango luke, kontzeptu ezezagunak edo konplexuak izango direlako.

Azkenik, "Amaitu dut". Kasu hau ulergarriena da, itzulitako esaldi batzuen amaieran agertzen da. NLLB modeloak itzultzen amaitu duela adierazteko modu bat izan liteke, hala ere, oso bitxia da esplizituki horrelakoak testua gehitzea. 4.4 irudian testuko batz besteko agerpen kopurua datu multzoka ikus daiteke. Gutxien agertzen den esalditxoa da eta Wikihow datu multzoan joera gehiago dagoela dirudi.

Bitxiak eta dibertigarriak badirudite ere, entrenamendua oker dezakete horrelako esalditxoek. "O, ez!"-en kasuan, zuzenean ezabatu daitezke, euskaraz ez baitu esanahi berezirik. Gainera, entrenatu ondorengo testu sorkuntzan eragin oso negatiboa du, ia beti "O, ez!" gehitzen baitu modeloak testu amaieran. Beste bien kasuan, agerpen guztiak ezin dira bermatu akats batengatik direnik, euskaraz esaldi horiek esanahia dutelako eta erabiltzen direlako. Hauen kasuan, entrenatu ondorengo testu sorkuntzan ez dute eragin berezirik izan, orokorrean ez ditu testuetan gehitzen. Gainera, "Ez dut ulertzen" eta "Amaitu dut" esalditxoak gutxiago agertzen dira corpusean, entrenamendua zikintzeko aukera gutxiago izanik. Azken bi esalditxoak kentzeak ez dio onurarik ekarriko entrenamenduari, eragina eta agerpen kopuruak txikiak direlako. Beraz, "O, ez!"-en agerpen kasuak ezabatu

dira, baina, beste bienak mantendu dira.

4.2.3 Corpusaren deskribapena

Itzulitako datu multzoak bilduta, euskarazko abstrakzio-laburpenerako Corpora sortu da. Corpus hau osatzeko, Arxiv, Pubmed, Reddit TIFU, Wikihow, AESLC eta BillSum datu multzoak euskarara itzuli dira. 4.3 taulan ikus daitezke Corpora osatzen duten testu-laburpen pareak.

Itzultzeko izan diren gorabeherak kontuan hartuta, datu multzo bakoitzean ez da eskura zegoen dena itzuli edota azpimultzoak bakarrik itzuli dira. Jarraian datu multzo bakoitzetik itzuli denaren deskribapen laburra:

Arxiv-en kasuan, testuen luzera dela eta testu guztiak ez dira osotasunean itzuli. Oro har, testu originalen esaldien azken % 16a ez da itzuli. Hau da, testuak 100 esaldi baditu, lehen 84 esaldiak itzuli dira. Hala ere, testu gehienek token kopurua 1024etik gorakoa da, gaur egun erabiltzen diren modeloentzat egokiak izanik. Laburpenen kasuan, ez da mozketarik egin eta osotasunean itzuli dira.

Pubmed-en kasuan, testuak bere osotasunean itzuli dira, Arxiv-ekin alderatuz motzagoak eta gutxiago zirelako. Laburpenen kasuan ere testu guztiak itzuli dira.

Reddit TIFUn, hainbat datu daude testu-laburpen pare bakoitzeko. Horietatik, laburpen atazarako garrantzitsuak direnak itzuli dira, testuak eta laburpenak hain zuzen ere. Long motatako testuetan, laburpen gisa TLDR deitzen dioten testua itzuli da eta short-etan berriz, izenburua. Gainontzeko informazioa, estekak, puntuaketak etab. ez dira itzuli, beraz, ez daude corpusean.

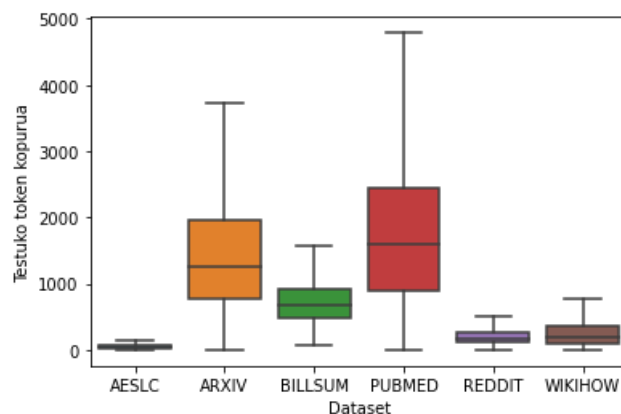
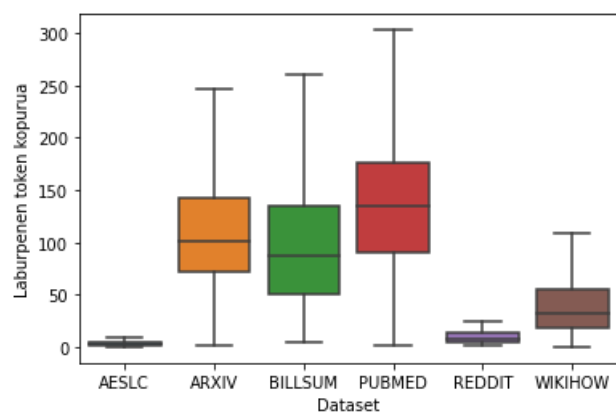
Wikihow-etik entrenamenduko zatia bakarrik itzuli da. HuggingFace-tik datuak lortzeko izandako arazo baten ondorioz izan da. 170 mila testu inguru itzuli dira. Datu multzo honetatik testuak eta "headline" edo burukoak deitzen dioten laburpen gisa erabiltzen diren testuak itzuli dira. Izenburuak ez dira itzuli.

AESLC testuekin arazoak daude orokorrean. 4.1 irudian ikus daiteke BLEU balioek joera arraroa dutela. Azalpen argirik ez dago, baina, testu batzuk hutsa direla edo itzulpenean arazoak egon direla eta ez direla ongi osatu ikusi da. Hala ere, kasu gutxi izanik, ez da garbiketarik egin. Datu multzo honen kasuan, datu guztiak itzuli dira, "email body"-ak testu bezala harturik eta "subject line"-ak laburpen gisa.

BillSum-en testuak eta laburpenak itzuli dira eta gainontzeko informazioa ez da itzuli. Zati guztietako datuak hartu dira.

Corpuseko testuen luzerak 4.5 irudian ikus daitezke. Argi ikusten da Pubmed eta Arxiv-eko testuak luzeenak direla eta Reddit TIFU edota AESLCren kasuan berriz motzagoak. Erreferentziazko laburpenen luzerarekin erlazionatuta daude testuen luzerak, 4.6 irudian ikus daiteke laburpen luzeenak Arxiv eta Pubmed-ekoak direla. Ereduak ere korrelazio hori ikasiko du eta testuak luzeagoak diren kasuetan laburpenak ere luzeak izan behar dutela ikasiko du.

Corpusaren banaketa ia kasu guztietan HuggingFace-n dagoen bezala egin da. Hala ere, bi kasu berezi daude, lehena Wikihow-rena da. Entrenamendu zatia bakarrik itzuli denez, eskuzko banaketa egin da, Train-Dev-Test zatiek %80-%10-%10-eko banaketa izanik. Bigarrena, Reddit TIFUren kasua, datu multzo honek ez du banaketarik ingelesezko bertsiotan,

Irudia 4.5 Testuetako token kopuruen diagramak.**Irudia 4.6** Laburpenen token kopuruen diagramak.

beraz, eskuz egin da berriro ere, %80-%10-%10 proportzioan. Banaketan kontuan hartu dira short eta long motatako Reddit testuak daudela banaketa egiterako orduan. Banaketa hauetatik atera dira 4.3 taulan ikusten diren kopuruak.

Dataset	Train	Dev	Test	Guztira
AESLC	14.436	1.960	1.906	18.302
Arxiv	203.037	6.436	6.440	215.913
BillSum	18.949	1.237	3.269	23.455
Pubmed	119.924	6.633	6.658	133.215
Reddit TIFU	97.503	12.188	12.188	121.879
Wikihow	172.292	21.536	21.537	215.365
Guztira	626.141	49.990	51.997	728.128

4.3 Taula: Datu multzo bakoitzeko banaketen testu kopurua.

4. INGURUNE ESPERIMENTALA

Irudia 4.7 PEGASUS-Large modeloa entrenatzeko hiperparametroak [2].

Dataset	Learning rate	Label smoothing	Num of steps	Batch size	Beam size	Beam alpha	Max input tokens	Max target tokens
XSum(C4)	1e-4	0.1	130k	256	8	0.8	512	64
XSum(HugeNews)	1e-4	0.1	80k	256	8	0.8	512	64
CNN/DailyMail(C4)	5e-5	0.1	220k	256	8	0.8	1024	128
CNN/DailyMail(HugeNews)	5e-5	0.1	170k	256	8	0.9	1024	128
NEWSROOM	4e-4	0.1	104k	256	8	0.8	512	128
Multi-News	5e-5	0.1	80k	256	8	0.9	1024	256
WikiHow	8e-4	0.1	50k	256	8	0.6	512	256
Reddit TIFU	1e-4	0.1	12k	256	8	0.6	512	128
BIGPATENT	5e-3	0.1	300k	256	8	0.7	1024	256
arXiv	8e-4	0.1	74k	256	8	0.8	1024	256
PubMed	2e-4	0.1	100k	256	8	0.8	1024	256
Gigaword	8e-4	0.1	90k	256	8	0.6	128	32
AESLC	2e-4	0.1	16k	256	8	0.6	512	32
BillSum	2e-4	0.1	100k	256	8	0.8	1024	256

4.3 Ereduak eta entrenamenduak

4.3.1 Entrenamenduen azalpen orokorra

Entrenamendurako MT0 ereduak erabili dira. Lehen esan bezala, modelo hauek kodetzailde-deskodetzailde motako *transformer*-ak dira eta euskara ere ikasi dute. Gainera, hainbat atazatan entrenatuta daude, horietatik abstrakzio-laburpenak bat izanik.

Erabilgarri dauden MT0 modeloetatik MT0-base erabili da orokorrean, honek 580M parametro ditu. Tokenizatzailea ere MT0 modeloena erabili da.

Corpuseko token kopurua ikusita, entrenamenduan *prompt* berdina erabiltzea corpuseko testu guztientzat ez dela egokia izango ondoriozta daiteke. Horregatik MT0ren artikuluan aurkezten dituzten hainbat *prompt*-etatik testu mota bakoitzarentzat egokiena dena aukeratu da. Zerrendatuta erabakitako *prompt*-ak:

Arxiv, pubmed: First, read the Basque article below. {{testua}} Now, please write a short abstract for it in Basque.

Pubmed: First, read the Basque article below. {{testua}} Now, please write a short abstract for it in Basque.

Wikihow: {{testua}} TL;DR in Basque:

Reddit TIFU: {{testua}} How would you rephrase that briefly in Basque?

AESLC: {{testua}} How would you rephrase that briefly in Basque?

Billsum: First, read the Basque article below. {{testua}} Now, please write a short abstract for it in Basque.

Hiperparametroak PEGASUSeko artikulutik hartutakoak dira. Hala ere, PEGASUSen ez bezala, gure entrenamenduetan corpus osoa erabiltzen da, PEGASUSen kasuan testu multzo bakoitza modu independente batean entrenatzen zen. 4.7 irudian PEGASUS modeloen hiperparametroak ikus daitezke.

Arxiv, Pubmed eta BillSum datu multzoen kasuan, 1024 luzerako sarrerak erabiltzen dituzte, gure kasuan MT0ak 512 gehieneko luzera erabilia aurre-entrenatu ziren. Honek arazo bat suposa dezake 1024 luzerako datuak entrenatzen hastean. PEGASUSekoek arazo berdina dute, luzera ezberdineko datuak erabilia entrenatzeko. Horren aurrean, suposizio bat egiten dute, posiziodun hitz-bektoreak egokituko direla entrenamenduan zehar sarrera luzeagoak kodetzeko. Suposizio bera eginda, gure corpuseko luzera 1024ra zehaztu da.

4.7 irudian erabili duten *batch* tamaina 256koa dela ikusten da. Gure kasuan, baliabideen mugez ondorioz ezinezkoa zaigu *batch* tamaina berdina erabiltzea. Konponbide gisa, gradiente metaketa teknika erabili dugu, horrela *batch* tamaina txikiagoa erabiltzen da (8 izan da erabiliena) eta behar adina aldiz galera balioa metatzen da 256ra iritxi arte.

Entrenamenduaren luzera *epoch*-etan neurtu da, nahiz eta pausuak izan orokorrean erabiltzen den neurria. Baliabideen mugak kontuan hartuta, ezin genituen entrenamendu luzeak egin eta 5 *epoch*-etara mugatu zen. Galera funtzio gisa, besterik ezekoa mantendu da, *Negative Log Likelihood* izanik. Ebaluazio automatikoa egiteko, ROUGE-F1 erabili da.

4.3.2 Entrenamenduko erabakiak

Oinarrizko erabakiak hartu ondoren, probako entrenamendu bat egin da eta jarraian azalduko diren arazo eta erabakiak hartu dira.

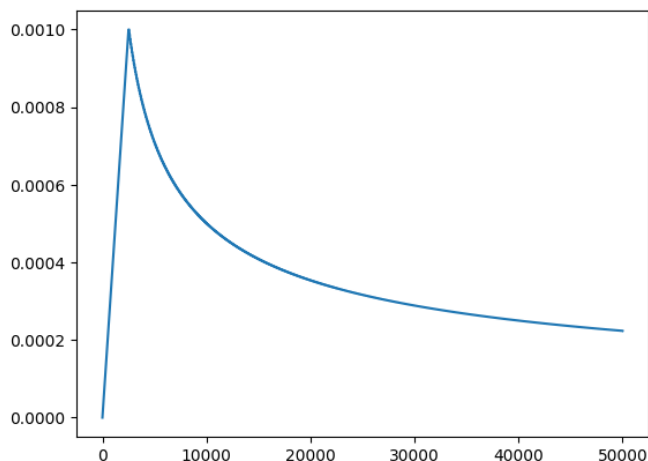
4.3.2.1 LR schedulerra

HuggingFaceko *transformer* liburutegiak hainbat ikaskuntza tasaren planifikatzaile onartzen ditu. Gainera, eskuz inplementatzeko aukera ere badago pytorch-eko LRScheduler objektu moduan. PEGASUS artikuluan *Inverse Squared Root with Warm Up* planifikatzailea erabili zutela zehazten dute, beraz, gure kasurako ere planifikatzaile bera erabili dugu.

Planifikatzaile honek arazoak sortu dizkigu entrenamenduetan. PEGASUS entrenatzean, 256 *batch* tamaina erabiltzen dutenez, orokorrean egiten dituzten pausu kopurua txikiagoa da gure entrenamenduetan baino. Adibidez, PEGASUS entrenatzean 500.000 kasu erabili balituzte, $500.000/256 = 1954$ pausu egiten dituzte eta pausu bakoitzeko, planifikatzaileak ikasketa tasa (LR) eguneratzen du. Gure kasuan, 500.000 kasu izanik eta *batch* tamaina 8 erabilia, $500.000/8 = 62500$ pausu dira. Hori ikusita, argi dago gure planifikatzailearen kasuan LRa azkarrago eguneratuko dela eta azkarrago jaitsiko dela.

Arazoa ezagututa, konpontzeko honako bi aldaketa egin dira planifikatzailean:

- Lehenik, planifikatzailearen eguneraketak atzeratu behar dira. *Batch* tamaina 8 bada, 32 aldiz atzeratu behar da planifikatzailearen eguneraketa. Atzeratzeko eta kodean aldaketa gehiegi ez egiteko, LRa behin kalkulatu eta gradiente metaketa balio aldiz erabiltzen da LR bera. Horrela, 256 kasu ikusi arte LRa ez da eguneratuko eta *batch* tamaina handiagoa simulatu dezakegu.
- Bigarren, nahiz eta LRa gradiente metaketa balio aldiz erabili, kalkulu egokia egiten dela bermatu behar da. PEGASUSen kasuan, 1.000 pausuko LR balioa, gure modeloko 32.000 pausuko LRa da. Hau da, 32 (Gradiente metaketa balioa) aldiz pausu gutxiagoa 8ko *batch* tamaina izanda. Horretarako, LRaren kalkulu formula egokitu behar da, 32 aldiz txikiagoa den pausuko LRa kalkula dezan.

Irudia 4.8 Entrenamenduetan erabilitako LR planifikatzaileak sortzen duen grafika.

Aldaketak egin ondoren, 4.8 irudian ikus daiteke LR planifikatzaileak egingo duen grafikoa. Egokiagoa da gure entrenamendurako eta *batch* tamaina handiagoak simulatzea ahalbidetzen digu oztoporik gabe.

4.3.2.2 Optimizatzailea

Transformers liburutegiko *Trainer* objektuak besterik ezean AdamW (Loshchilov et al.) optimizatzailea erabiltzen du. Optimizatzaile hau oso ezaguna eta erabilia da [30] artikuluan aurkeztu zuten eta emaitza oso onak ematen ditu.

Entrenatuko dugun modeloa kodetzaile-deskodatzaile motakoa da eta 256 *batch* tamainara gerturaten saiatuko gara. Horretarako, memoria erabilera handia egingo dugu eta ahal dugun neurrian aurreztea komeni zaigu. Hori kontuan izanik, Adafactor (Shazeer et al.) [31] bezalako optimizatzaile bat egokia da guretzako, memoria gutxiago erabiltzen baitu AdamW-ren pareko emaitzak lortuz.

Adafactor-ek hainbat parametro ezberdin jasotzen ditu, modu ezberdinetara erabil daitekeelako optimizatzailea. Hala ere, T5 edo eratorriak diren modeloak entrenatzeko, aholku batzuk daude HuggingFace-ko dokumentazioan. Zehazki, [HuggingFace-ko forotik](#) ateratako gomendioak dira eta honakoa zehazten dute:

- LR planifikatzailea erabiltzean *Warm Up*-ak egitea oso gomendagarria da.
- *Clip Threshold* balioa 1era zehaztu.
- Adafactor-ek erabil ditzakeen "Relative Updates" pausu motak desgaitu
- *scale_parameter* aukera desgaitu.
- Gainontzeko optimizatzaile eragiketarako, *gradient clipping* adibidez, ez erabiltzea gomendatzen da.

Nahiz eta Adafactor T5 moduko modeloentzako diseinatua egon [31], bi entrenamendu egingo dira bi optimizatzaileak konparatzeko. Hori entrenatutako modeloak atalean erakutsiko da.

Entrenamendua	Denbora (ordutan)
Train 1	36
Train 1P	24
Train 2Adafactor	40
Train 2Adam	43
Train 3Long	79
Train 4Large	118

4.4 Taula: Entrenamendu bakoitzak behar izan duen denbora.

4.3.3.1 Promptaren eragina aztertzekeo entrenamenduak

Bi eredu entrenatu dira ia modu berean, baina, ezberdintasun batekin, *prompt*-aren erabilera. Train 1 modeloa entrenatzerakoan ez da *prompt*-ik erabili eta Train 1P entrenatzean berriz, *prompt* bakarria erabili da testu mota guztietan 4.2.2 atalean zehazten direnetatik Arxiv, Pubmed eta BillSum-en erabilitakoak. Entrenamendu hauen helburua *prompt*-aren eragina neurtzea izan da eta 5 kapituluaren emaitzak alderatuko dira.

Entrenamenduak corpusaren azpiatal batekin eginak daude. Guztira 356.094 testu erabili dira eta 4.5 taulan datu multzo bakoitzeko erabilitako testu kopurua agertzen da.

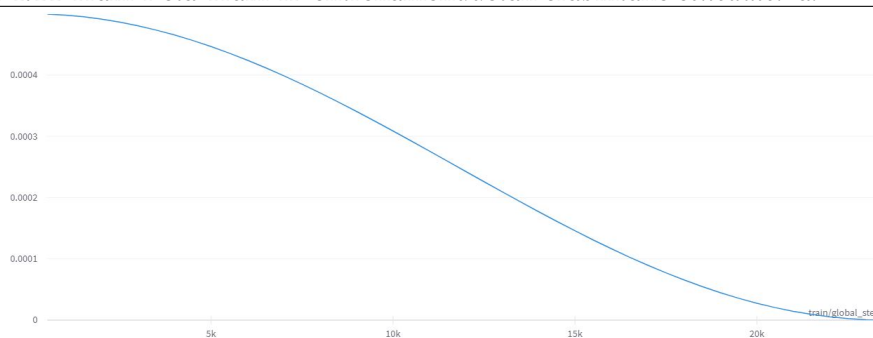
Dataset	Train	Dev	Test	Guztira
AESLC	14.436	1.960	1.906	18.302
Arxiv	203.037	6.436	6.440	215.913
Reddit TIFU	97.503	12.188	12.188	121.879
Guztira	314.976	20.584	20.534	356.094

4.5 Taula: Erabilitako testu kopurua datu multzoka.

Oinarri bezala, MT0-base eredu hartu da. Hiperparametroak berriz, 4.6 taulan ikus daitezke. *Batch* tamaina 4koa erabili da eta 16 aldiz metatu da gradiente metaketa aplikatuz. Ikasketa tasa $5e-4$ balioarekin hasi eta Kosinuaren planifikatzailea erabiliz jaisten da 4.11 irudian ikus daitekeen moduan. 5 *epoch*-etan entrenatu da, gutxi gorabehera 80 orduko exekuzioa izanik.

Hiperparametroa	Balioa
Batch tamaina	4
Gradiente metaketa balioa	16
Ikasketa tasa	$5e-4$
Epoch	5
Planifikatzailea	Kosinuarena
Optimizatzailea	AdamW

4.6 Taula: Train 1 eta Train 1Pren hiperparametroak.

Irudia 4.11 Train 1 eta Train 1P entrenamenduetan erabilitako *scheduler*-a.**4.3.3.2 Optimizatzailearen eragina aztertzen**

Entrenamendu hauetan, bi optimizatzaile ezberdin erabilia egin dira entrenamenduak. Lehena, AdamW optimizatzailearekin (Train 2Adam), aurrekoetan erabili dena eta bigarren, Adafactor optimizatzailea (Train 2Adafactor).

Erabilitako testuen inguruan, corpus guztia erabili da 700 mila testu-laburpen pare inguru. *Prompt*-ak erabili dira, 4.2.2 atalean zehaztutakoak.

Kasu hauetan ere, MT0-base eredu hartu da oinarritzat. Hiperparametroak eta gai-nontzeko erabakiak, 4.3.1 eta 4.3.2 ataletan zehaztutakoak erabili dira. Hau da, PEGASUS modeloen antzera, 256 *batch* tamaina simulatu da eta LR balioa hasieran $1e-3$ ko da eta 4.3.2.2 ataleko ikasketa tasaren planifikatzaile bitartez jaisten da. 4.7 taulan ikus daitezke argiago erabilitako hiperparametroak.

Hiperparametroa	Balioa
Batch tamaina	8
Gradiente metaketa balioa	32
Ikasketa tasa	$1e-3$
Epoch	5
Planifikatzailea	4.3.2.2 atalekoa
Optimizatzailea	AdamW/Adafactor

4.7 Taula: Train 2Adam eta Train 2Adafactor entrenamenduetako hiperparametroak.

4.3.3.3 Entrenamenduak luzatzearen eragina

Entrenamendu honetan, denbora luzeagoan entrenatzen utziz gero ereduaren duen eragina aztertu nahi da. Horretarako Train 3Long izeneko entrenamendua egin da.

Modelo hau Train 2Adafactor deituriko ereduaren antzera entrenatuta dago. Hau da, MT0-base oinarritzat hartuta eta hiperparametro, datu eta optimizatzaile berdina erabilia. Baina, 5 epoch entrenatu beharrean 10 *epoch*-etan entrenatu da. 4.7 taulan daude Train 2Adafactor entrenatzeko erabili diren hiperparametroak.

4.3.3.4 HE handiagoa erabiltzearen eragina

Entrenatutako azken modeloa da hau, Train 2AdaFactorren antzera entrenatutakoa da, [4.7](#) taulan agertzen diren hiperparametroekin. Datu berdinekin, hiperparametro berdinekin, 5 epochetan entrenatuta eta planifikatzaile berdinarekin, baina, MT0-base modeloari *finetuning*-a egin beharrean, MT0-large modeloa aukeratu da.

MT0 large modeloak 1200 milioi (1,2B amerikar) parametro ditu, ia MT0-basearen kopurua bikoiztuz. Modelo honi *finetuning* egiteko denbora dezente gehiago behar da, Train 2AdaFactor entrenatzeko 41 ordu behar izan badira, Train 4Large entrenatzeko 118 ordu. Memoria aldetik ere aldaketa bat nabari da, Train 2AdaFactorrek *batch* tamaina 8 izanik, GPUaren %62,5 inguru erabiltzen du eta Train 4Largek *batch* tamaina 4 izanik %77.

Emaitzak

Hainbat modelo entrenatu dira, eredu bakoitza ikerketa galdera bati erantzuteko entrenatu da. Kapitulu honetan, entrenamenduen emaitzak aztertuko dira, galerak, metrika bidezko ebaluazioak eta eskuzko ebaluazioak alderatuko dira.

5.1 Ereduen galera funtzioaren azterketa

5.1.1 Promptaren eragina aztertzekeko entrenamenduak

Entrenamendu hauetan *prompt*-a erabili gabe (Train 1) eta *prompt*-a erabilia (Train 1P) ereduan dakarren eragina aztertu nahi izan da.

Entrenamenduko galerak alderatzeko, entrenamenduan zeharreko balioak 5.1 irudiko grafikoan ikus daitezke. Train 1Pren galerak txikiagoak direla ikus daiteke, orokorrean 0,1en eta 0,05en arteko ezberdintasuna izanik. Garapen galerak alderatuta ere, 5.2 irudian ikusgai daudenak, ezberdintasun antzekoa ikusten da, kasu honetan 0,1-0,15eko aldea dago modelo batetik bestera. Ikasketa azken *epoch*-ean moteldu egiten da, LR balioak ere txikiagotzen joaten direlako seguruenik.

Ondorio gisa, *prompt*-ak erabiltzeak onura dakarrela entrenamenduan ikus daiteke, orokorrean galera balio txikiagoak lortu direlako agindu hauen bitartez. Hala ere, jarraian dauden ataletan zehaztuko da bi modeloen artetik laburpen egokienak zeinek sortzen dituen. Gainontzeko entrenamenduetarako, *prompt*-ak erabiltzea erabaki da galeren kurbak hobeak direlako.

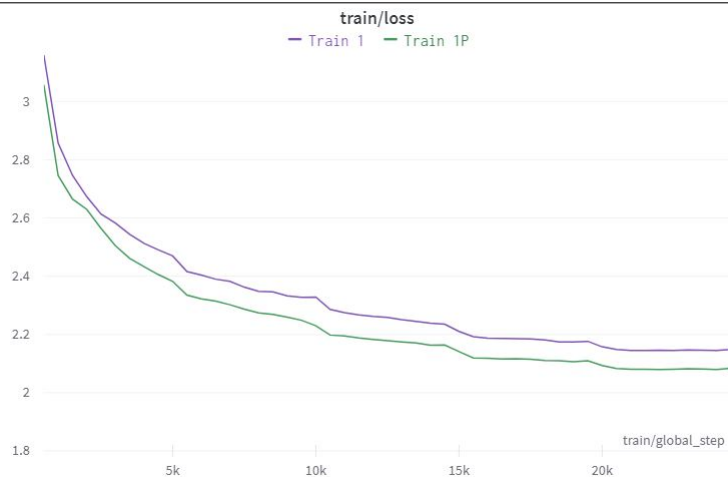
5.1.2 Optimizatzailearen eragina aztertzen

Entrenamendu hauetan bi optimizatzaile ezberdin erabili dira ereduan dakarren eragina aztertzeke. Lehena AdamW izenez ezagutzen dena (Train 2Adam) eta bigarrena Adafactor (Train 2Adafactor).

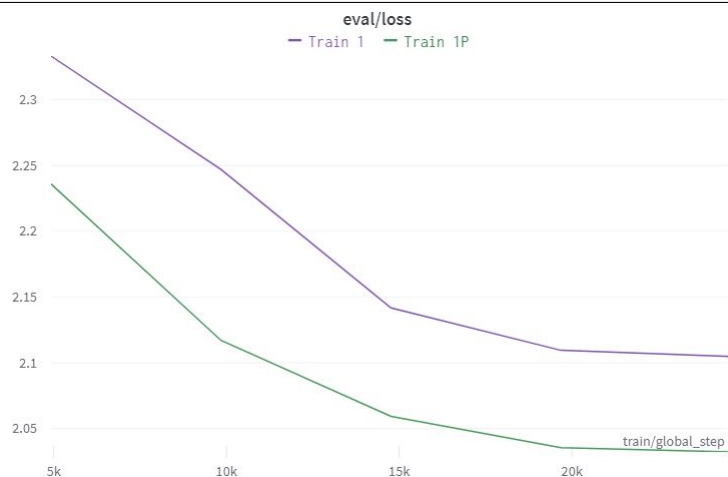
5.3 irudian entrenamenduetan izandako galerak alderatzen dira. Ikus daiteke Adafactor erabilia galeraren balioa apur bat txikiagoa dela, azken *epoch*-ean Adafactor-ek 0,06 puntu gutxiago izanik. Hala ere, biek balio antzekoak dituzte eta orokorrean ez da ezberdintasun handirik antzematen.

5. EMAITZAK

Irudia 5.1 Train 1 eta Train 1P entrenamenduetako galera balioak.

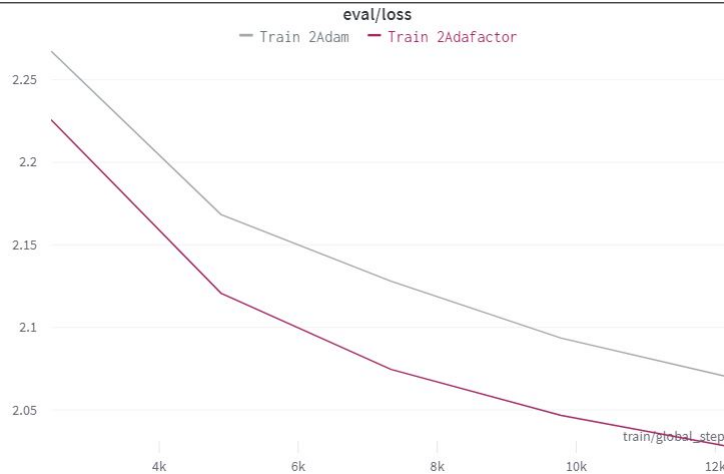


Irudia 5.2 Train 1 eta Train 1P garapeneko galera balioak.



Irudia 5.3 Train 2Adam eta Train 2Adafactor entrenamenduko galera balioak.

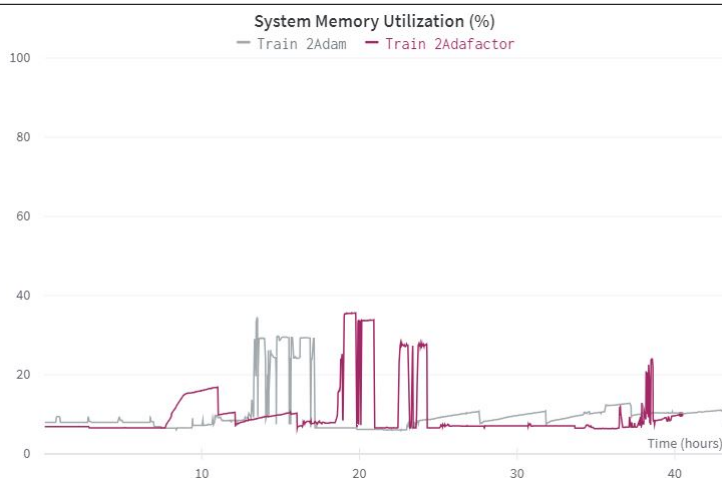
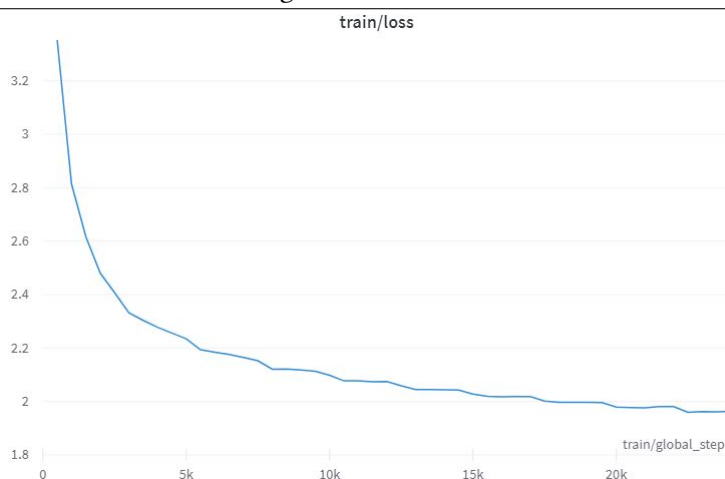


Irudia 5.4 Train 2Adam eta Train 2Adafactor garapeneko galera balioak.**Irudia 5.5** Train 2Adafactor eta Train 2Adam entrenamenduen GPUaren memoria erabilera.

AdamW eta Adafactor entrenamenduen garapeneko galeren artean ere entrenamenduko joera berdintsua ikusten da 5.4. Bietan balio antzekoak dituzte, baina, 0,05 puntu gutxiago izanik Adafactor-ek gutxi gorabehera. Berrito ere, ez da ezberdintasun handirik antzematen.

Adafactor eta AdamW-ren arteko ezberdintasunik nabariena memoriaren erabilera da. Adafactor-ek memoria gutxiago erabiltzen duen algoritmoa inplementatzen du eta hori entrenamenduan nabaritu da. 5.5 irudian GPU memoria erabileren balioak daude eta Adafactor-ek % 5 inguru aurrezten du memoria erabileran. Gutxi badirudi ere, kontuan hartu behar da errendimendu antzekoa (edo apur bat hobea) lortzen dela memoria aurreztuz eta hori abantaila handi bat dela entrenamenduetan. RAM memoriaren erabileran jokaera oso antzekoa erakusten dute, badirudi, Adafactor erabiltzeak onura txiki bat duela orokorrean, baina, ez da GPUan bezalako aldaketarik nabari. 5.6 irudian alderatzen dira bi optimizatzaileen RAM memoria erabilerak.

Memoriaren erabilera egokiagoa eta emaitza antzekoak edo apur bat hobea lortzen direnez Adafactor erabilita, optimizatzaile hau erabili da gainontzeko entrenamenduetan. Bi ereduen laburtzeko gaitasuna hurrengo ataletan alderatuko da, ataza honetarako

Irudia 5.6 Train 2Adafactor eta Train 2Adam entrenamenduen RAM memoria erabilera.**Irudia 5.7** Train 3 entrenamenduko galera balioak.

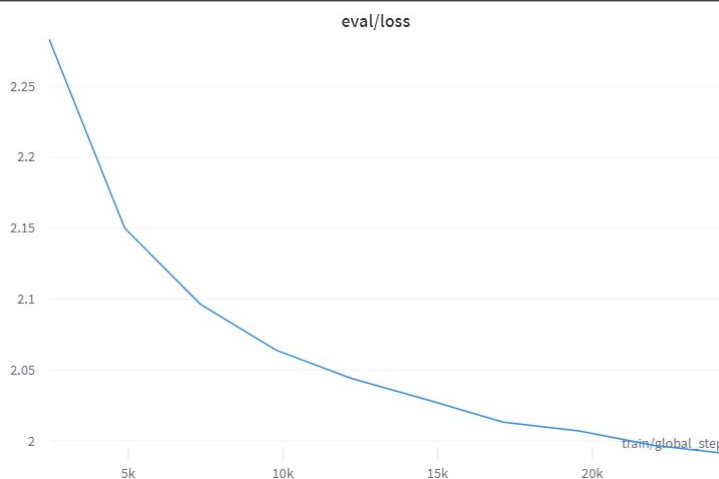
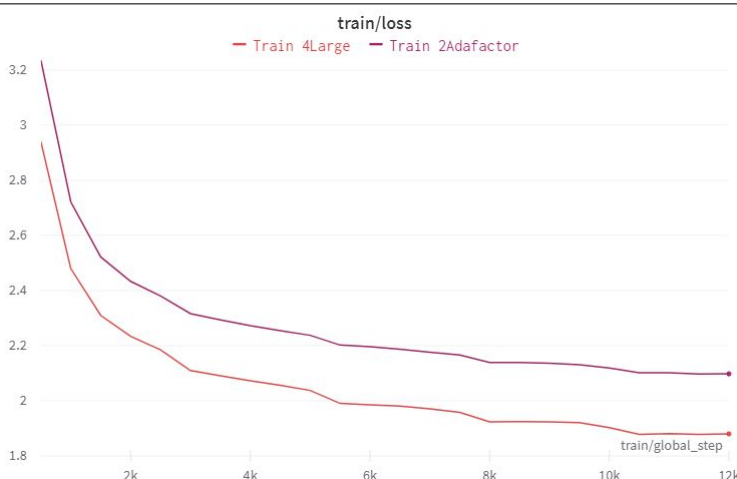
optimizataile egokiena aukeratuz.

5.1.3 Entrenamenduak luzatzearen eragina

Gainontzeko entrenamenduen garapeneko zatiaren galerak ikusita, entrenamendu zatiaren gaintze ez dela gertatzen ikus daiteke, bi galeren grafikek beherantz joateko joera mantentzen dutelako. Beraz, entrenamendu luzeagoa eginez gehiago ikasten jarraituko lukeela suposatzen da eta ondorioz, modelo hau sortu da (Train 3Long).

Entrenamenduko galera balioak aztertuz, txikiagoak lortu dira 5 *epoch*-ekin entrenatutako modeloekin alderatuz. 5.7 irudian grafika ikus daiteke, baina, azken *epoch*-n ikasketa prozesua motelagoa da, hau LR planifikatzailearen eragina izan daiteke, azkar xamar jaisten baitu LRa. Modu berean, garapeneko galerak aztertuz, 5.8 irudian ikus daiteke, entrenamenduko joera antzekoa duela, balio txikiagoak beste modeloekin alderatuz, baina, azken *epoch*-etan ikasketa motelagoa izan da.

Galera balio txikiagoak lortu dira gehiago entrenatuta, baina, ereduaren laburpen gaitasunean eraginik izan duen aztertzeke, 5.2 atalean metrika automatikoak erabiliko dira.

Irudia 5.8 Train 3 entrenamenduko garapeneko zatiko galera balioak.**Irudia 5.9** Train 4 entrenamenduko galera balioak Train 2Adafactorrekin alderatuz.

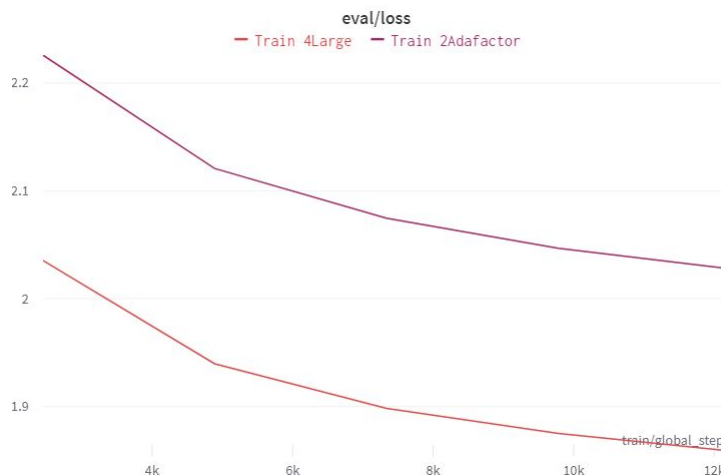
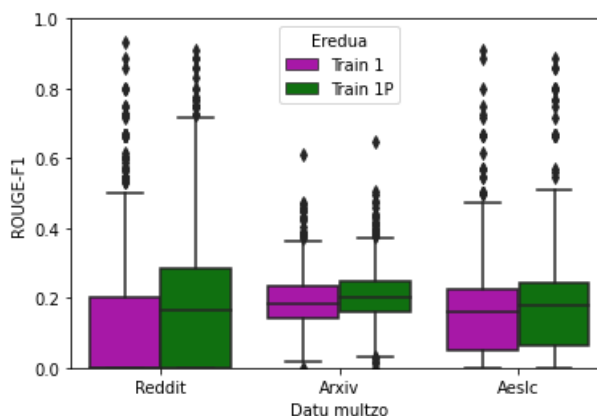
5.1.4 HE handiagoa erabiltzearen eragina

Entrenamenduko galera balioak Train 2Adafactor-ekin alderatuz, hobeak dira, 5.9 irudian ikus daiteke galera txikiagoak direla. Azken *epoch*-aren amaieran 0,22ko alde dago Train 2Adafactor eta Train 4Large modeloen artean. Garapeneko galerak alderatuta ere, alde handia nabari da. 5.10 irudiko grafikan ikus daitezke Train 4Largeren garapeneko galera balioak eta azken *epoch*-ean 0,17 puntuko ezberdintasuna dago.

Train 4Large eta Train 3Long entrenamenduen artean, galera baxuenak modelo handienak lortzen dituela ikusten da. Modelo handiagoa erabilia emaitza hobeak lortzen direla ematen du, baina, 5.2 atalean ikusiko da zein den egokiena.

5.2 Metrika automatikoen bidezko ebaluazioa

Modeloen entrenamenduak 4 ataletan banandu daitezke 4.3.3 atalean bereizi diren bezala. Lehenak, *prompt*-a erabili edo erabili gabe duen eragina aztertzea du helburu. Bigarrenak,

Irudia 5.10 Train 4 entrenamenduko garapen zatiko galerak Train 2Adafactor-ekin alderatuz.**Irudia 5.11** Train 1P eta Train 1 ereduak lortutako ROUGE-F1 balioak.

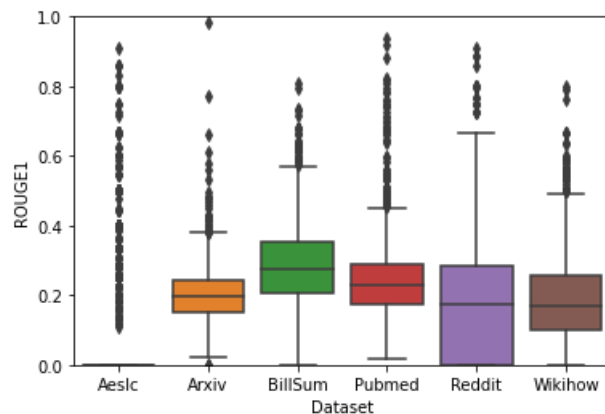
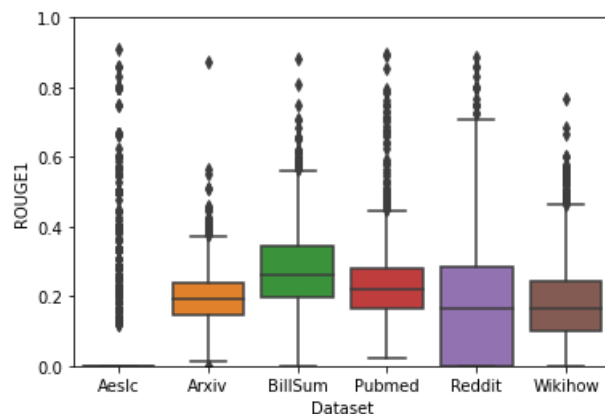
Adafactor eta AdamW optimizatzailetik egokiena aukeratzea. Azken biak, modelorik onena entrenatzeko helburuarekin egin dira, aurreko ataletan ikasitakoa erabiliz.

Konparaketak burutzeko, ROUGE-F1 metrika erabili da, hau da, unigramak konparatzen dituen eta ausaz datu multzo bakoitzeko Testeko 2000 testu aukeratu dira. Test guztia erabil daiteke, baina, 52 mila testuren laburpenak sortzea lan oso neketsua da eta hainbat ordu beharko lirateke. Horren aurrean, ausaz azpimultzo bat erabiltzea hautatu da.

5.2.1 Promptaren eragina aztertzeko entrenamenduak

Train 1 eta Train 1P modeloen arteko ezberdintasun bakarra, *prompt*-aren erabilera dago. MT0 ereduak laburpenetarako *prompt*-ak erabiliz ikasi dute laburtzen. *Prompt* ezberdinak erabilia, mota ezberdinetako testuak itzultzen ikasi dute, artikulua, testu informala, testu juridikoak etab. Bi ereduaren ROUGE balioak 5.11 grafikan ikusi daitezke. Batzuetan 0,1827ko balioa lortu du Train 1P ereduak eta 0,1467koa Train 1ek.

Hala ere, *prompt*-ik erabili gabeko entrenamenduarekin alderatuta, *prompt*-ak erabiltzeak ez du onura handirik ROUGE metrikaren arabera Arxiv eta AESLC testuetan.

Irudia 5.12 Train 2Adafactor modeloak lortutako ROUGE1 balioak.**Irudia 5.13** Train 2AdamW modeloak lortutako ROUGE1 balioak.

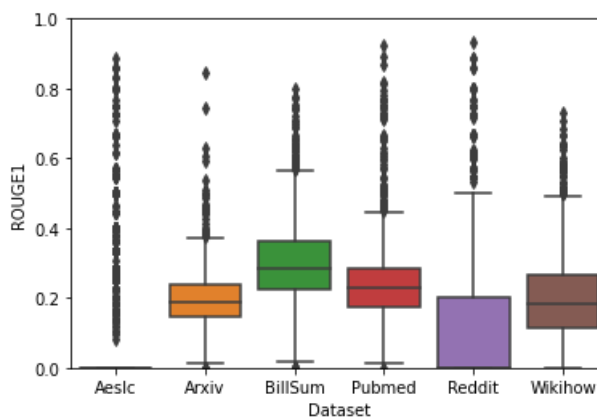
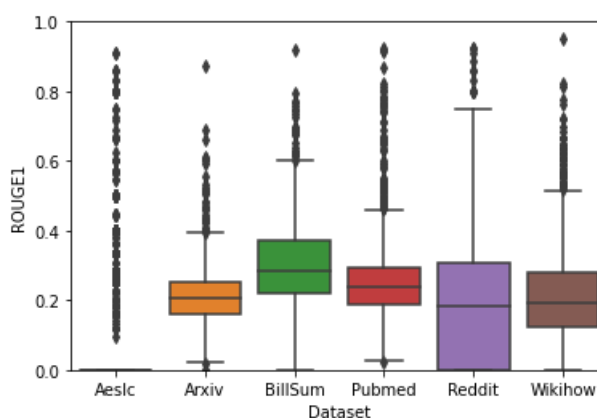
Reddit-en kasuan berriz, *prompt*-ak erabilia egokiago ikasi duela dirudi.

ROUGE metrikari jaramon eginda, bi ereduak antzeko joera dute testuen laburpenak egiterakoan. Alde handiena Reddit TIFUko testuetan dago, *prompt*-ak erabiltzeak onura dakar testu hauen laburpenean. Gainera, testuak irakurrita, egokiagoak dirudite *prompt*-ak erabiliz sortzen diren laburpenak. Beraz, *prompt*-ak erabiltzeak testu mota guztietan onura dakarrela eta ereduak testu mota ezberdinen arteko aldea errazago ikasiko duela suposatuko da.

5.2.2 Optimizatzailearen eragina aztertzen

AdamW oso erabilia den optimizatzailea da, baina, Adafactor-ek memoria gutxiago erabiltzen du eta hau onura garrantzitsua da modelo handiak entrenatzerakoan.

Gure kasuan, 4.3.3.2 atalean erakusten den bezala, memoria gutxiago erabili da Adafactor optimizatzailearekin eta galera balioak apur bat txikiagoak dira. 5.12 irudian ikus daitezke Adafactor erabilia lortu diren emaitzak eta 5.13 irudian AdamW erabilia. Apur bat hobeak lortzen dira orokorrean Adafactor erabilia eta batz bestea, 0,00485ko aldea dago ROUGE balioan, Train 2Adafactor-ek 0,2 izanik eta Train 2Adamek 0,19535. Honek, Adafactor-en erabilera hobe dela erakusten digu, bai memoria aldetik, bai lortutako emaitzengatik.

Irudia 5.14 Train 3Long modeloak lortutako ROUGE-F1 balioak.**Irudia 5.15** Train 4Large modeloak lortutako ROUGE-F1 balioak.

5.2.3 Entrenamenduak luzatzearen eragina

Prompt-ak erabilia eta Adafactor optimizatzailearekin denbora gehiago entrenatuta Train 3Long izeneko modelo garatu da. Emaitza antzekoak lortu dira ROUGE metrikaren arabera, 5.14 irudian ikus daitezke datu multzoka balioak eta orokorrean Train 2Adafactor-en antzekoak. Arxiv, Pubmed eta Wikihow-en kasuan apur bat hobetago egin du, baina besteetan okerrago. Batz bestea 0,191ko balioa lortu du, oraingoz modelorik okerrera izanik ROUGE-F1 metrika kontuan hartuta.

5.2.4 HE handiagoa erabiltzearen eragina

Azkenik, MT0-large oinarri hartuta entrenatutako modelo. Train 3Long-ekin alderatuta, emaitza hobek lortu ditu datu multzo guztietan, 5.15 grafikan ikus daitekeen bezala eta batz bestea 0,2136ko ROUGE-F1 balioa du.

Azken bi entrenamenduak konparatuz, modelo handiagok erabiltzea hobea dela ikusi da, txikiagoa den bat gehiago entrenatuta baino. Hori [32] artikuluan egindako ikerketa baten bitartez ere baieztatzen dute. Hala ere, baliabide gehiago behar dira modelo handiagok entrenatzeko, Train 4Large modelo entrenatzeko 118 ordu behar izan dira eta Train 3Long entrenatzeko berriz, 79 ordu.

5.2.5 Modeloen alderaketak PEGASUSekin

5.1 taulan ikus daitezke modelo bakoitzak datu multzo bakoitzeko lortutako ROUGE-F1 balioak. Aurreko atalean ikusi den moduan, Train 4Large modelo da metrika honen arabera modelorik onena, datu multzo ia guztietan balio onenak lortu ditu eta. Harrigarria bada ere, AESLC datu multzoko emaitzarik onenak Train 1P modeloak lortu ditu. Zentzuzko joera da, AESLC corpuseko zati txiki bat delako, zehazki corpuseko % 3a, honek eragin negatiboa du ikasketan. Train 1Pren kasuan, testu kopuru txikiagoa izanik (340 mila gutxi gorabehera), pisu gehiago hartzen dute AESLCko testuek.

Modeloa	AESLC	Arxiv	BillSum	Pubmed	Reddit	Wikihow	Bataz beste
Train 1	0,1520	0.1901	–	–	0,1120	–	0,1467
Train 1P	0,1662	0,2064	–	–	0,1819	–	0,1827
Train 2Adam	0,0803	0,1973	0,2749	0,2341	0,1838	0,1804	0,1954
Train 2Ada-factor	0,0783	0,2017	0,2844	0,2435	0,1828	0,1876	0,2002
Train 3Long	0,0840	0,1940	0,2984	0,2401	0,1186	0,1972	0,1911
Train 4Large	0,0879	0,2127	0,3	0,2519	0,1937	0,2104	0,2136

5.1 Taula: Modelo guztiak datu multzoka duten ROUGE balioa. Letra lodiz datu multzo bakoitzeko lortu diren balio hoberenak.

Corpus hau ingelesetik itzultako testuak bilduz sortu da eta PEGASUS modeloak entrenatzeko ere erabili ziren. Modeloak aurkezten dituzten artikuluan ere datu multzo bakoitzeko lortutako ROUGE-F1 balioak ere ikusgai daude. 3.14 irudian daude balioak eta Train 4Large modeloarekin alderatuz emaitza hobeak lortu dituzte PEGASUSekin, bai Large baita Base erabilia ere.

PEGASUS ereduaren aurreko artearen egoerarekin eta Transformer BASerekin alderatuta, Reddit TIFU datu multzoan emaitza hobeak lortu dira, gainontzekoetan berriz balio okerragoak. AESLCren kasuan, Train 1P modeloak emaitza hobeak lortzen ditu Transformer Baserekin alderatuta.

Metrikaren aldetik entrenatu diren ereduak ez dira oso erabilgarriak PEGASUSekin alderatuz. Baina, eredu hauek euskararako sortu dira eta corpusa ingelesetik itzultako testuekin sortu da, beraz, alderaketa hori ez da guztiz bidezkoa. Euskarazko aurrekaririk ez dagoenez datu multzo hauetarako, ez dago bidezkoa den alderaketa egiterik. Hori gutxi balitz, ROUGE balioak ez dira hizkuntzen artean alderatzeko egokiak.

Euskaraz: $ROUGE-F1(\text{euskara}, \text{euskaraz}) = 0.0$

Ingeleseaz: $ROUGE-F1(\text{basque}, \text{in basque}) = 0.66$

Euskaraz deklinabidea hitzari lotuta idazten denez atzizki moduan, unigramak osatzean

hitz osoa kontuan hartzen du, atzizkia eta erroa ezberdindu gabe, unigrama bakarra sortuz. Ingeleseztik, bi hitz erabiltzen direnez, unigramak osatzean bi sortuko dira. Ondoren kontaktak egitean, euskaraz akats moduan kontaktuko da, baina ingelesez bi unigrametatik bat asmatu dela. Beraz, posible da sortutako laburpen-ereduak ingelesezko PEGASUS ereduaren antzerako errendimendua izatea.

5.2.5.1 ROUGE metrika tokenizatzaileekin

Ingeleseztik eta euskarazko ROUGE metrikak ezin dira zuzenean alderatu ikusi den bezala. Horri aurre egiteko, tokenizatzaile bat erabili da metrika kalkulatzeko garaian. ROUGE-TOK izena eman zaio metrika honi eta formulak berdin mantentzen dira.

ROUGE metrika estandarrak besterik ezean hitzak banatzeko hutsunea erabiltzen du banatzaile gisa. Kasu honetan, tokenizatzaile bat erabili da, ingelesezko eta euskarazko balioak gerturatzeko asmoz. Zehazki, bai PEGASUS bai Train 4Large ebaluatzen MT0ren tokenizatzailea erabiltzea aukeratu da. Tokenizatzaile bera erabiliz, sortu daitezkeen alborapenak edo ezberdintasunak txikiagotu nahi izan dira, gainera elebiduna izanik arazorik ez dago. Jarraian tokenizatzailearen adibide bana ikus dezakegu bi hizkuntzetan:

Tokenizatzailea(Kaixo, zer moduz?) $\rightarrow \{ 'K' 'aixo' ' ' 'zer' 'modu' 'z' ' '? \}$

Tokenizatzailea(Hello, how are you?) $\rightarrow \{ 'Hello' ' ' 'how' 'are' 'you' ' '? \}$

Metrika berri honen bitartez, ROUGE metrika estandarrak euskaraz 0 balioa ematen dien esaldi batzuetan, balio egokiagoak lortzen dira. Lehen aipatutako adibidearen kasuan,

Euskaraz: $ROUGE-TOK(\text{euskara}, \text{euskaraz}) = 0.8$

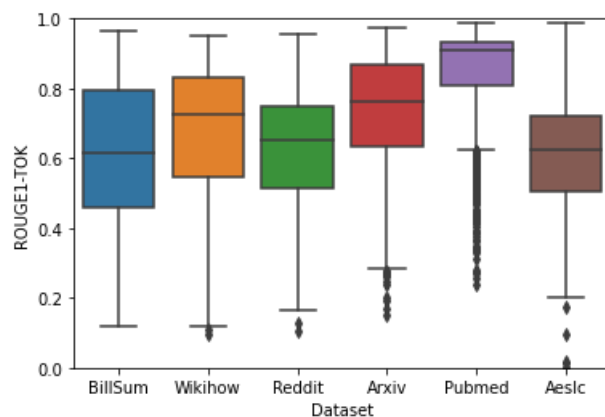
Ingeleseztik: $ROUGE-TOK(\text{basque}, \text{in basque}) = 0.8$

emaitza berdinak lortu dira.

Train 4Largeraren ROUGE-TOK balioak 5.16 irudian ikus daitezke eta 5.15 irudiaren aldean hobekuntza oso nabaria da. PEGASUS ereduaren kasuan, denbora faltagatik, 2 datu multzoko kalkuluak bakarrik egin dira. PEGASUSek ingeleseztik bakarrik dakienez, ingelesezko testuetan ebaluatu da. Train 4Large ereduak euskarazko testuetan ebaluatu da eta PEGASUS ingeleseztik. 5.2 taulan ROUGE-F1 eta ROUGE-TOK balioak ikusten dira, PEGASUS eta Train 4Large ereduaren.

ROUGE-TOK erabilia bi ereduaren metrika balioak igo dira, tokenak asmatzea errazagoa izango delako hitz osoak baino. Train 4Large ereduak metrika honen arabera PEGASUS baino hobea da datu multzo guztietan, gainera, alde ikaragarri batekin. Ondorioz, Train 4Large ereduak oso ongi laburtzen duela ondoriozta daiteke, balio altuak lortu direlako. Hala ere, ezin daiteke esan PEGASUS baino hobea denik, nahiz eta tokenizatzailea erabili ezin baita konparaketa hau bidezkoa dela esan.

ROUGE-TOK metrikaren erabilera egokiagoa dela dirudi euskarazko ROUGE metrika estandarra baino. Euskararen izaera dela eta hutsune banatzaileekin erabilia ezin da N-gramak modu egokian alderatzen eta horrek balio okerragoak lortzea eragiten du. Irtenbide egokiena ez izanik ere, tokenizatzailea erabilia bidezkoagoak diren balioak lor daitezke.

Irudia 5.16 Train 4Large modeloak lortutako ROUGE-TOK balioak.

Modeloa	AESLC	Arxiv	BillSum	Pubmed	Reddit	Wikihow	Bataz beste
Train 4Large (ROUGE-F1)	0,088	0,213	0,300	0,252	0,194	0,210	0,217
Train 4Large (ROUGE-TOK)	0,610	0,737	0,626	0,848	0,623	0,677	0,692
PEGASUS (ROUGE-F1)	0,377	0,447	0,572	0,455	0,265	0,431	0,424
PEGASUS (ROUGE-TOK)	–	0,581	0,409	–	–	–	0,501

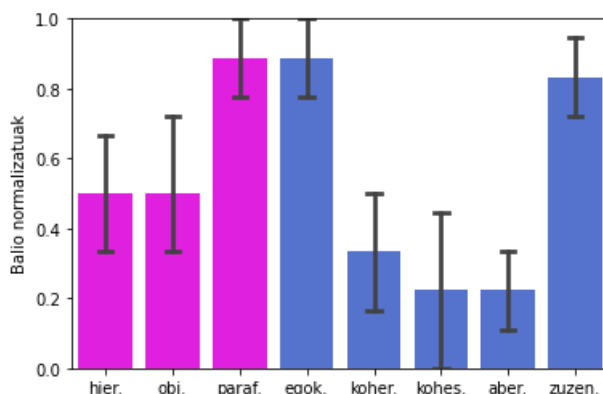
5.2 Taula: Train 4Large eta PEGASUS ereduaren ROUGE-F1 eta ROUGE-TOK balioak.

5.3 Eskuzko ebaluazioa

Eskuzko ebaluaziorako 3.3.2 atalean deskribatutako irizpideak erabili ditu ebaluatzaileak. Irizpideetatik egokitasuna, koherentzia, kohesioa, aberastasuna eta zuzentasuna HABEko azterketak ebaluatzeko ohikoak dira. Laburpenak ebaluatuko direnez, hierarkia, objektibitatea eta parafraisi kantitateak ere kontuan hartuko dira, beharrezkoak baitira ebaluazio egokiagoa egiteko.

Eskuzko ebaluazioa automatikoa ez bezala, oso neketsua eta zaila da. Gainera, irizpideak testu guztien gainean modu berean aplikatu direla bermatzea ez da erraza. Baina, ebaluatzailearen eskutik esfortzu berezia egin da ahal den modurik objektiboenean ebaluatzeko, hasieran hartutako erabakiak mantenduz.

Eskuzko ebaluazioak behar duen denbora kontuan harturik, eredu bakarra ebaluatzea erabaki da. Hipotesi gisa, ROUGE metrika eta eskuzko ebaluazioak korrelazioa izango

Irudia 5.17 Wikihow-eko testu hobereenen bataz besteko irizpideen balioak.

dutela kontsideratu da, zenbat eta erreferentziaren antzekoagoa izan sortutako laburpena, orduan eta laburpen hobeak izango dira. Ondorioz, eredia ROUGE metrika bidez aukeratu da, balio hobereenak lortu dituen, Train 4Large deiturikoa, izanik.

Eredua ebaluatzeko 24 testu erabili dira guztira. 20 Wikihow datu multzotik hartu dira eta 4 elikagaiei buruzko Unai utzitako testuak. Elikagaiei buruzko testuak, informazio-testu ulertterazak dira, bataz beste 472,5 token dituzte, beraz, ereduak ikusitako testu luzeretara egokitzen dira. Wikihow-ekoak aldiz, ROUGE metrika erabiliz aukeratu dira. Ereduek metrikari jaramon eginda, egokien eta okerren laburtzen dituen 20 testu hautatu dira, 10 onen eta 10 okerren izanik.

Jarraian testu multzo bakoitzetik ateratako ondorioak eta emaitzak azalduko dira.

5.3.1 Wikihow-eko testu onenak

5.17 grafikan irizpide bakoitzetik lortutako bataz besteko notak ikus daitezke. Notak normalizatuta daude, konparaketak egokiago egiteko. Euskararen maila ona dela ikus daiteke egokitasuna eta zuzentasunak oso altuak direla eta, baina, koherentzia, kohesioa eta aberastasunean eskas dabil eredia. Laburpenerako irizpideetan berriz, parafasian bikain bat lortu du eta hierarkian eta objektibitatean berriz 5eko bat.

Orokorrean, ebaluatzailearen iritziz, erreferentziako testua eta ereduak sortutakoa oso antzekoak izan dira testu hauetan. Kasu askotan erreferentzien parafasiak izanik, horrek ROUGE balio altuak izatea arrazoitzen du.

Testu bakoitzari nota zehatzak jarri dizkio irizpide bakoitzari pisuak emanda, 5.3 taulan nota hauek ikus daitezke. Bataz bestean, 0,56ko balioa lortu da. Testu batzuek bikain bat lortu dute eta beste batzuk berriz gutxiegi, baina, nota hauek ez dute benetako pisurik laburpenak ebaluatzeko orduan. Kasu gehienetan, laburpenen esaldiak oso motzak eta sinpleak direlako, giza laburpenetatik urrun geratuz.

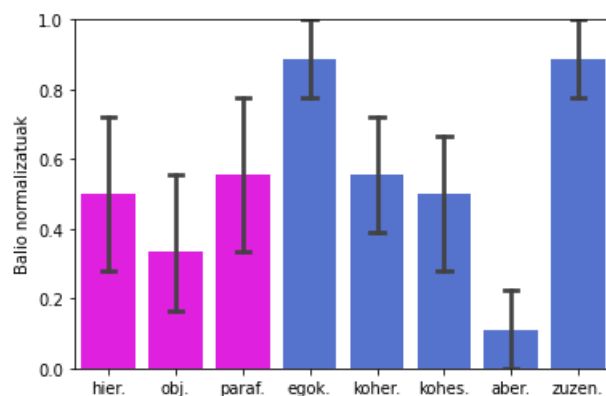
Irizpideen ebaluazioak ikusita, bi ondorio nagusi atera daitezke. Lehena, euskararen erabilera egokia egiten duela, ereduak euskaraz ongi idazten daki eta sortutako laburpenetan islatzen da kalitatea. Bigarren, nahiz eta nota ona lortu, giza laburpen maila ez da lortu. Esaldi asko oso motzak eta zehatzak dira, baina adibidez,

Ez edan alkoholik oheratu baino ordu batzuk lehenago. Lasaitu zaitetz oheratu aurretik.

WHO1	WHO2	WHO3	WHO4	WHO5
0,58	0,95	0,7	0,58	0,54
WHO6	WHO7	WHO8	WHO9	WHO10
0,35	0,38	0,38	0,38	0,84

5.3 Taula: Wikihow datu multzoko ROUGE balio altuena duten testuen eskuzko ebaluazioko nota.

Irudia 5.18 Wikihow-eko testu okerrenen batz besteko irizpideen balioak.



Saiatu alde batera jartzen lo egiten duzunean.

laburpenean, euskara maila egokia da, baina, ez du giza laburpenen antzik. Wikihow-eko erreferentziako laburpenak ere honelako esaldiak dituzte, beraz, ulergarria da ereduaren joera.

5.3.2 Wikihow-eko testu okerrenak

Kasu honetan, ROUGE metrikaren arabera testu okerrenak direnak aukeratu eta ebaluatu dira. 5.18 irudian notak ikus ditzakegu eta orokorrean Wikihow-eko testu onenen irizpideen balioak gogorarazten ditu.

HABE-n erabiltzen diren irizpideen kasuan, kohesio eta koherentzia hobeak dutela dirudi, baina, aberastasunean beherakada egon da. Laburpen irizpideen kasuan, parafraasi kantitate gutxiago dago eta objektibitatean ere beherakada nabari da.

Badirudi, ROUGE balioak eta laburpenen kalitateak korrelazioa dutela, Wikihow-eko testu onenen kasuan laburpen irizpideen balioak hobeak direlako. Hala ere, testu okerrenen kasuan hizkuntzaren aldetik, aberastasunean izan ezik, hobeak dira testu egokiak baino. Baina, ondorio hauek kontuz hartzekoak dira, guztira 20 testu ebaluatu direnez, zaila da ROUGE eta laburpen kalitateak korrelazio zuzena dutela esatea eta ROUGE eta hizkuntzaren kalitateak alderantzizko korrelazioa.

5.4 taulan Wikihow-eko testu okerrenen notak ikus daitezke, irizpideei pisu ezberdinak emanda. Batz bestean, 0,56 balioa lortu da, hoberenen nota berdina izanik. Arrazoieta-ko bat, lehen esan bezala, hizkuntza aldetik testu okerrenak egokiagoak direla eta testu hoberenak laburpen aldetik izan daiteke.

WHT1	WHT2	WHT3	WHT4	WHT5
0,56	0,73	0,46	0,73	0,73
WHT6	WHT7	WHT8	WHT9	WHT10
0,30	0,62	0,22	0,68	0,64

5.4 Taula: Wikihow datu multzoko ROUGE balio baxuena duten testuen eskuzko ebaluazioko nota.

5.3.3 Elikagaien laburpenen ebaluaketa

Azkenik, elikagaien 4 testu ebaluatu dira. Testu hauek ereduak inoiz ikusi gabeko testu motak dira, entrenamenduan ez du antzeko testurik ikusi. Hori dela eta *prompt*-ekin hainbat proba egin dira egokiena aukeratzeko. Informazio-testuak direnez eta testu luzera ertainekoak direnez, Arxiv, Pubmed eta BillSum datu multzoetan erabilitako *prompt*-a erabili da azkenean.

Irizpideak aplikatzerako orduan, arazoak zeuden aurreko bi testu multzoekin alderatuta. Wikihow-eko testuetan erabilitako irizpideen puntuaketen esanahia mantentzea zaila da, testu mota oso ezberdinak izanik. Wikihow-eko testu eta laburpenen idazkera, esaldi laburrak eta elkar lotura gehiegi gabekoak dira eta elikagaien testuak berriz, testu izaera gehiago dute, esaldi ertainak, elkar lotuta edota paragrafo ezberdinak etab. izanik. Hala ere, ebaluatzailea saiatu da bi testu moten arteko ebaluazioa ahalik eta parekatuen egiten, ondoren emaitzak konparagarriak izan daitezten.

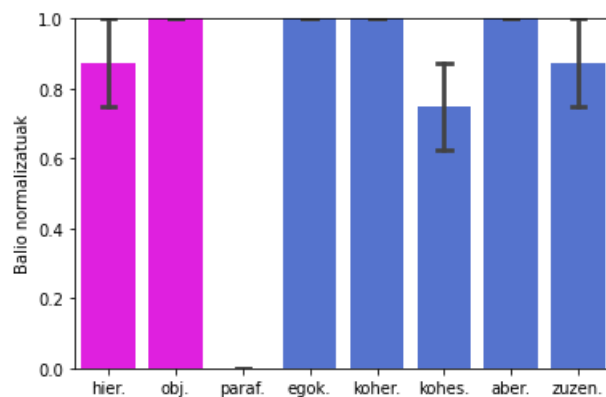
5.19 grafikan 8 irizpideen notak ageri dira. Elikagaien laburpenek euskararen erabilera ona eta hierarkia eta objektibitatea ere oso onak dituzte. Hala ere, parafrasi kantitatean huts larria egin du, honen zergatia, token-ratio izeneko ezaugarri batetik ondoriozta daiteke.

Token-ratioak sarrerako testuaren eta laburpenaren arteko token kopuruaren proportzioa adierazten du. 5.5 taulan token-ratio balioak agertzen dira testu bakoitzeko. Elikagaien testuetan, 0,5eko batz besteko proportzioa dago token kopuruaren artean. Lehen bi laburpenen kasuan, laburpenek ia testu guztia barne hartzen dute, ondorioz, euskararen maila egokia izango da sarrerako testuaren hizkuntza egokia delako eta hierarkia eta objektibitatea ere oso ongi dago jatorrizko testuko ideia guztiak biltzen baititu. Gainontzeko bi testuen kasuan, nahiz eta token-ratio balio txikiagoa izan, ebaluatzailearen iritziz egoera berdina errepikatzen du ereduak.

Testua	Testuaren token kopurua	Laburpenaren token kopurua	Token-ratio
Eli 0	285	221	0,78
Eli 1	332	256	0,77
Eli 2	559	190	0,34
Eli 3	714	83	0,12
Bataz beste	472,5	187,5	0,5

5.5 Taula: Elikagaien testuetako token-ratio balioak.

5.6 taulan elikagaien testuetako notak ikus ditzakegu eta espero den bezala nota oso altua lortu dute. Baina, ezin dira laburpen gisa hartu, lehen aipatu bezala, laburpena ia testuaren kopia baita.

Irudia 5.19 Elikagaien testuetako bataz besteko irizpideen balioak.

Eli0	Eli1	Eli2	Eli3
0,89	0,73	0,73	0,89

5.6 Taula: Elikagaien testuen eskuzko ebaluazioko nota.

Arazoaren arrazoia erabilitako *prompt*-a izan liteke. Artikuluetan erabili den *prompt*-arekin laburtu dira elikagaien testuak. Artikuluetak laburpenak, *abstract*-ak dira eta ohikoa izaten da sarreraren antzeko informazioa edukitzea. Ondorioz, *prompt* hori erabiltzean, kopia egin behar duela ikasi du.

Badirudi ereduak inoiz ikusi gabeko testu motetan laburpenak egiteko zailtasunak dituela, baina, Wikihow-eko testuak laburtzen ikasi duenez horiek egoki itzultzeko gai dela, aipatutako arazoak kontuan hartuz.

Ondorioak eta etorkizuneko lanak

GrAL honetan euskarazko abstrakzio-laburpenak egiten dituen sistema garatu nahi izan da. Horretarako kodetzaile-dekodemak moduko *transformer*-ak diren MT0 izeneko HEEi *fine-tuning*-a egin zaie. Entrenamendurako corpora PEGASUS ereduak entrenatzeko erabilitako datu multzoak itzulpen automatiko bitartez euskaratuz sortu da. Guztira 6 datu multzo itzuli dira, 728 mila testu-laburpen pare dituen corpora sortuz. Hainbat entrenamendu egin dira, laburpen-eredu egokiena entrenatzeko moduak esploratuz. Laburpen-ereduak alderatzeko, ROUGE metrika automatikoa erabili da eta ondoren, eredu onenari eskuzko ebaluazioa egin zaio Unai Atutxaren tesiko irizpideak erabiliz.

Proiektuan zehar hainbat ondorio eta ekarpen egin dira eta hurrengo ataletan zehaztasunez azalduko dira.

6.1 Itzulitako testuak eta corpora

Corpora osatzeko, PEGASUS ereduak entrenatzeko erabilitako datu multzoak, baina, MT0k ikusi gabeak euskaratu dira. Baliabide eta denbora faltagatik, 6 datu multzo itzuli dira Wikihow, Reddit TIFU, Arxiv, Pubmed, BillSum eta AESLC. Corpusak guztira 728.128 testu-laburpen ditu.

4.2.2 atalean itzulitako testuen BLEU balioak kalkulatu eta aztertu dira. Ondorioz, itzulpenak onargarriak direla ikusi eta entrenamendurako egokiak direla suposatu da. Azkenik, itzulpenetan agertzen diren akats pare bat identifikatu eta aztertu dira.

Sortutako corpora oso erabilgarria izan daiteke laburpenak ez diren atazetarako. Nahiz eta itzulpen automatiko bidez itzuliak izan, eruedetan ez du eragin negatiborik sortzen dituzten testuen hizkuntza mailan. Ondorioz, corpora ekarpen garrantzitsua dela esan daiteke eta itzulpen automatiko bidez itzulitako datu multzoak erabilgarriak izan daitezkeela ikusi da.

6.2 Entrenatutako ereduak

4.3.3 atalean azaldu den bezala, guztira 6 eredu entrenatu dira sortutako corpora erabilita. Entrenamenduetan *prompt*-aren eragina, optimizatzaileen aldea, entrenamendu luzeen

eragina eta eredu handiagoak erabiltzearen aldea aztertu da. Jarraiko ondorioak atera dira entrenamendu hauetan.

1. *Prompt*-en erabilerak eragin positiboa du entrenamenduan. Bai galera aldetik, bai ROUGE metrikaren aldetik emaitza hobekuntza lortu dira. Gainera, sarrerako testu mota zehazteko erabili da lan honetan, datu multzo bakoitzari *prompt* bat esleituz.
2. Adafactor optimizatzailea erabiltzeak onura asko dakartza entrenamenduan. AdamW-rekin alderatuta, memoria erabilera txikiagoa da eta emaitza parekoak edo apur bat hobekuntza lortu dira.
3. Entrenamendu luzeagoa egiteak galera aldetik hobekuntza dakar, baina ez da gehiegi hobetu. Erabili den ikasketa-tasaren planifikatzaileak balioa azkar jaisten du, ondorioz, ikasketa luzeen azken zatian, ikasketa motelagoa da. Metrika aldetik ere ez da hobekuntza handirik lortu eta batzuetan emaitza okerragoak lortu dira.
4. Eredu handiagoa erabilia emaitzarik onenak lortu dira. Galera eta metrika aldetik hobekuntza ere eredu txikiagoak luzaroan entrenatzea baino.

Eredurik handiena, Train 4Large, izan da metrika aldetik emaitzarik hoberenak lortu dituenak. ROUGE-F1 metrika estandarra erabilia, PEGASUS edo artearen egoera diren ereduetatik urrun dago Train 4Large. Hala ere, bi hizkuntzen arteko balioak konparatzea zaila da ROUGE metrika erabilia, ez baitira hizkuntzen ezaugarriak kontuan hartzen. Konparagarriak egiteko saiakera bat egin da tokenizatzailea erabiliz 5.2.5 atalean eta PEGASUS baina emaitza hobekuntza lortu dira metrika berri honen arabera. Badirudi tokenizazioak eragin handia duela metrikak ematen dituen emaitzetan. Sarrera tokenizatzean nabari gogortzen direla balioak, ingelesekoetara hurbilduz. Dena den, ez dira konparagarriak, baina bi hizkuntzen arteko konparaketa egokiago bat egiteko modua ematen du.

Unai Atutxak eskuzko ebaluazioa ere egin du, bere doktorego tesian oinarritutako irizpideak kontuan hartuta. Ebaluazioaren denbora kantitatea kontuan hartuta, eredu bakarra ebaluatu da, ROUGE baliorik hoberenak eman dituenak eta Wikihow datu multzoko 20 testu aukeratu dira. Testu horiek ROUGEren arabera egokien eta okerren laburtutako 20 dira. Gainera, Unai Atutxak erabilitako elikagaiei buruzko 4 testu ere ebaluatu dira. Ebaluaziotik 3 ondorio nagusi atera dira.

1. Euskaraz ongi idazten daki ereduak, nahiz eta sarrerako testuaren kalitatea eskasa izan.
2. Entrenamenduan ikusi dituen datu multzoak, Wikihow adibidez, ongi laburtzen ikasi du orokorrean. Datu multzo horretako idazkera barneratu du, nahiz eta ROUGE metrikaren arabera oso txarrak izan.
3. Ikusi ez dituen datu multzoko testuak, elikagaienak adibidez, laburtzean arazoak ditu. Elikagaiekin erabili den *prompt*-a artikulua laburtzeko erabili denez, modu berean laburtzen ditu, hau da, sarrerako testua oso luzea denez, laburpena ere luzea da. Ondorioz, kasu batzuetan laburpenak eta testuak oso antzekoak dira.

Laburbilduz, entrenatutako ereduak ikasitako datu multzoetan ongi laburtzen du. Ikusi gabeko testu motetan berriz, ez du hain emaitza onik lortzen. Ondorioz, giza hurbilpen mailatik urrun dagoela esan daiteke.

6.3 Prompt-aren garrantzia

Train 1 eta Train 1P modeloen metrikaren aldea, Reddit TIFU testuetan dago, Arxiv eta AESLCren kasuan antzeko joera izanik. Honek, *prompt*-ak erabiltzeak onura gutxi dakarrela dirudi, baina, testuak irakurrita, laburpen hobekak dirudite Train 1P modeloek sortutakoak, gainera, bibliografia aztertuz, artikulua asko idatzi dira *prompt*-en onura azalduz. Adibidez, [33] artikuluan *few-shot* moduko entrenamenduak *prompt*-ekin egiteak duen onura azaltzen dute.

Entrenamenduetan erabilitako *prompt*-ek modeloari testu motak ezberdintzen erakutsi diete. Hori dela eta, testu bakoitzerako erabiliko den *prompt*-a egoki aukeratzeak pisu handia du laburpenak sortzeko garaian. Adibidez, Unairen elikagaien eli3 testuarekin argi ikusten da *prompt*-en eragina, 6.1 taulan agertzen dira *prompt* ezberdinekin sortutako testuak.

Laburpena	<i>Prompt</i> -a
Elikagai erregulatzaileak ezinbestekoak dira organismoak funtzionamendu zuzena izan dezan. Mantenuz garrantzitsuenak bitaminak, gatz mineralak eta ura dira. Oro har, kopuru txikitik behar izaten ditu gorputzak ura izan ezik, ura ugari hartu behar izaten baitu. O, ez!	First, read the Basque article below. {{testua}} Now, please write a short abstract for it in Basque. (Arxiv, Pubmed eta BillSum datu multzoetan erabilitako <i>prompt</i> -a)
Ulertu elikagaien funtsezko osagaiak	{{testua}} TL;DR in Basque: (Wikihow datu multzoan erabilitako <i>prompt</i> -a)
Elikagai erregulatzaileak	{{testua}} How would you rephrase that briefly in Basque? (AESLC eta Reddit TIFU datu multzoetan erabilitako <i>prompt</i> -a)

6.1 Taula: Elikagaien testuen laburpenak *prompt* ezberdinekin.

Bi adibideen arteko aldea ikaragarria dela ikusten da eta hau datu multzo ezberdinetako *prompt*-ak nahastuta ere nabari da. Adibidez, AESLCko testu bati *prompt* ezberdinak jarrita sortzen diren laburpen ezberdinak 6.2 taulan ikus daitezke. Jatorrizko laburpena 2 hitzekoa da eta *prompt* ezberdinak erabiliz testu luzeagoak ere sor daitezke. Laburpenak irakurritik, lehenengoa AESLCrako aukeratutako *prompt*-arekin egindakoa da eta kasu honetan egokiagoa dirudi gainontzeko bi *prompt*-ak erabiltzea laburpenak sortzeko.

Ondorioz, *prompt*-en garrantzia ikaragarria dela nabarmentzen da. Modelo hauek la-

Sortutako laburpena	<i>Prompt-a</i>
Programa aldaketak	<i>Erreferentziako laburpena</i>
Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak	{{testua}} How would you rephrase that briefly in Basque (AESLC eta Reddit TIFU datu multzoetan erabilitako <i>prompt-a</i>)
2001eko urtarrilaren 1ean, Enron-ek bertan behera utzi zituen ClickAtHome eta Opariak batzen eta boluntarioak bultzatzeko programak. O, ez!	First, read the Basque article below. {{testua}} Now, please write a short abstract for it in Basque. (Arxiv, PubMed eta BillSum datu multzoetan erabilitako <i>prompt-a</i>)
Enron-eko trantsizio aldiaren, hainbat programa bertan behera uztea beharrezkoa da.	{{testua}} TL;DR in Basque: (Wikihow datu multzoan erabilitako <i>prompt-a</i>)

6.2 Taula: AESLCko testu bat *prompt* ezberdinak erabiliz laburtuta.

burpen sistema moduan erabili nahi badira, *prompt* ezberdinekin jokatzeari oso garrantzitsua izango da ahal den laburpen egokiena lortzeko.

6.4 Mugak eta etorkizuneko lanak

Baliabide edota denbora faltagatik, egiteke edo aztertzeke geratu dira ataza edo lan batzuk. Kapitulu honetan, egin gabe geratu diren lanen ideiak eta izandako arazoak zerrendatuko dira.

6.4.1 Corpusaren garbiketarako eta eskuzko kalitate ebaluazioa

Corpusean itzulpenen akatsen ondorioz, 3 esalditxoren agerpenak aztertu dira 4.2.2.1 atalean. Hala ere, itzulpen prozesuak bestelako akatsak sortu ditu corpusean, horrek corpusaren kalitatea jaisten du. Kasu batzuetan ez litzateke zaila izango konponketak egitea, baina, denbora eta baliabide faltagatik ez dira konponketak egin.

Aurkitu eta identifikatu diren akats ohikoak zerrendatu dira.

- Hitz tekniko edo bitxiak gaizki itzulita daude batzuetan. Hitz erdi batzuk hizkuntza batean jartzen ditu eta gainontzeko zatia beste batean. Adibidez, "Veteranoen", "oscilazio" eta "espectroskopikoen". Konpontzeko, corpuseko hitzak hiztegi tekniko batekin aldera daitezke eta [Levenshtein distantzia](#) erabiliz zuzendu.

- Latex-eko agindu edo formatu bereziak agertzen dira bereziki Arxiv eta Pubmed datu multzoetan. Kasu batzuetan, formulak ere gehituta daude testuetan. Kasu honetan, zuzenean ezabatu litezke, adierazpen erregularren bitartez adibidez.
- Esaldi batzuk ez dute zentzurik euskaraz. Itzulpenean, esaldi batzuen esanahia galdu da, hizkuntza ezberdinetan aditz edo hitz ezberdinak erabiltzen direlako deskribapenetan. Adibidez, "Jarri kola pixka bat txikitan eta inguruko lekuan." esaldian "txikitan" hitzak ez du gaiarekin zerikusirik. Kasu honetan, azterketa sakona beharko luke eta oso zaila izango litzateke honelako akatsak zuzentzea.
- Ebaluatzailearen hitzetan, irakurri dituen testu batzuk ez dira ulergarriak. Ereduak sortutako laburpenak ulergarriak ziren, baina, sarrerako testua bera oso zaila zen ulertzeko. Kasu hau ezin da guztiz konpondu, oso zaila da gaizki idatzitako esaldiak konpontzea.

Akatsak alde batera utziz, interesgarria izango litzateke datu multzo bakoitzeko testu batzuk hartzea eta eskuzko ebaluazio txiki bat egitea. Ereduen eskuzko ebaluazioan bezala, testu gutxi batzuk ebaluatzea nahikoa izan daiteke ideia orokor bat egiteko. Baina, ebaluatzaile bat beharko litzateke denbora luzez ebaluatzen, ondorioz, egiteke geratu da.

6.4.2 Datu multzoen eragina entrenamenduetan

Entrenamendu ia guztietan corpus osoa erabili da entrenamenduak egiteko. PEGASUSen kasuan adibidez, datu multzoak banaka ikusi ditu ereduak. Hau da, lehenik datu multzo entrenatzen dute, ondoren bigarren datu multzoarekin, denak ikusi arte. Modu honetara, testu moten eragina ikustea errazagoa da eta hiperparametro ezberdinak erabil daitezke, testu mota bakoitzari zuzenduta.

GrAL honetan, interesgarria litzateke itzulita dauden datu multzoen arteko eragina aztertzea. Adibidez, Arxiv-eko testuak erabiliz entrenatzea eta ondoren Pubmed-en ebaluatzea. Ikerketa aldetik, interesgarria izan daiteke ekarpen handiena zein datu multzok duten ikertzea. Baina, denbora eta baliabideen arazoagatik ezin izan da burutu.

6.4.3 Datu multzo gehiago itzultzea

Corpusa osatzeko 728 mila testu-laburpen pare inguru itzuli dira, hala ere, hasierako helburua 3,5 milioi pare itzultzea zen, baldintzak betetzen zituzten eta eskuragarri zeuden testu kopurua zelako. Baina, denbora izugarria behar da hainbeste testu itzultzeko, beraz, 700 mila inguru bakarrik itzuli dira.

6.4.4 Prompt ezberdinak sortzea

Entrenamenduetan erabili diren *prompt*-ak MT0 entrenatzeko erabili zirenak izan dira. Datu multzo bakoitzari prompt bat esleitu zaio, ondorioz, datu multzo barneko testu mota ezberdinak ez dira kontuan hartu. Interesgarria izango litzateke *prompt*-ak eskuz sortzea eta datu multzoka esleitu beharrean, testu motaren edo gaiaren araberako esleipena egitea. Gainera, MT0ren *prompt*-ak ingelesez daude, beraz, euskarazko *prompt*-ak sortzea egokiago izango litzateke GrAL honetarako. Berriro ere, denbora eta baliabide faltagatik ezin izan da *prompt*-etan gehiegi sakondu.

Plangintza eguneratua

Proiektuaren hasieran egindako plangintzak desbiderapen batzuk jasan ditu atazak buru ahala. Aldaketarik handiena, ataza bakoitzari eskainitako denboran nabaritu da, corpora sortzeak, ereduak entrenatzeak eta ebaluatzeak uste baina ordu gehiago behar izan ditu eta.

A.1 taulan ataza bakoitzari benetan eskainitako denbora ikus daiteke eta hasieran egin zen estimazioarekin alderatzen da. Taulan, aldaketa nabariak izan dituzten lan-paketeak bakarrik agertzen dira, gainontzekoetan estimazioa nahiko zehatza izan baita. ”*” bidez adierazitako atazak, automatikoki makinetan egin diren exekuzioak adierazten dituzte, beraz, ez dira adierazitako ordu guztiak eskuzko lan orduak izan. Ondorioz, ordu asko behar izan badira ere, eskuzko lana gutxiago izan da.

Lan-paketea	Ataza	Denbora estimazioa(ordutan)	Denbora erreala(ordutan)
Memoria	M.1	5	10
	M.2	30	50
	M.3	5	10
	Guztira	40	60
Testuak itzultzea	TI.1	5	10
	TI.2	50	200*
	Guztira	55	210
Exekuzioak	Ex.1	100	380*
	Ex.2	20	20
	Guztira	120	400
Ebaluaketak	Eb.1	15	50*
	Eb.2	2,5	2,5
	Guztira	17,5	52,5
Guztira, proiektu osoari eskainitako denbora		312	802

A.1 Taula: Ataza bakoitzari benetan eskaini zaion denbora.

A. PLANGINTZA EGUNERATUA

Atazen denboraldien kasuan, ez da orokorrean desbiderapenik izan. Exekuzio eta ebaluazio kontuak pixka bat atzeratu dira eta ekaineko lehen astean ere azken ukituak eman zaizkio atal horri. A.2 taulan kolore gorritz adierazita ikus daitezke egon diren desbiderapenak. IB.2 ataza berriz hartu behar izan da automatikoki ebaluazioa egiterako orduan, metrikak ulertu eta gainontzeko ereduekin ongi alderatzeko.

Ataza	urt.	ots.	mar.	apr.	mai.	eka.
PL.1	█					
PL.2	█					
PL.3	█					
PL.4						█
JK.1		█	█	█	█	█
JK.2		█	█	█	█	█
M.1	█	█	█	█	█	
M.2				█	█	
M.3					█	█
BI.1	█					
BI.2	█					
TI.1		█				
TI.2		█	█	█		
CA.1		█	█	█		
CA.2				█		
KG.1		█				
KG.2		█	█			
Ex.1			█	█	█	█
Ex.2			█	█	█	█
Eb.1			█	█	█	█
Eb.2			█	█	█	█
IB.1	█	█				
IB.2					█	█
BeO.1	█	█	█	█	█	█
BeO.2	█	█	█	█	█	█
LP.1	█					
LP.2	█					

A.2 Taula: Ataza bakoitzaren denboraldiak.

Sortutako laburpenen adibideak

Eranskin honetan, Train 4Large ereduak sortutako laburpenen adibideak agertzen dira.

B.1 Datu multzoak

B.1.1 BillSum

Testua:

First, read the Basque article below. 1. atala Izenburu laburra Lege honi "2008ko beteranoen konpentsazioaren bizi kostuen egokitzapen legea" deritzo. Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Bi, bi, bi, bi, bi. Desgaitasunaren ordainketen eta mendekotasunaren eta indemnizazioaren tasaren igoera a) Tasa-aldaketa -2008ko abenduaren 1ean indarrean, beteranoen gaietako idazkariak, c) azpitoluaren arabera, gehituko ditu 2008ko azaroaren 30ean indarrean zeuden dolarreko zenbatekoak, b) azpitoluan zehazt b) Gehitu beharreko zenbatekoak -Azal (a) honekin gehitu beharreko dolarreko zenbatekoak honako hauek dira: (1) Gerra garaiko ezintasunaren kalte-ordaina -Estatu Batuetako Kodearen 38. artikuluen 11.14 atalean xedatutako dolar bakoitza. (2) Zaindariantzako konpentsazio gehigarria -Dolarezko zenbateko bakoitza, 1115 (1) artikuluen arabera. (3) Jantzientzako kuota -Dirua, 1162 atalaren arabera. (4) Bizirik dagoen senargaiaren mendekotasuna eta kalte-ordaina -Dolarreko zenbateko bakoitza, atal horretako 1311 ataleko (a) - (d) azpitoluen arabera (5) Haurren mendekotasuna eta kalte-ordaina -Dolarreko zenbateko bakoitza 1313a eta 1314 artikuluen arabera da. c) Gehikuntzaren zehaztapena - (1) Ehuneko -2. paragrafoan ezarritakoa salbu, b) letran aipatzen den dolarreko zenbateko bakoitza, Gizarte Segurantzaren Legearen II. tituluaren arabera ordaindu beharreko prestazioen ehunekoaren ehuneko berberarekin gehituko da. Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak 401 eta hurrengoak) 2008ko abenduaren 1etik aurrera handitzen dira, lege horren 215. atalaren arabera zehaztutakoaren ondorioz (42 U Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak 415 (i) (2) Biribiltzea - (1) paragrafoaren arabera gehitutako dolarreko zenbateko bakoitza, dolarreko zenbateko osoa ez bada, hurrengo dolarreko zenbateko osoa izango da. d) Arau berezia - Beteranoen Sailak administrazio aldetik zuzendu ahal izango ditu, a) atalean ezarritako gehikuntzen

arabera, 85-857 Lege Publikoaren 10. atalaren arabera pertsona gaixoei ordaindu beharreko ezgaitasun-ordainaren tasak 1263) ez baitute indemnaziaziorik jaso 38 tituluko 11. kapituluaren arabera e) Tasa egokituak argitaratzea –Alderdi Zaharren Saileko idazkariak (Federal Register) argitaratuko ditu (b) idatz-zatian zehaztutako zenbatekoak, atal horren arabera gehituak, Gizarte Segurantzaren Legearen 215. atalean zehaztutako gaiak Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak 415 ((i) ((2) ((D)) argitaratu behar dira, 2009ko ekitaldian lege horren 215 ((i) atalaren arabera egindako erabakia dela eta. Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak Hiru 2007ko KODIFIKATUA Bizi-kostuen zuzenketa desgaitasunaren kalte-ordainaren eta mendekotasunaren eta indemnazioaren tasetan a) Beteranoentzako kalte-ordainak –Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak b) Bizirauteko pertsonentzako konpentsazio gehigarria –Azal horretako 1115 (1) atala aldatzen da– (1) A atalean, “\$139’ ezabatzea eta “\$142’ sartzea; (2) B atalean, “\$240’ ezabatzea. c) Aldaketa handiak dituzten beterano batzuentzako arropa-prestazioa –Titulu horretako 1162. atala aldatzen da, "662"ezabatzearekin eta "677"sartzearekin". d) Bizirik geratzen diren senideen mendekotasun- eta kalte-ordain-ordaina – (1) Lege berria –A, 1. paragrafoan, » 1.067"e
Now, please write a short abstract for it in Basque.

Erreferentzia zko laburpena:

2008ko beteranoen konpentsazioaren eta bizi kostuen egokitzapen-legea - 2008ko abenduaren 1etik aurrera beteranoen alorreko idazkariari agintzen zaio beteranoen ezintasun-ordainaren tasak handitzea, mendekoent Gizarte Segurantzaren Legearen II. atalaren (Zahartzaro, Biziraupen eta Desgaitasun Aseguruak) arabera emandako prestazioen ehunekoaren ehuneko bera izan behar du igoera horietako bakoitzak, indarrean dagoen egun 2007ko Legearen arabera hobetutako tasak kodifikatzen ditu, beteranoen konpentsazioaren eta bizi kostuen aldaketaren arabera.

Train 4Large ereduak sortutako laburpena:

2008ko beteranoen konpentsazioaren bizi kostuen egokitzapen legea - Veteranoen gaietako idazkariari agintzen dio 2008ko azaroaren 30ean indarrean zeuden dolarreko zenbatekoak handitzea, Gizarte Segurantzaren Legearen II.

B.1.2 Wikihow

Testua:

Kola ere erabil dezakezu, eskola-kola zuria, txikleak kentzeko. Jarri kola pixka bat txikitatu eta inguruko lekuan. Ziurtatu kola nahikoa lodia dela zati osoa estaltzeko Ez erabili kola berehalakoa. Hau ez da zure larruazalaren gainetik jausiko eta orbanak zure azalean trabatuko dira kendu beharrean. Saia zaituz argizariarekin ilea kentzen edo argizari-zirriborroak erabiltzen kola bezala. Hasi aurretik, garbitu eta lehortu eskuak eta txiklearen inguruko aldea. Kola guztiz lehortu behar da kendu aurretik, bestela ez da zizarean itsatsiko. Utzi kola azalean 30 minutu edo ordubete. Begiratu ea lehor dagoen. Kola lehor dagoenean, ez da itsaskorra edo bustia sentituko. Kola lehorra dagoela ziurtatu ondoren, hartu kola ertzetik eta tiratu izurria zure azalean sartu den norabidean. Atera poliki eta modu uniformearekin. Pitzadurak itsatsita dauden bitartean, tira egin behar da. Kola kendu ondoren, aztertu kola, ea txistua harrapatuta dagoen. Gainera, aztertu behar duzu ea azalean orbainik geratzen den. Horrela bada, agian prozesua errepikatu beharko duzu edo beste metodo bat probatu.
TL;DR in Basque:

Erreferentziazko laburpena:

Jarri kola txatal horretan. Utzi kola lehortzen. Kendu kola. Begiratu ea aztarnarik badagoen.

Train 4Large ereduak sortutako laburpena:

Jarri kola pixka bat. Utzi kola lehortzen. Kendu txiklea.

B.1.3 Reddit***Testua:***

Nire anaia oftalmologora joan da gaur eta nire betaurrekoak daramatzala ikusi du. Hiru urtez Nolabait aldatu egin ziren lehen aldiz hartu genituenean. *How would you rephrase that briefly in Basque?*

Erreferentziazko laburpena:

Nire anaiak hiru urte daramatza betaurrekoekin.

Train 4Large ereduak sortutako laburpena:

Nire betaurrekoak aldatzen.

B.1.4 Arxiv***Testua:***

First, read the Basque article below. @xmath0 scuti- motako pulsatzailak izarrak dira, masekin 1 artean. 5 eta 2 5 @xmath1 cepheid klasikoaren eta sekuentzia nagusiaren ezegonkortasun lerroaren elkargunean kokatua A eta F espektro-mota dute, eta tarte hori 0-ra doa. 5 eta 6 ordu, eta, oro har, pulsatzeko, erradialak eta ez-erradialak, @xmath2-mekanismoak piztuta, bigarren helioko ionizazio eremuarekin lotuta. Oszilazio modu hauek izarraren barruko sakontasun desberdinetara murgiltzen dira. Hortaz, @xmath0 izarrek aukera ona ematen dute tarteko masa duten izarren barne egitura aztertzeko. @xmath0ko izar gehienak aldi motzekoak dira 20 mmag-ko anplitude fotometrikoarekin, eta beren oscilazioak erraz antzeman daitezke lurretik. Izan ere, @xmath0 izar hegalaria batzuk aurkitu dira istripuz, @xmath0 izar hegalaria ezagunak behatzeko erreferentziazko izar gisa hartzen direnean. eta Ez dut ulertzen. Hala ere, lurretik euren oscilazio espektro aberatsa argitzeko, kalitate handiko denbora-sorta luzeak behar dira, Lurreko luzera-lerroetan banatutako gune ezberdinetako behaketen bidez bakarrik lor daitezkeenak (e eta Ez dut ulertzen. beraz, @xmath0 scuti izar berri baten pulsazio espektroaren ezaugarria izateko, behaketa-ahalegin osagarriak behar dira. Paper honetan aurkezten ditugu @xmath0 izar berri baten behaketa fotometriko eta espektroskopikoen emaitzak zeinak azkenaldian Cygnus-en deskubritu baitira. @xcite-k @xmath3 sao 51294 izarra @xmath3 sao 51294, bd@xmath442 4207, hip 107557 izar berria @xmath0 scuti izar gisa aurkitu zuen 2007ko irailaren 27ko gauean A mot Ez, ez dut uste. Johnson 1 zenbakia. San Pedro Martir behatokiaren 5 metroko teleskopioa, Baja California, Mexiko Izarraren aldakortasuna nabaria zen, 2 ordu baino gutxiagotan ikusi bazen ere. 2007ko irailaren 30eko gauean 6 orduz behaketa fotometrikoak egin ziren Behatoki bereko 84 metroko teleskopioak bere aldagarritasuna baieztatu du. @xcite-k lortutako @xmath7 fotometria diferentziala eta estandarra, aurreko behaketetan erabili ziren ekipamendu berberak erabiliz. Bereziki, izarra 11 eta 19 azaro artean 11 orduz behatua izan zen. @xcitek frogak aurkitu ditu gutxienez bi frekuentzia gertutik, eta horrek argitasun kur-

ba bat azaltzen du. fotometria estandarra ondorengo da: (@xmath8, @xmath9@xmath10, @xmath11, @xmath12) = (8 Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Ehun eta berrogeita hamar Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak @xcitek Schlegel eta besteen mapa gorrietan izarraren gorritzea ere estimatu du. (1998) web orrialdeko kalkulagailuaren bidez (Nasa / IPACko datu base extragalaktikoak) ondoriozko koloreak ondorengoak dira: @xmath13, @xmath14@xmath15, @xmath16, @xmath17) = (Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak Eusko Jaurlaritzaren laguntzarekin egindako azpitoluak 002), eta @xmath18, @xmath19 mag, eta @xmath20 pc Hipparcosen katalogoak, berriz, 3 parallaxe ematen ditu. Ez dut ulertzen. 88 mas, eta handik 302 @xmath21 81 pc-ko distantziaren balioa kalkula daiteke. 7ko distantziako modulua da. 40 @xmath21 0 60 mag, @xmath22 0 balioa adierazten duena. 6 mg-ra baino gutxiago Beraz, hipparcos *Now, please write a short abstract for it in Basque.*

Erreferentziazko laburpena:

@xmath0 scuti pulsador berriaren behaketa kanpaina baten emaitzak aurkezten dira. Izarra behatu zen 2009ko abuztuaren 26tik irailaren 2ra, San Pedro Martir behatokian. 84 metroko teleskopioarekin (Mexiko) eta Teide behatokiarekin (80 metroko teleskopioa, Espainia) 53ko saio orokorra. Bi guneetatik 8 orduz datu erabilgarriak bildu ziren. Lau osziladoreen maiztasuna %99a baino handiagoa da. Emaitza hauek izarraren pulsazio natura multiperiodikoa baieztatzen dute, aurreko behaketen arabera, datu eskasekin. 2009an egindako behaketa espektroskopikoek bere espektro-mota eta argitasun-klasea atera eta bere biraketa-tasa kalkulatzeko ahalbidetu ziguten. ereduarekin alderaketa simple bat egiten da. izarrrak @xmath0 sct teknikak: izar fotometrikoak, izar espektroskopikoak, izar oszilatuzaileak, izar indibidualak, 97, beraz. Hogeita hamar Eusko Jaurlaritzaren laguntzarekin egin

Train 4Large ereduak sortutako laburpena:

@xmath0 scuti izar berri baten behaketa fotometriko eta espektroskopikoen emaitzak aurkezten dizkizuegu. Cygnus-en aurkitu zuten 2007ko irailaren 30ean, San Pedro Martir behatokiaren 5 metroko teleskopioarekin. Izarraren aldakortasuna nabarmena izan zen, 2 ordu baino gutxiagotan ikusi bazen ere. 6 orduko behaketa fotometrikoak egin ziren 8 metroko teleskopio batekin. Argiaren kurbak aldaketa nabarmenak erakusten ditu, bi aldiz handiagoa den anplitudea adierazten dutenak. Amplitudearen murrizketa nabaria ikusten da izarraren inguruko sakontasun desberdinetan. O, ez!

B.1.5 Pubmed

Testua:

First, read the Basque article below. Burke eta beste batzuek osasun mentalaren eta fisikoaren alorretan epe luzeko ondorio kaltegarriak aurreikusten dituzte. 2010; Ferguson eta besteak 2005; Odgers eta besteak 2008ko martxoa) Beraz, oso garrantzitsua da garunaren egituraren aldaketa txikiak egotea gizabanako hauen bizitzan. anomalia morfologikoen aurkikuntza koherenteak nahaste horien jatorria eta mekanismoak ulertzeko lagungarri izango lirarteke. CD eta ODDren irudien azterketa estrukturalak, haurtzaroan, bereziki baliagarriak izango lirarteke hiru arrazoi nagusirengatik. Lehenik eta behin, anomaliak bizitzan goizago agertzen badira, horrek adieraziko luke eragina duela eraginkortasun etiologiko batzuek goizago ere. eta, eragin genetikoak, eragin prenatalak eta/edo esperientzia goiztiarrak) Bigarrenik, garapen neuronala azkar garatzen delako haurtzaroan eta nerabe-

zaroan (Casey, 2013; Crone eta Dahl, 2012), garrantzitsua da garunaren egitura aldaketak aztertzea. Hirugarrenik, nerabezaroa baino lehen burmuinaren egitura desberdintasunak egiteari buruzko ikerketak gutxiago kezkatzen dira umearen portaera desberdintasunak burmuineko lesio traumatikoak izateko probabilitatea handitu dezakeela Haur eta nerabeen burmuineko egitura aldaketen aurreko ikerketen emaitzak ez dira koherenteak izan (Sterzer eta Stadler, 2009), baina hipotesi tentatibo batzuk onartzen dituzte. Aurreko ikerketetan, ume eta nerabeek, aldaketa nagusiak aurkitu zituzten amigdaletan, aurreko zuloetan eta lobulu tenporaletan. Hiru ikerketek amigdala- bolumena gutxitu dutela salatu dute CD sintomak dituzten nerabeetan normalean garatzen ari diren nerabeekin alderatuz (FairChild et al. 2011a; Huebner eta beste. 2008; Sterzer eta besteak, 2007b) portaera disruptiboen ikerketaren bigarren ardatza aurreko insularen egitura eta funtzioen aldaketak izan dira, prozesamenduan parte hartzen duena eta besteekiko kezka enpatikoarekin lotutako sare baten zati bat (2008; mutschler eta beste., hau garrantzitsua da, besteekiko kezka enpatikoaren gabeziak ohikoak direlako CD duten gazteengan (Frick and White, 2008; lahey et al. 1999ko martxoak Lau ikerketek insularen bolumena txikiagoa dela salatu dute jokabide-arazoak dituzten nerabeetan, kontrol osasuntsuetan baino (fairchild eta besteak). 2011b; Huebner eta beste. 2008; Sterzer eta besteak 2007a), baina jokabide-arazoak eta psikopata-ezaugarriak dituzten nerabeei egindako ikerketa batek ez zuen alderik aurkitu aurreko insularen aldean osasuntsuak diren kontrol taldeekin alderatuta (de brito et al., 2007a). 2009ko martxoak) Emozioen erregulazioan eta sarien eta zigorren prozesamenduan ere paper garrantzitsua jokatzen du (O'Doherty eta besteak). 2001ean (Eusko Jaurlaritza - Eusko Jaurlaritza) Hainbat ikerketek prozesu horien gabeziak nerabeen portaera-arazoekin lotzen dituzte (Fairchild eta besteak). 2013a; Huebner eta beste., 2008) eta nortasunaren nahaste antisoziala (raine et al. De Oliveira-Souza et al. (2011), eta psikopatia helduen artean (de Oliveira-Souza et al., 2011). Gainera, portaera disruptiboa duten haur eta nerabeen azterketek materia grisaren bolumen murriztua (gmv) agerian utzi dute aldizkarietan (Huebner eta besteak). 2004), nahiz eta hirugarren ikerketak bolumen handiagoak aurkitu dituen CD arazoak dituzten mutilen lagin komunitario batean (de brito et al. 2009ko martxoak) Pertsonalitate-arazoak dituzten heldu presoetan ere, lobulu tenporalak handiagoak izatea antzeman da (Dolan et al., 2008), 2002) eta preso diren psikopata held *Now, please write a short abstract for it in Basque.*

Erreferentziazko laburpena:

Burmuineko egitura-aldaketak zerikusia duten jakiteko garrantzitsua da haurren bizitzaren hasieran arazoak sortzen dituzten. 1111 911 urteko emakumezkoen datuak jaso ziren (58 neska eta 53 mutil), 43 oposizioaren aurka jartzen den nahasmendu eta/edo jokabide-nahasmenduaren irizpideak betetzen zituztenak, eta 68 kontrol osasuntsu. Voxelen oinarritutako morfometria erabili zen jokabide neurrien loturak aztertzeke materia grisaren bolumenekin. Aurreko ikerketetan ez bezala, materia grisaren bolumena aldatzeak ez du loturarik jokabide-arazo disruptiboaren diagnostikoarekin. 05 FWren zuzenketarekin Hala ere, portaera-arazoaren sintoma kopuruaren asoziazio ez-lineala materia grisaren bolumenarekin ezkerreko sulkus temporal gorenean adierazgarria izan da lagin osoan ($p < 0,05$, hego hemisferioan ($p < 0,001$ zuzendu gabe) Cd sintomak eta materia grisaren bolumenaren arteko lotura indartsuagoa ageri da ezkerreko sulcus

Train 4Large ereduak sortutako laburpena:

Burmuinaren egitura aldaketa txikiak izateak osasun mentalaren eta fisikoaren alorrean epe luzeko ondorio kaltegarriak aurreikusten ditu. Beraz, oso garrantzitsua da

Datu multzoa	ROUGE-F1	ROUGE-F2	ROUGE-L	ROUGE-TOK
BillSum	0.37	0.21	0.33	0.439
Wikihow	0.55	0.4	0.55	0.58
Reddit TIFU	0.22	0.0	0.22	0.48
Arxiv	0.28	0.04	0.17	0.4
Pubmed	0.19	0.05	0.1	0.36
AESLC	0.0	0.0	0.0	0.0

B.1 Taula: Adibideetako testuen ROUGE balioak.

burmuinaren egitura aldaketa txikiak izatea gizabanako hauen bizitzan. Haur eta nerabeen burmuineko egitura aldaketan aurreko ikerketen emaitzak ez dira koherenteak izan baina hipotesi tentatibo batzuk onartzen dira. Aurreko ikerketen arabera, haurtzaroan eta nerabezaroan burmuineko egitura aldaketa nabarmenak daude. Azterketa honen helburua haurren eta nerabeen burmuineko egitura aldaketak aztertzea izan da. material/metodoak azterketa honetan parte hartu zuten 18 haur eta nerabe zeuden (CD sintomak dituztenak) eta 20 pertsona osasuntsu (CD gabekoak). Estructura-anomaliak aztertu ziren materia grisaren bolumena (gmv), amigdalaren bolumena eta lobulu tenporalen bolumena erabiliz. Gmv eta amigdalaren bolumena handitu egin dira CD duten ne

B.1.6 AESLC

Testua:

Dakizuen bezala, hau inoiz gertatu ez den gertaera da Enronen historian. Gure langile programak moldatu behar ditugu gure enpresaren premia zuzenetara trantsizio aldi honetan. Bereziki zaila da honako programa hauen berehalako etetea iragartzea: ClickAtHome - Enronek bertan behera utzi du ClickAtHome programa Programak ez ditu parte-hartzaile berriak onartuko, ez ordenagailuen eskaerarik, ez Interneteko eskaerarik. 2001eko azaroaren 20a baino lehen bidali eta baieztatutako eskaerak onartuko dira. Enronek Interneteko zerbitzuak ere bertan behera utziko ditu. 2002ko urtarrilaren 1etik aurrera, Interneteko zerbitzu diruz lagunduta dauden langileak zerbitzuaren tarifa komert Opariak batzen eta boluntarioak bultzatzeko programa (VIP) - Enronen opariak batzen eta VIP programak Langile guztien premia larriak kontuan hartuta, egoera zail honetan, egokia da dolar-diru truke anto Opari egokiak eta VIP eskaerak azaroaren 20a baino lehenago jasoko dira. Sentitzen dugu aldaketa hauek egin behar izana. Transizio aldi, operazio gastuak murrizteko bideak bilatzen jarraitu behar dugu. *How would you rephrase that briefly in Basque?*

Erreferentziazko laburpena:

Programa aldaketak

Train 4Large ereduak sortutako laburpena:

Eusko Jaurlaritzaren laguntzarekin egindako azpigituluak

B.1.7 ROUGE balioak

Laburpenen ROUGE balioak ??aulan ikus daitezke.

B.2 Elikagaien testuak

B.2.1 Eli0

Testua:

First, read the Basque article below. Janari edo edari bakoitza, hau da elikagai bakoitza, substantzia edo mantenugai jakin batzuek osaturik dago. Mantenugai bakoitzak zeregin jakin bat du gure gorputzaren osaketan eta funtzionamenduan. Horregatik, janari batek beste batek baino hobeto erantzun diezaioke gorputzaren behar jakin bati. Betetzen duten funtzioaren arabera, elikagaiak hiru multzotan sailkatzen dira. Gorputzak, hazten ari denean, zelula eta ehun gehiago sortzen lagunduko duten mantenugaiak beharko ditu. Eginkizun hori, zelula berriak sortzea, alegia, betetzen lagunduko duten mantenugaiak dituzten elikagaiei eraikitzaile deritze. Bestalde, lasterka egiteko edo hotz handia dagoenean gorputza berotzeko, energia emango dion erregai bat beharko du gorputzak. Eginkizun hori betetzen duten mantenugaiak dituzten elikagaiei energetiko deritze. Azkenik, hirugarren taldea ezinbestekoa da organismoak funtzionamendu zuzena izan dezan. Talde honetako mantenugaiak dituzten elikagaiei erregulatzailerik deritze. *Now, please write a short abstract for it in Basque.*

Train 4 Large ereduak sortutako laburpena:

Janari edo edari bakoitza, hau da elikagai bakoitza, substantzia edo mantenugai jakin batzuek osaturik dago. Mantenugai bakoitzak zeregin jakin bat du gure gorputzaren osaketan eta funtzionamenduan. Horregatik, janari batek beste batek baino hobeto erantzun diezaioke gorputzaren behar jakin bati. Betetzen duten funtzioaren arabera, elikagaiak hiru multzotan sailkatzen dira. Gorputzak, hazten ari denean, zelula eta ehun gehiago sortzen lagunduko duten mantenugaiak beharko ditu. Bestalde, lasterka egiteko edo hotz handia dagoenean gorputza berotzeko, energia emango dion erregai bat beharko du. Azkenik, hirugarren taldea ezinbestekoa da organismoak funtzionamendu zuzena izan dezan. Talde honetako mantenugaiak dituzten elikagaiei erregulatzailerik deritze.

B.2.2 Eli1

Testua:

First, read the Basque article below. Gehiago ala gutxiago, lehenago ala beranduago, hazi egingo zara. Hazteko gorputzak duen bidea zelulak eta ehun berriak sortzea da (ehunak, luzatu eta sendotzeko, giharrek handitzeko...). Baina hazitakoan ere, gorputzak izaten du zelula eta ehun berrien beharrik; esate baterako, ilea eta azazkalak luzatzeko edo zauri baten ondorioz hondatutako ehunak berreraikitzeke. Gainera, suntsitutako zelula zaharrak ordezkatzeke berriak osatu behar ditu. Horretarako, gure gorputzak proteinak dituzten elikagaiak behar ditu, proteinak baitira mantenugai eraikitzaile nagusiak. Etxe bat eraikitzeke adreiluak behar diren bezala, proteinak behar ditugu gure gorputza eraikitzeke eta luzatzeko. Bestalde, proteina gehiegi duten dietek gaixotasunak eragin ditzakete. Horixe gertatzen ari da zenbait herri aurreratutan. Proteinak dituzten jakiak landare-jatorria nahiz animalia-jatorria izan dezakete. Babarrunak, dilistak, babak, soja, txitxirioak eta fruitu lehorrak, adibidez, landare-jatorrikoak dira, eta, esnea, esnekiak, arrautzak, haragia, arraina eta urdaia, berriz, animalia-jatorrikoak. *Now, please write a short abstract for it in Basque.*

Train 4 Large ereduak sortutako laburpena:

Gorputzak zelula eta ehun berriak sortzen ditu (ehunak, luzatu eta sendotzeko, giharrek handitzeko...). Baina hazitakoan ere, gorputzak izaten du zelula eta ehun berrien beharrik; esate baterako, ilea eta azazkalak luzatzeko edo zauri baten ondorioz hondatutako ehunak berreraikitzeke. Gainera, suntsitutako zelula zaharrak ordezkatzeko berriak osatu behar ditu. Horretarako, gure gorputzak proteinak dituzten elikagaiak behar ditu, proteinak baitira mantenugai eraikitzaile nagusiak. Etxe bat eraikitzeke adreiluak behar diren bezala, proteinak behar ditugu gure gorputza eraikitzeke eta luzatzeko. Bestalde, proteina gehiegi duten dietek gaixotasunak eragin ditzakete. Horixe gertatzen ari da zenbait herri aurre-ratutan. Proteinak dituzten jakiak landare-jatorria nahiz animalia-jatorria izan dezakete. Babarrunak, dilistak, babak, soja, txitxirioak eta fruitu lehorrak,

B.2.3 Eli2

Testua:

First, read the Basque article below. Elikagai energetikoeke bi eratako mantenugaiak izan ditzakete: karbohidratoak eta koipeak. Karbohidratoak. Giharrek lan egiteke beharrezko energia karbohidratoetatik datorkigu, hein handian. Janari askoren bidez lor ditzakegu karbohidratoak, eta karbohidratoetatik lortutako mantenugaiak gibelean pilatzen dira eta, gorputzak eskatzen duenean, odolera igarotzen dira. Karbohidratoak bi taldetan bana daitzeko: almidoia eta azukreak, biak landare-jatorrikoak. Landareek eguzkiaren energia erabiliz sortzen dituzte. Patata, fruitu lehorrak, azukrea, arroza, garia, artoa, garagarra, irinez egindako janariak (ogia, gailetak, pastak...) dira karbohidrato gehien duten jakiak. Kultura bakoitzak iturri jakin batetik lortzen du almidoia; hala nola, ogiaren bidez Mediterraneo aldean, arrozaren bidez Ekialdean, patataren bidez Iparraldean, eta artoaren eta gaztainaren bidez antzinako Euskal Herrian. Energia lortzeko berehala erabiltzen ez diren karbohidratoak, glukogeno bihurtuta pilatzen dira gibelean eta giharretan. Glukogeno hori gorputzak behar duena baino gehiago denean, koipe bihurtzen da eta azalpean pilatzen da gizentasuna eraginez. Koipeak. Koipeak dira, proportzioan, energia gehien ematen diguten mantenugaiak; hau da, gramo bakoitzeko energia gehien sortzen dutenak. Karbohidratoekin gertatzen den bezala, gure jarduerak eskatzen duena baino koipe gehiago jaten badugu, gorputzak gorde egiten du, erreserba moduan; hau da, gizendu egiten gara. Koipe batzuk landare-jatorriko elikagaietan daude: oliba, kakahuete, ekilore eta arto-olioetan, esaterako; fruitu lehorretan (almendrak, intxaurrak, hurrak...); margarinan, etab. Beste batzuk, berriz, animalia-jatorriko jakietan: zerri, behi, ardi eta bestelakoan gantzetan; esnean eta esnekietan, etab. Adituek diotenez, gorputzarentzat animalia-jatorriko koipeak baino osasuntsuagoak dira landare-jatorrikoak. *Now, please write a short abstract for it in Basque.*

Train 4 Large ereduak sortutako laburpena:

Elikagai energetikoeke bi eratako mantenugaiak izan ditzakete: karbohidratoak eta koipeak. Giharrek lan egiteke beharrezko energia karbohidratoetatik datorkigu, hein handian. Janari askoren bidez lor ditzakegu karbohidratoak, eta karbohidratoetatik lortutako mantenugaiak gibelean pilatzen dira eta, gorputzak eskatzen duenean, odolera igarotzen dira. Koipe batzuk landare-jatorriko elikagaietan daude: oliba, kakahuete, ekilore eta arto-olioetan, esaterako; fruitu lehorretan (almendrak, intxaurrak, hurrak...); margarinan, etab. Adituek diotenez, gorputzarentzat animalia-jatorriko koipeak baino osasuntsuagoak dira landare-jatorriko koipeak.

B.2.4 Eli3

Testua:

First, read the Basque article below. Elikagai erregulatuak dituzten mantenugaiak ezinbestekoak dira organismoak funtzionamendu zuzena izan dezan. Mantenugai garrantzitsuenak bitaminak, gatz mineralak eta ura dira. Oro har, kopuru txikitik behar izaten ditu gorputzak ura izan ezik, ura ugari hartu behar izaten baitugu. Bitaminak. Izaki bizidun batek bere funtzionamendu arrunterako behar-beharrezkoak dituen mantenugaiak dira bitaminak. Bizidunen oinarriko funtzioetan parte hartzen dute; adibidez, organismoaren mineral-orekan, eta zenbait egitura eta ehunen kontserbazioan (hezurrak, hartzak, odola, azala, etab.) .Bitaminen urritasunak abitaminosi izeneko gaixotasun larria eragin dezake, eta bitamina gehiegi kontsumitzeak, berriz, hiperbitaminosi izeneko eritasuna. Bitamina hainbat taldeetan sailkatzen dira eta talde bakoitzari letra bat dagokio: A, B, C, D, E eta K taldeak bereizten dira. Ezinbestean, elikagaien bidez lortzen dira bitaminak (D taldeko bitaminak eguzki-argiaren bidez ere jasotzen ditugu).Gatz mineralak.Gure gorputzeko zelulentzat eta ehulentzat behar-beharrezko osagaiak dira mineralak. Horregatik, gure elikadura osoa izan dadin, mineralak ere hartu beharko ditugu. Jaki askok dituzte mineralak, askotan gatz mineralen moduan. Horregatik, gehienetan ez da beharrezkoa izaten aparteko elikagai edo gauza gisa hartzea.Gatz mineralen artean ezagunena sukaldian erabiltzen den gatzak dugu. Baina gure gorputzak behar duen mineral bakarra ez da hori, beste hainbat ere beharrezkoak baititu, esate baterako, kaltzioa, fosforo eta burdina bezalako elementuak. Elementu horiek hezurren eta odoloko hemoglobinarekin osagarriak dira, eta gure gorputzarentzat nahitaezkoak, beraz. Beste elementu batzuk ere, barazkien bitartez geureganatzen ditugun magnesioa eta potasioa, esaterako, gure gorputzerako behar-beharrezkoak ditugu.Ura.Ura izaki bizidunen organismoaren funtzionamendurako funtsezko osagaia da. Besteak beste, hainbat substantzia uretan disolbatzen dira eta, horrela, hobeto asimilatzen ditu gorputzak. Bestalde, toxinak eta bestelako hondakinak kanporatzeko ere ezinbestekoa da, gurearen eta izerdiaren osagai nagusia baita.Gure gorputza bezala, janari asko urarekin osatuak daude, nagusiki: barazkiek eta frutek, esate baterako, ur kantitate handia dute. Behar dugun ura, zuzenean ura edanez ez ezik, barazkiak eta frutak janez ere eskura dezakegu. *Now, please write a short abstract for it in Basque.*

Train 4Large ereduak sortutako laburpena:

Elikagai erregulatuak dituzten mantenugaiak ezinbestekoak dira organismoak funtzionamendu zuzena izan dezan. Mantenugai garrantzitsuenak bitaminak, gatz mineralak eta ura dira. Oro har, kopuru txikitik behar izaten ditu gorputzak ura izan ezik, ura ugari hartu behar izaten baitu.

Bibliografia

- [1] Unai Atutxa. Laburpen-gaitasunaren garapena eta eskolako laburpen-testuen prozesamendua. Ikusi [viii](#), [2](#), [26](#), and [27](#) orrialdeak.
- [2] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020. Ikusi [ix](#), [20](#), and [42](#) orrialdeak.
- [3] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022. Ikusi [2](#), [22](#), and [33](#) orrialdeak.
- [4] Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2023. Ikusi [11](#), [12](#), and [17](#) orrialdeak.
- [5] Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. Ikusi [12](#) orrialdea.
- [6] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010. Ikusi [12](#) orrialdea.
- [7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. Ikusi [12](#) orrialdea.
- [8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. Ikusi [12](#) orrialdea.
- [9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. Ikusi [13](#) orrialdea.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. Ikusi [14](#), [17](#) orrialdeak.
- [11] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020. Ikusi [15](#) orrialdea.
- [12] Benyamin Ghojogh and Ali Ghodsi. Attention mechanism, transformers, bert, and gpt: Tutorial and survey. 2020. Ikusi [16](#) orrialdea.
- [13] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019. Ikusi [16](#), [19](#) orrialdeak.
- [14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. Ikusi [16](#) orrialdea.
- [15] Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*, 2023. Ikusi [17](#) orrialdea.

- [16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. Ikusi [17](#), [18](#) orrialdeak.
- [17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. Ikusi [17](#), [20](#) orrialdeak.
- [18] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. Ikusi [18](#) orrialdea.
- [19] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021. Ikusi [19](#) orrialdea.
- [20] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958. Ikusi [19](#) orrialdea.
- [21] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017. Ikusi [20](#) orrialdea.
- [22] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. Ikusi [22](#) orrialdea.
- [23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. Ikusi [23](#), [24](#) orrialdeak.
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. Ikusi [25](#) orrialdea.
- [25] William C Mann and Sandra A Thompson. *Rhetorical structure theory: Description and construction of text structures*. Springer, 1987. Ikusi [26](#) orrialdea.
- [26] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. Ikusi [32](#) orrialdea.
- [27] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 2021. Ikusi [32](#) orrialdea.
- [28] Maja Popović. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015. Ikusi [32](#) orrialdea.
- [29] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021. Ikusi [36](#) orrialdea.
- [30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. Ikusi [44](#) orrialdea.
- [31] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. Ikusi [44](#) orrialdea.
- [32] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of trans-

- formers. In *International Conference on machine learning*, pages 5958–5968. PMLR, 2020. Ikusi [56](#) orrialdea.
- [33] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. Ikusi [67](#) orrialdea.