Laburpengintza automatikoa aztertzen

Xabier Irastorza Urbieta

Abstract

Artikulu hontan laburpengintza automatikoan murgiltzeko jarraitu den proiektu baten nondik norakoak azaltzen dira. Proiektu horretan bi laburtzaile automatiko sinple garatu dira, batek erauzketa bidez egiten ditu laburpenak eta bigarrenak abstrakzio bidez. Sistema sinple horiek nola sortu diren eta nola hobetu daitezkeen ere irakurri daiteke artikuluan.

1 Sarrera

Laburpengintza automatikoaren helburua gizakien testuak laburtzeko gaitasuna makinetara eramatea da eta ahalko balitz, hura hobetzea. Gero eta informazio uholde handiagoak dituen mundu honetan indarra hartzen ari da informatikaren adar hau, erakunde eta norbanakoentzat oso interesgarria izan daitekeelako, informazio esanguratsua gainerakotik bereizi eta denbora aurrezteko. Alabaina, oraindik ere asko du hobetzeko laburpengintza automatikoak, haren garapena duela hamarkada gutxi hasi baitzen arloaren konplexutasunak eraginda eta orain arte sortu diren sistema guztiak gizakien gaitasunetik oso urruti daude.[1]

Proiektu honetan laburpengintza automatikoko tenkina ezberdinak aztertu dira, haiek erabiliz pare bat laburtzaile automatiko sinple sortzeko. Laburpengintza automatikoaren historian begiratuta, bi bide jarraitu dira laburtzaile automatikoak sortzerako orduan: erauzketa eta abstrakzioa. Proiektu honetan bi bideak ibili dira eta bakoitzetik sistema bana sortu da, horrela bien arteko alderaketak egiteko.

2 Sistemen arkitektura

Esan bezala, bi bide jarraitu dira eta bakoitzean oinarrituta sistema bat sortu. Sistema hauen arkitektura zeharo desberdina da.

2.1 Erauzketa bidezko sistema

Erauzketa bidezko sistema garatzeko SFNet arkitektura baliatu da, laburpengintza automatikorako lehenago ere erabilia.[2] Arkitekturaren muina bi informazio-iturri erabiltzea da, batetik esaldien informazioa aztertuko duen LSTM sare errekurrentea eta bestetik, eskuz diseinatutako esaldien ezaugarri erauzle bat. Arkitektura honi zenbait aldaketa egin zaizkio eta hainbat ekarpen ere bai, jarraian azalduko den bezala.

Erauzketa bidezko laburpengintzak zertan datzan labur azaltzeko, jatorrizko testuko esaldi edo pasarte garrantzitsuak zeintzuk diren atzematean datza, ondoren zati horiek aldaketarik egin gabe elkartzeko eta laburpena osatzeko. Hortaz, berez, laburpena baino gehiago pasarte garrantzitsuenen bilduma bat osatzea da erauzketa bidezko laburpenginztaren helburua.

2.1.1 LSTM sarea

Ereduak esaldiak jasotzen dituenez sarrera bezala, LSTMari hitzak sartuko zaizkio urrats bakoitzean sarrera bezala. Aurreko urratsetan sartu zaizkion hitzak gogoratzeko sare errekurrenteen ezaugarri diren egoera-bektoreak erabiliko ditu eta hauetako bakoitza gorde egingo da. Egoera-bektore hauen azpilaginketa egingo da maximoa erabiliz (*max-pooling*) eta hau hartuko da LSTMaren esaldiaren errepresentazio gisa hurrengo pausoetarako.

2.1.2 Ezaugarri erauzlea

Ezaugarri erauzleak esaldiak jasoko ditu sarrera gisa eta esaldi bakoitzeko bere ezaugarriak bilduko dituen tentsorea emango du irteeratzat. Ezaugarri erauzleak hiru motako ezaugarriak hartuko ditu kontuan:

1. Ohiko ezaugarriak: digitu kopurua esaldiko, parentesi kopurua, esaldiaren kokapena, TFIDF indizea (esaldiaren barnean agertzen diren hitzen maiztasuna dokumentuan

agertzen direnen maiztasunarekiko arraroa edo arrunta den jakiteko), etab. Hauek informatikaren literatuan ohikotasunez topatu daitezke.[2]

2. Hitz mota ezberdinen maiztasunak: sintaxia ardatz hartuta, aditz, izen, artikulu eta bestelako hitz motak zein maiztasunekin agertzen diren esaldian neurtuko da. Ezaugarri honek aditz kopuru ezohikoko eta izen kopuru ezohikoko esaldiak identifikatzen lagunduko du, izen ugariko esaldiak laburpenean agertu beharrekoak izan ohi direlako, adibidez. Spacy tresna erabiliko da horretarako.

3. Esaldi arteko informazioa gordetzen arduratuko direnak: sistemak esaldika lan egitearen arazo nagusia beste esaldien informazioa eskuragarri ez izatea da. Oso zaila da testuko gainerako esaldien informazioa eduki gabe esaldi bat laburpenean egon beharko lukeen edo ez erabakitzea. Lan hori errazteko, esaldi bakoitzak bere esaldian agertzen diren hitzak aurreko esaldietan zenbait aipatu diren jakingo du. Adibidez, "Zaraitzu"izen berezia aurreko esaldietan ez bada inoiz agertu eta oraingoan bi aldiz, ezaugarri hauek hitz honen maiztasun diferentzia bikoa dela jasoko dute, ondorengo formularen arabera:

$$hitzMaiz = \frac{Un}{Au + 1} \quad (1)$$

Non Un uneko esaldiko hitzaren agerpen kopurua den (kasu honetan bi) eta Au aurreko esaldietan izan duen agerpen kopurua (kasu honetan zero). Zerorekiko zatiketa saihesteko gehi bat egiten da izendatzailean. $hitzMaiz$ zenbat eta txikiagoa izan, orduan eta gehiagotan agertu da hitza aurreko esaldietan eta beraz, uneko esaldian informazio errepikakorra egoteko aukera gehiago egongo da. Sistemak soilik ez du jasoko ezaugarri tentsorean zein hitz agertu den zehazki gehiago edo gutxiago aurreko esaldietan, baina jakingo du hitz batzuk asko agertu direla edo gutxi. Horrek informazio minimoa emango dio aurreko esaldien inguruan.

Ezaugarri guztiak entrenamendu fasean aurrez konputatuak izan dira, horrela etengabe ez egoteko kalkulatzen ezaugarri tentsore berdinak. Test fasean ereduari deitu baino lehen egiten da konputazioa.

2.1.3 Eredua

Ereduari berari dagokionean, behin ezaugarri tentsoreak eta LSTMaren irteera-tentsoreak izanda bakoitza geruza dentso batera igarotzen ditu eta horko irteerak kateatu egiten ditu ondoren beste geruza dentso batera sartzeko. Geruza dentso horretan diluzioa aplikatzen du eta azken geruza dentso baten bidez irteerak lortzen ditu. Azken geruzan sigmoide aktibazio-funtzioa aplikatzen denez, irteerak 0 eta 1 artekoak izango dira, laburpenean ez sartzeko ziurtasun osoa 0 izanda eta laburpenean sartzeko ziurtasun osoa 1. Geruza dentso bakoitzaren ostean, azkenekoan izan ezik, sorta normalizazioa eta ReLU aktibazio-funtzioa aplikatzen dira.

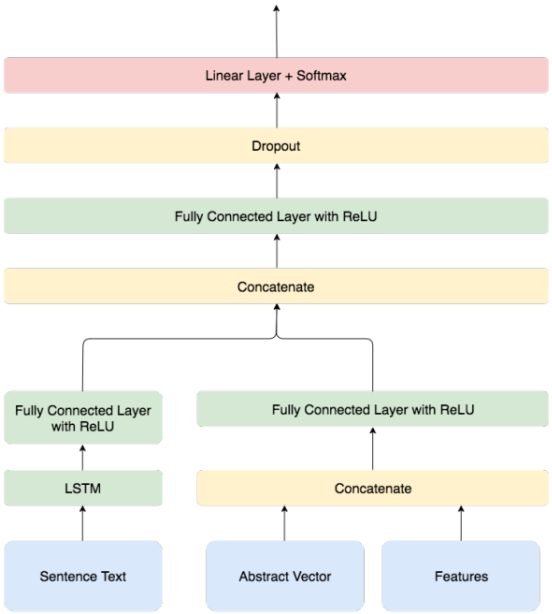


Figure 1: SAFNet (SFNet-en aldaera) arkitekturaren ohiko egitura.[2] Erauzketa bidezko sistemak ez du jasotzen laburpenaren bektorea sarrera gisa, gainerakoan antzekoa da.

Ereduak bere irteerak eman ostean, laburpena sortzeko, puntuazio altueneko esaldiak aukeratzen dira eta horiek denak elkartu.

2.2 Abstrakzio bidezko sistema

Abstrakzio bidezko sistema bat garatzeko, Google enpresak entrenatutako T5 eredua[3] erabili da proiektu honetan. Abstrakzio bidezko laburpegintzan sistemari testuaren gaineko ulermen sakona, benetako esanahia edo edukia erauzteko gaitasuna, esakatzen zaio eta horrek berarekin batera konputazio gaitasun handiagoaren beharra dakar. Hori izan da laburpengintza mota hau historian zehar, azken

hamarkadan izan ezik, nahiko alboratuta egotearen arrazoietako bat. Konputazio gaitasun murritzarekin egin denez proiektu hau, *T5-small* izeneko eredua berrentrenatu da, jatorrizko T5a baino gaitasun txikiagokoa.

3 Erabilitako datu-basea

Sistemak entrenatzeko *multi-news* izeneko datubasea erabili da. Datu-base honek *newser.com* ingelesezko egunkariaren webguneko artikuluak ditu bere baitan, haietako bakoitzari pertsona batek eskuz idatzitako laburpenekin batera. Datu-baseak guztira 56.216 adibide ditu.

Erauzketa bidez dabilen sistemari datu-base berezi bat sortu zaio, ereduak sarrera bezala testuko esaldiak jasotzen dituelako. Literaturan beste batzuetan egin den bezala,[2] ROUGE-L indizea atera zaie testuen esaldi guztiei eta ondoren, testu bakoitzeko, ROUGE-L indize onenak zituzten esaldien %20a eta indize kaskarrenak zituzten esaldien %20a hautatu dira entrenamenduan adibide izateko. ROUGE-L indize altuko esaldiak informazio handikoak izan ohi dira eta kontrakoa indize baxua dutenak, hori dela eta, ereduak entrenamenduan laburpenean egon beharko luketen eta egon beharko ez luketen esaldiak izan zitzan, hautaketa hori egin da. Esaldien gainerako %60a ez da hautatu, zalantzagarriagoak izan daitezkeelako.

4 Emaitzak

Jarraian sistema bakoitzak eman dituen emaitzak aztertuko dira.

4.1 Galerak

Erauzketa bidezko sistemaren entrenamendua 34 arotan egin da, azpiko taulan ageri diren hiperparametroekin. Gainera, ikaskuntza tasaren planifikatzaile bat erabili da, hiru aldiz segidan garapen galera okertuz gero ikaskuntza tasa hamarren batera jaisteko. Gelditze goiztiarra ere erabili da.

HiperP	bs	hd	nl	cd	lM	hT	dP
Balioa	64	128	3	256	0.2	3	0.1

Non bs sorta-tamaina, hd dimentsio-ezkutu tamaina, nl geruza kopurua, cd kateatze geruzaren tamaina, lM luzera maximoa, hT historia tamaina, eta dP diluzio probabilitatea diren. Lehen ikaskuntza-tasa 0.01 eta pisuen gainbehera 1e-5 izan da. 2. irudian grafikoki ikus daitezke galerak eta azpiko taulan aro jakin batzuetakoak zehatz.

Aroa	0	15	25	33
Ent.	0.095	0.059	0.054	0.053
Gar.	0.073	0.058	0.056	0.057
Asm-tas.	97.18	97.74	97.85	97.82
Ik. tas.	0.01	0.001	0.0001	1e-6

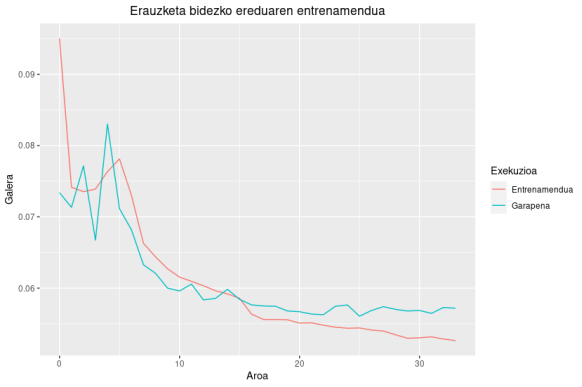


Figure 2: Erauzketa bidezko sistemaren entrenamendua.

Irudian eta taulan ikusten den bezala, galerak behera egin du bai entrenamenduan eta baita ere garapenean. Ikaskuntza tasaren planifikatzaileak lagundu egin du horretan nabarmen. Asmatze-tasari dagokionean, zeinak asmatutzat jotzen duen, adibidez, labela 1 denean iragarpena 0.5 baino handiagoa bada, ez da aldatu nabarmen aroen artean. Horrek esan nahi du lehen arotik jada ereduak nahikoa ondo bereizten zituela laburpenean egon beharreko esaldiak besteetatik. Hori ordea bere ikaskuntzako lehen pausoa baino ez da, izan ere, laburpenean egon beharko luketen esaldi horien artetik garrantzitsuenak zeintzuk diren ikastea da helburu nagusia, garrantzitsuenak ipintzeko laburpenean. Ez dirudi gainegokitzapenik dagoenik. Gorde den eredua 25. arokoa izan da. Testeko galera 0.589 izan da.

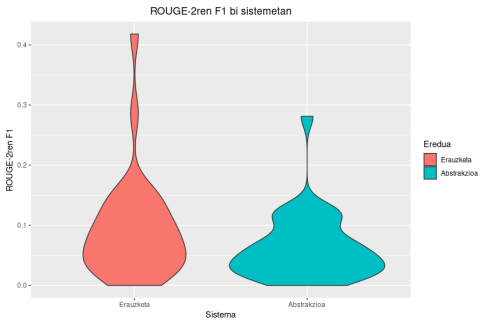
Abstrakzio bidezko sistema entrenatzeko 25 aro erabili dira eta T5 ereduarekin erabili ohi den entrenamendu prozedura.

4.2 ROUGE-2 indizeak

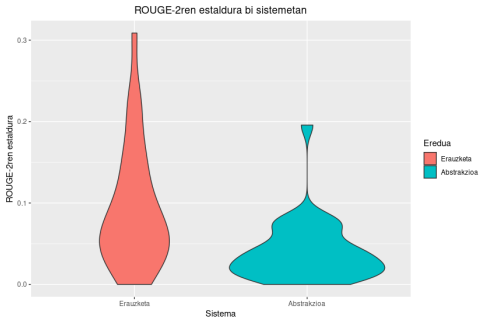
Ereduek sortzen dituzten laburpenen kalitatea neurtzeko modu errazena, automatikoa delako, ROUGE-2 [4] indizea kalkulatzea da. Honek, berezko laburpenean eta sortutako laburpenean agertzen diren bigrama kopurua neurtzen du. Bigrama kopurua zenbat eta handiagoa izan, orduan eta egokiagotzat jotzen da laburpena. Azpiko taulako datuak 30 laburpen ezberdinetan ROUGE-2 indizeak kalkulatzearen emaitza dira, laburpen horietako indizeen batez bestekoa. Biolinak sortzeko ere datu

berdinak erabili dira.

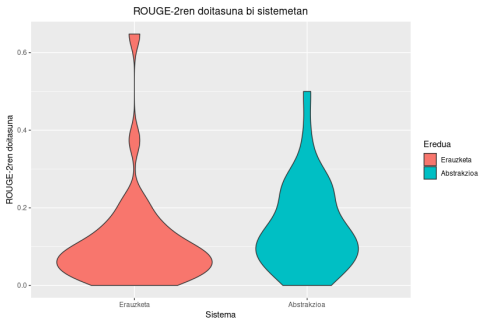
Sistema	F1	Estaldura	Doitasuna
Erauzketa	0.093	0.092	0.110
Abstrakzioa	0.063	0.04	0.150



(a) F1



(b) Estaldura



(c) Doitasuna

Figure 3: Erauzketa bidezko eta abstrakzio bidezko ereduen arteko alderaketa, ROUGE-2 indizea erabiliz egina. Gorriz erauzketa bidezko sistema, urdinez abstrakzio bidezkoa.

Taulan eta irudian (3. irudia) ikusi daitekeen bezala, ROUGE-2 indizeak ez dira oso altuak. Erabili dugun datu-basearekin ez, baina beste batzuekin entrenamendua egitean 0.15eko ROUGE-2 indizeak lortu dira, erraz. Beraz, badago zer hobetua. Hala ere, erabilitako datu-basearen urrezko estandarreko laburpenen eragina izaten du indize honek, ondorioetan zabalago azaldu den moduan.

Biolinek (3. irudia) 30 laginen (laburpenen) ROUGE-2 indizeen banaketa adierazten dute, horrela, lagin dentsitate handiena zein balio-tartetan kokatzen den ikusi daiteke. Biolinetan ikusi daiteke doitasunean izan ezik, beste bi grafikoetan erauzketa sistemak emaitza hobeak eman ohi dituela gehienetan. Erauzketa bidezko sistemak gainera, hiru grafikoetan, muturreko balio gehiago sortzeko joera du, ziur aski izen berezi konposatu (erakunde izenak, alderdien izenak, etab.) ugari agertzen diren artikuluetan izen horiek maiztasun handiarekin kopiatzen dituelako. F1 indizea izan ohi da sistemen gaitasuna neurtzerakoan gehien begiratzen dena eta hor ikusi daiteke laginen dentsitate esanguratsua duela erauzketa bidezko sistemak 0.0 eta 0.2 ROUGE-2 indizeen artean, abstrakzio bidezko sistemak aldiz 0.15 arte zabaltzen du dentsitate handiko eremu hori eta lagin gehien 0.0 eta 0.1 artean biltzen dira. Gainera, muturreko balioak erauzeta bidezko sistemak urrunago kokatzen ditu.

Hori horrela, esan daiteke erauzketa bidezko sistemaren emaitzak hobexeagoak direla abstrakzio bidezkoarenak baino, baina hala ere, 30 laburpen dituen lagin txiki bat aztertu denez, ezin da ezer ziurtatu. Biolinek erakutsi digute batez bestekoetan argiago ikusten ziren desberdintasunak tentuz hartu behar direla, erauzketa bidezko sistemak muturreko balio gehiago ematen dituenez, bere batez bestekoak igotzen dituelako.

4.3 Eskuzko azterketa

Ausaz bost laburpen hautatu dira testeko partiziotik eta horien ezaugarriak eskuz aztertu. Laburpen horiek *Material gehigarria* atalean eskuragarri daude. Bosgarren testutik hasita, zeinaren laburpenek ROUGE-2 indize onenak lortu dituzten, ikusi daiteke bai erauzketa bidezko eta baita ere abstrakzio bidezko sistemak laburpen dexente bat sortu dutela jatorrizko testutik. Puntu aipagarriak honakoak dira:

1. Bi sistemek berezko laburpena baino laburpen motzago bat egin dute.

2. Bi sistemek jatorrizko testuan dauden zitazioak ez dituzte ipini eta haiek esandakoa modu pasiboan adierazi dute, laburpenetan egin ohi den bezala.

3. Abstrakzio bidezko sistemak esaldi berriak sortu ditu, baina hainbat akats gramatikal ere egin ditu. Testua uler daitekeen arren, akatsek zaildu egiten dute. Esaldi berri horietan erakunde izenak ere laburtu ditu, adibidez, *U.S Coast Guard* izena *coast guard* bihurtu du.

Lehen testuaren laburpenek ROUGE-2 indize arruntak lortu dituzte batez bestekoekin alderatuta. Ikusi daiteke nola abstrakzio bidezko metodoaren laburpenak zentzu gutxi duen eta hasiera eta bukaera gaizki idatziak daudela. Nabaria da atazarako egokitzapen falta. Bestalde, erauzketa bidezko metodoak laburpen txukunagoa egin du, badirudi konturatu dela albistegietako artikulu gehienetan lehen esaldiek duten garrantziaz, laburpenean ipini ohi baititu.

5 Ondorioak

Proiektu honetan laburpengintza automatikoaren arlorako sarrera bat egin da, gaur egungo egoera eta etorkizunean izan ditzakeen onurak ezagutzeko. Proiektua egin bitartean laburtzaile automatikoak sortzearen konplexutasuna begi bistan geratu da eta emaitzetan hori ere islatu da. Hala ere, neuronasareak alor honetan aurrera egiteko tresna egokiak izan daitezkeela ere ikusi da, beti ere, nahikoa denbora eta konputazio gaitasun edukita.

Garatutako sistemei dagokienean, badirudi erauzketa bidezko sistemak kalitate altuagoko laburpenak egiten dituela. Sistema honek landu duen ataza abstrakzio bidezko sistemarena baino errazagoa izan da eta horrek ziur eduki duela eragina. Abstrakzio bidezko sistema garatzeko *T5-small* eredua erabili da oinarri bezala, horren ordez *T5-base* edo *T5-large* erabiliz gero eta entrenamendudenbora nahikoa utzita, sistema honen kalitatea asko handituko litzateke. Gainera, hiperparametroen doikuntza egoki batekin (laburpenen gehieneko luzera altuagoa, sorta tamaina ezberdinak, etab.) emaitza hobeagoak lortu daitezkeen begiratu beharko litzateke.

Hobekuntzei dagokienean, erauzketa bidezko sistema erraz hobetu daiteke, lehenik, hiperparametroen doitzea eginez eta ondoren, datu-base egokiago baten gainean entrenatuz. Izan ere, datu-basea ez da egokia erauzketa bidezko sistemak behar bezala entrenatzeko, edo ez behintzat proiektu honetan landu dena bezalakoa entrenatzeko. Sistema honen urrezko estandarrean esaldi bakoitza laburpenean egon beharko litzatekeen edo ez agertu behar da eta erabilitako datu-baseak ez zuen esaldien inguruko informazio hori. Datu-basea egokitu egin da entrenamendua egin ahal izateko, ROUGE-L indizean oinarrituta esaldi bakoitzari etiketa bat ipiniz, laburpenean egon behar lukeen edo ez dioena. Horrek ROUGE-Lren fidagarritasunaren mende utzi du urrezko estandarraren zuzentasuna eta gainera, inplementazioan ikusi daitekeen bezala, adibideen aniztasuna murriztu du. Horrek eragin nabarmena izan du, ziur aski, ereduaren gaitasunean.

Bestalde, ebaluazioari dagokionean, datu-baseko urrezko estandarreko laburpenak ez dira egokienak erauzketa bidezko sistema ebaluatzeko, laburpen horiek ez dituztelako jatorrizko testuko esaldiak bere horretan ipinita, eraldaketa sintaktiko eta gramatikoak jaso dituzten testuak dira (gizakiek eginak). Horrek alderaketak zailtzen ditu.

Azken ondorio bezala, proiektuaren helburu nagusietako bat bete dela esan daiteke, laburpengintza automatikoaren munduan murgiltzea mundu horretako teknikak imitatu eta erabiltzen saiatuz. Sortutako ereduek gabeziak dituzte, baina haiek sortzeari esker hurrengo lanetarako errazagoa izango da sistema berriak sortzea, nondik jo eta nondik ez jo ikusi ahal izan delako lan honetan.

References

[1] Guruh Fajar Shidik Edi Noersasongko Abdul Syukur Affandy Affandy eta De Rosal Ignatius Moses Setiadi Adhika Pramita Widyassari, Supriadi Rustad. Review of automatic text summarization techniques methods. *Science Direct*, Apirila 2022.

[2] Ed Collins eta Isabelle Augenstein eta Sebastian Riedel. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Abuztua 2017.

[3] Adam Roberts Katherine Lee Sharan Narang Michael Matena Yanqi Zhou Wei Li eta Peter J. Liu Colin Raffel, Noam Shazeer. Exploring the limits of transfer learning with a unified text-to-text transformer. Uztaila 2020.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *aclanthology*, Uztaila 2004.

6 Material gehigarria

- Sistemen inplementazioa azaltzen duen koadernoa.
- Sistemek sortutako bost laburpenak.
- Gaian sakontzeko hainbat esteka baliagarri: lehena, bigarrena.