IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

III. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2019ko maiatzaren 27, 28 eta 29
Baiona, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Euskaraz hitz egiten ikasten duten makina autodidaktak

Asier López Zorrilla,
Mikel de Velasco Vázquez
eta *Raquel Justo Blanco*

125-132 or.
https://dx.doi.org/10.26876/ikergazte.iii.03.16

Euskaraz hitz egiten ikasten duten makina autodidaktak

López Zorrilla, Asier; deVelasco Vázquez, Mikel eta Justo Blanco, Raquel

Euskal Herriko Unibertsitatea UPV/EHU

asier.lopezz@ehu.eus

Laburpena

Lan honetan sare neuronalen bidez euskaraz hitz egiten ikasten duen elkarrizketa sistema automatikoa aurkezten dugu. Horretarako, Turingen testaren ideia era konputazionalean inplementatzen duten sare neuronal sortzaile aurkariak erabili ditugu. Normalean erabiltzen diren ingelesezko corpusak baino bi magnitude ordena txikiagoa den euskarazko corpus batekin halako sareak entrenatzea badagoela frogatzen dugu. Amaitzeko, euskararen morfologia kontuan hartzen duen aurreprozesamendua erabiltzea komenigarria dela erakusten dugu. Dakigunaren arabera, sare neuronaletan oinarrituta dagoen euskarazko lehen elkarrizketa sistema aurkezten dugu.

Hitz gakoak: elkarrizketa sistema automatikoak, sare neuronalak, sare neuronal sortzaile aurkariak, euskara

Abstract

This work presents a neural dialogue system capable of learning Basque. To this end, we build upon generative adversarial networks which implement the idea of the Turing test. We demonstrate that training such a dialogue system with corpora two orders of magnitude smaller than usual English corpora is feasible. Finally, we also found that preprocessing the Basque language according to its morphology helps training these neural models. To the best of our knowledge, this is the first attempt to develop a neural dialogue system in Basque.

Keywords: dialogue systems, deep learning, generative adversarial networks, Basque

1. Sarrera eta motibazioa

Elkarrizketa sistema automatikoek pertsona eta makinen arteko komunikazioa eta interakzioa ahalbidetzen dute, lengoaia naturalaren bidez. Oro har, bi motatako elkarrizketa sistema desberdintzen dira: helburudunak eta helbururik gabekoak edo eremu irekikoak. Lehenengo kategorian erabiltzailearen nahi espezifikoak asetzeko eraikitako sistemak sartzen dira. Adibidez, busen ordutegiak eta lineak kontsultatzeko (Olaso Fernández eta Torres, 2017) eta jatetxeetan edo hoteletan erreserbak egiteko (Bordes eta Weston, 2016) balio duten sistemak helburudunak dira. Hauetaz gain, azken urteotan hedatu diren laguntzaile birtualak, hala nola Siri, Cortana, Google Assistant edo Alexa, elkarrizketa sistema helburuduntzat ere har ditzakegu, normalean euren lana erabiltzailearen aginduak burutzea baita, esate baterako dei bat egitea edo Interneten biharko eguraldiaren iragarpena bilatzea.

Beste aldetik, eremu irekiko elkarrizketetan ez dago aldez aurretik definitutako helbururik ezta gairik. Hau da, erabiltzaileak eta makinak ez diote elkarri hitz egiten helburu espezifiko batekin; interakzioa bera naturala eta zentzuduna izatea da helburua. Horretarako, sistemak esaldi ahal bezain logiko, koherente eta informatzaileekin erantzun behar dio erabiltzaileak esaten duenari. Beste modu batean esanda, sistemak era gizatiarrean hitz egin behar du. Lan honetan elkarrizketa sistema mota horietan zentratuko gara.

Era gizatiarrean hitz egitearen ideiarekin lotuta, Alan Turing matematikariak 1950. urtean bere test famatua aurkeztu zuen: Turingen testa (Turing, 1950). Testaren ideia nagusia honakoa da: sistema automatiko bat kalitatezkoa edo adimenduna dela esateko, sistema hori eta pertsona bat elkar bereizezinak izan behar dute haiekin hitz egiterako orduan. Sistema batek halako propietatea betetzen duen egiaztatzeko, Turingek hainbat epaile zenbait makinekin hitz egiten jartzea proposatu zuen, makina batzuen atzean sistema automatikoak eta besteen atzean pertsonak daudelarik. Egoera horretan epaileek ehuneko altu[1] batean usteko balute sistema automatikoa pertsona bat

[1]Eztabaida handia dago sistema batek Turingen testa gainditzeko behar duen portzentajearen inguruan. Erreferentzia gisa, 2011. urtean Indian Institute of Technology Guwahati institutuan ospatutako Turingen test batean, epaileek pertsonak pertsona moduan sailkatu zituzten ebaluazioen % 63,3-an.

dela, orduan sistema hori erabat adimenduna dela esan liteke.

Denbora pasa ahala, Turingen testa gainditzearen ideiak gero eta ikerketa gehiago bultzatu zituen adimen artifizialaren arloan. Adibidez, 1966. urtean ELIZA programa (Weizenbaum, 1966) aurkeztu zuten MIT-eko ikerlariek. Programaren funtsa hitz gakoak detektatzean eta horien arabera aurredefinitutako esaldi bat aukeratzean datza. Algoritmo hori sinplea izan arren, hainbat epailek pertsonatzat hartzea lortu zuen.

Hurrengo hamarkadetan ikerketek aurrera jarraitu zuten arren, benetan Turingen testa gainditzeko gai zen sistemarik ez zen lortu. 2011. urtean Turingen test batean inoiz lortu diren emaitzarik onenak Cleverbot sistemak[2] lortu zituen, berarekin hitz egin zuten 1.334 epaileetatik % 59,3-ak pertsonatzat hartu zuenean. ELIZA-k ez bezala, Cleverbot-ek ez ditu aurredefinitutako esaldiak erabiltzen. Horren ordez urteetan zehar pertsonekin edukitako elkarrizketak erabiltzen ditu erantzuterako orduan. Hitz gutxiekin esanda, esaldi bati erantzuteko Cleverbot-ek esaldi hori edo antzeko bat esan duenean zein erantzun jaso duen bilatzen du, eta erantzun horretaz abiatuz sortzen du bere erantzuna. Ideia hau interesagarria bada ere, konputazionalki nahiko konplexua da, denbora zein memoriaren ikuspegitik, datu-base oso handi batean bilaketak egitea baitakar. Are gehiago, datu-basea gero eta handiagoa izan, orduan eta denbora eta memoria gehiago beharko du halako sistema batek erantzun bat lortzeko.

Eragozpen horiek saihesteko, baita adimen artifizialaren beste arloetan izan duten emaitzengatik, azken urteetan sare neuronalak elkarrizketa sistemak eraikitzeko teknologia nagusia bilakatu dira. Sare neuronalak datuetatik eredu konputazional konplexuak lortzeko balio duten paradigma konputazional bat dira, bereziki eraginkorra datuen kantitatea oso handia denean. Ulertzekoa da, beraz, arloko autore gehienek ingelesez dauden datu-baseekin lan egitea, normalean hauek baitira handienak, eta hortaz sare neuronalak hobeto funtzionatuko dutelako. Baina, zer gertatzen da baliabide gutxiagoko hizkuntzekin? Ba al dago sare neuronaletan oinarrituriko elkarrizketa sistema automatikoak eraikitzerik euskaraz?

Lan honetan erakusten dugu baietz, badagoela. Normalean erabiltzen diren datu-baseak baino bi magnitude ordena txikiagoak diren datu-baseak erabiliz modu koherente eta zentzudunean euskaraz hitz egiten duen elkarrizketa sistema automatikoa aurkezten dugu.

2. Arloko egoera eta ikerketaren helburuak

Sare neuronalen bidezko eremu irekiko elkarrizketa sistemak itzulpen automatikorako erabiltzen diren sareetan oinarritzen dira, hots, sekuentziatik-sekuentziarako sareetan (Sutskever *et al.*, 2014; Cho *et al.*, 2014) (*Sequence to sequence networks* ingelesez). Sare neuronal horiek luzera arbitrarioko bektore segida bat har dezakete sarrera moduan, eta era berean beste luzera arbitrarioko segida bat sortu. Hala, eremu irekiko elkarrizketak sortzearen problema transdukzio problema bat bezala planteatzen badugu, sare horiek erabili ditzakegu. Hori egiteko, sarrera erabiltzaileak esandako hitzen segida izango da, eta irteera sistemaren erantzunari dagozkion hitzen sekuentzia.

Sare horiek entrenatzeko, edo euren parametroak doitzeko, ikasketa metodo gainbegiratuak erabili ohi dira, aipatutako sarrera-irteera bikoteez osaturiko corpusen bat erabiliz. Adibidez, lan honetan filmen azpitituluak erabiliko ditugu corpus hau eratzeko: sarrera bakoitza aktore batek esandako esaldi bat izango da, eta dagokion irteera beste aktore batek emandako erantzuna. Metodologia hau erabiliz emaitza interesgarriak lortu ahal diren arren, askotan horrela entrenatutako sareek informaziorik gabeko erantzun orokorrak sortzeko joera dute, hala nola *I don't know* edo *I'm sorry*[3] (Sordoni *et al.*, 2015; Serban *et al.*, 2016). Tuan eta Lee autoreek (2019) adierazten duten moduan, ikasketa metodo gainbegiratuek irteera bakarra esleitzen diote sarrera bakoitzari, baina horrek ez ditu elkarrizketen propietateak behar bezala jasotzen. Izatez, hitz egiten dugunean, norbaitek esan duenari erantzuteko hamaika esaldi ezberdin erabili ahalko genituzke, guztiak onargarriak. Horrela, esaldi askoren erantzuna izan daitezkeen esaldi generikoak probabilitate handiarekin sortzen ditu sareak.

Arazo hori konpontzeko, ikasketa gainbegiratuaren ordez sare sortzaile aurkariak (*Generative adversarial networks* ingelesez) (Goodfellow *et al.*, 2014) erabiliko ditugu lan honetan. Sare sortzaile aurkariek Turingen testaren ideia era konputazionalean aplikatzea ahalbidetzen dute. Kasu honetan, erantzunak sortzen dituen sareari (sare sortzailea hemendik aurrera) ez zaio adieraziko zein irteera dagokion sarrera bakoitzari. Horren ordez, beste sare batek, sare diskriminatzaileak, ebaluatuko ditu sare sortzaileak emandako erantzunak, zein punturaino gizatiarrak diren esanez, Turingen testaren epaile batek egingo lukeen modu berean. Sare sortzailearen helburua sare diskriminatzaileak berari emandako ebaluazioa ahal bezain beste hobetzea izango da. Sare diskriminatzailearena, aldiz, pertsonek sortutako eta sare sortzaileak sortutako esaldien artean bereiztea izango da. Modu honetan, bi sareak

[2] `https://www.cleverbot.com/`, azken bisita 2019ko martxoaren 22an.

[3] Arazo hori deskribatzen duten lanek ingelesez egiten dituztenez saiakuntzak, ingelesez ere ipini ditugu haiek erakutsitako adibideak.

iteratiboki entrenatuko dira; sortzailea saiatuko da diskriminatzaileak hura pertsonatzat hartzen, diskriminatzaileak sortzailearen eta pertsonen artean bereizten ikasten duen bitartean.

Halako optimizazio prozesua burutzea, dena den, ez da sinplea, sareak entrenatzeko normalean erabiltzen diren gradienteetan oinarritutako optimizazio metodoak ez baitira zuzenean aplikagarriak. Xehetasunetan sartu gabe, sare diskriminatzailearen irteera ez da diferentziagarria sare sortzailearen parametroekiko, sortzaileak sortutako hitzak diskretuak dira eta (Yu *et al.*, 2017). Errefortzu bidezko ikasketa erabili daiteke gradienteetan oinarritutako metodoen ordez (Li *et al.*, 2017; Hori *et al.*, 2019), baina horrek entrenamenduaren konbergentzia zaildu dezake (Sutton *et al.*, 1998). Beste aukera bat *straight-through Gumbel-softmax* (Bengio *et al.*, 2013; Jang *et al.*, 2016) zenbateslearen bidez gradientearen hurbilketa bat egitea da, Lu *et al.* (2017) eta Shetty *et al.* (2017) autoreek erakusten duten moduan. Azkenik, lan honetako autoreek guztiz diferentziagarria den sare sortzaile aurkari bat aurkeztu berri dute (López Zorrilla *et al.*, 2019)[4], hitzen errepresentazio bektorial hurbilduak erabiltzen dituena, ondoren azalduko dugun moduan.

Testuinguru honetan, lan honen ekarpenak hiru dira: alde batetik, López Zorrilla *et al.* autoreek (2019) proposatutako sare sortzaile aurkaria balioztatzen dugu, ingelesez gain euskaraz ere eraginkorra dela frogatuz; bigarrenik, modu koherente eta zentzudunean hitz egiten duen sare neuronaletan oinarritutako elkarrizketa sistema automatikoa euskaraz eraikitzea badagoela frogatzen dugu; eta amaitzeko lematizazio prozesu baten bidez corpusaren tamaina txikiagoagatik sortutako desabantailak nola leundu daitezken erakusten dugu.

3. Ikerketaren muina

Atal honetan, hasteko, erabilitako sare neuronalen egiturak azalduko ditugu. Ondoren, bi sareen parametroak doitzeko erabilitako algoritmo iteratiboa aurkeztuko dugu. Jarraitzeko, euskaraz dagoen corpusa nola aurreprozesatu dugun deskribatuko dugu. Azkenik, burututako saiakuntzak eta lortutako emaitzak erakutsi eta aztertuko ditugu.

3.1. Sare sortzaile eta diskriminatzailea

Sare sortzailea sekuentziatik-sekuentziarako sare bat da, *long short-term memory* (LSTM) (Hochreiter eta Schmidhuber, 1997) kodetzaile eta deskodetzaile errekurrente independenteekin (Sutskever *et al.*, 2014) eta arreta modulu batekin (Bahdanau *et al.*, 2015; Luong *et al.*, 2015). Sare honek T luzera arbitrarioko hitzen errepresentazio bektorialen (Mikolov *et al.*, 2013) segida bat hartuko du sarrera moduan: $\mathbf{v} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T$. Sarrera hori prozesatu ostean, irteera moduan beste τ luzera arbitrarioko segida bat bueltatuko du, elementu bakoitza sareak sor ditzakeen hitz guztien arteko probabilitate-banaketa delarik: $\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_\tau$.

Bestalde, sare diskriminatzailea beste bi kodetzaile errekurrentez osaturik dago, guztiz konektatutako geruza batzuez jarraituak, Kannan eta Vinyals autoreen (2017) antzera. Kodetzaile bakoitzak esaldi bat hartzen du sarrera moduan. Batek erabiltzailearen mezua $\mathbf{v}$ prozesatuko du, eta besteak erantzuna $\mathbf{u}$. Sistemaren irteera 0 eta 1-en arteko zenbaki erreal bat, a, izango da, erabiltzailearen mezuari emandako erantzuna zein punturaino gizatiarra den adierazten duena. Irteera zenbat eta baxuagoa, orduan eta gizatiarragoa izango da erantzuna, sarearen irizpidearen arabera. Bi sarrerak, berriz ere, hitzen errepresentazio bektorialen moduan hartuko ditu sareak.

Bi sareen gainontzeko xehetasunak (López Zorrilla *et al.*, 2019) lanean aurkitu daitezke.

3.2. Ikasketa algoritmoa

Bi sareak entrenatzeko, hiru optimizazio prozesu era iteratiboan burutuko ditugu. Lehenago aipatu dugun bezala, alde batetik sare sortzailea entrenatuko dugu diskriminatzaileak hura pertsonatzat hartzeko, hau da, diskriminatzailearen irteera minimizatzeko. Bigarrenik, diskriminatzailea entrenatuko dugu sare sortzaileak sortutako erantzunak eta corpusetik hartutako erantzunak desberdintzeko. Amaitzeko, Li *et al.* autoreen (2017) legez, sare sortzailearen parametroak ikasketa metodo gainbegiratuen bidez doituko ditugu ere, prozedura guztiaren konbergentzia bermatzeko.

Hiru optimizazio prozesu hauek definitzeko, horietako bakoitzean gradienteetan oinarritutako optimizazio metodoekin minimizatuko ditugun galera-funtzioak zehaztuko ditugu.

[4] Apirilean argitaratuko da lana.

Sare sortzailearen parametroen egiantz handieneko estimazioa

Sare sortzailearen parametroak ikasketa gainbegiratuaren bidez doituko ditugu egiantz handieneko estimazio baten bidez. Hau da, corpuseko sarrera-irteera bikote bakoitzarentzat, sareak sarrera prozesatzean irteera desiratua sortzeko duen probabilitatea maximizatuko dugu. 1 ekuazioan agertzen den galera-funtzioa erabiliko dugu.

$$L_{EH} = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{v},s \in \mathcal{C}} \frac{1}{|s|} \sum_{t=1}^{|s|} -\log \mathbf{p}_t[s_t] \,, \tag{1}$$

non $\mathcal{C}$ $\mathbf{v}$ sarreretaz eta s irteera desiratuez osatutako corpusa den, s_t irteera desiratuaren t-garren hitzari dagokion indizea, eta $\mathbf{p}_t[s_t]$ sareak t-garren denbora unean s_t hitzari esleitutako probabilitatea den. Sarearen irteera, $\mathbf{p}$, $\mathbf{v}$ sarreraren funtzioa da noski, baina mendekotasun hori ez dugu esplizituki adierazi notazioa ez korapilatzeko.

Sare diskriminatzailearen galera-funtzioa

Sare diskriminatzailearen entrenamendua burutzeko lehenik eta behin corpus berri bat sortu beharko dugu, $\mathcal{C}_D$, $\mathcal{C}$ corpusetik abiatuz eta sare sortzailea erabiliz. Diskriminatzaileak pertsonek emandako eta sare sortzaileak sortutako erantzunen artean desberdintzen ikasi behar duenez, bi eratako laginak behar ditu euren artean diskriminatu ahal izateko. Horretarako, bi motatako hirukoteez osatuko dugu $\mathcal{C}_D$ corpusa. Hirukote bakoitza erabiltzaileak bidalitako mezu batez, erantzun batez eta 0 edo 1 izan daitekeen etiketa batez osatuta egongo da. Lehenengo motako hirukoteek gizakiek emandako erantzunak edukiko dituzte, eta beraz etiketa 0 izango da. Hirukote hauek lortzeko $\mathcal{C}$ corpuseko bikoteak erabili genituen zuzenean. Bestalde, bigarren motako hirukoteek sare sortzaileak sortutako erantzunak edukiko ditu, eta ondorioz etiketaren balioa 1 izango da. Hirukote hauek eratzeko, $\mathcal{C}$ corpusetik hartu dira erabiltzailearen mezuak, gero hauek sare sortzaileari pasa sarrera moduan, eta sarearen irteera erabili erantzun moduan. $\mathcal{C}_D$ eraiki ondoren, entropia gurutzatuko galera-funtzioa erabili dugu sare diskriminatzailearen parametroak doitzeko (2 ekuazioa).

$$L_D = \frac{1}{|\mathcal{C}_D|} \sum_{\mathbf{v},\mathbf{u},l \in \mathcal{C}_D} -\left[l \cdot \log a + (1-l) \cdot \log\left(1-a\right)\right] \,, \tag{2}$$

non $\mathbf{v}$ erabiltzailearen mezuaren hitzen errepresentazio bektorialen segida den, $\mathbf{u}$ erantzunarena, l erantzuna pertsona batena edo sare sortzailearena den adierazten duen eskalarra, eta a diskriminatzailearen irteera. Berriro ere, a-k $\mathbf{v}$ eta $\mathbf{u}$-rekiko duen mendekotasuna ez dugu esplizituki adierazi.

Sare sortzailearen galera-funtzio aurkaria

Azkenik, sare sortzailea diskriminatzailearen irteera minimizatzeko galera-funtzioa definitzea erraza da, diskriminatzailearen irteera bera baita, 3 ekuazioan ageri den bezala.

$$L_S = \frac{1}{|\mathcal{C}_S|} \sum_{\mathbf{v} \in \mathcal{C}_S} a \,, \tag{3}$$

non $\mathcal{C}_S$ corpusa $\mathcal{C}$ corpusean dauden sarrera mezuez osatuta dagoen, $\mathbf{v}$ horietako bakoitza delarik. a diskriminatzailearen irteera da.

3 ekuazioko galera-funtzioa gradienteetan oinarritutako optimizazio metodoekin minimizatu ahal izateko, a sare sortzailearen parametroekiko diferentziagarria izan behar du. Sare sortzaileak $\mathbf{v}$ sarrera $\mathbf{p}$ irteeran era guztiz diferentzialean transformatzen du. Era berean, sare diskriminatzaileak bere bi sarrerak, $\mathbf{v}$ eta $\mathbf{u}$, era guztiz diferentzialean transformatzen ditu a irteeran. Hortaz, diferentziagarritasuna ez galtzeko $\mathbf{p}$ $\mathbf{u}$-n transformatu behar da transformazio diferentziagarri baten bidez. $\mathbf{p}$-ko elementu bakoitza, hots, $\mathbf{p}_t$, sareak esan ditzakeen hitz guztien arteko probabilitate-banaketa bat da. Normalean $\mathbf{p}_t$-ko maximoaren argumentua hartuko genuke sareak t-garren denbora unean esan duen hitza bezala. Baina argmax operazioa ez da deribagarria.

Arazo horri irtenbidea emateko, López Zorrilla *et al.* autoreen (2019) prozedura berdina erabiltzen dugu lan honetan. $\mathbf{p}_t$-ri dagokion errepresentazio bektoriala, $\mathbf{u}_t$, lortzeko, $\mathbf{p}_t$-ko k elementurik handienak hartzen ditugu, *top-k* operazio baten bidez. Horrela elementu horien $\widetilde{\mathbf{p}}_t$ balioak eta $\mathbf{k}_t$ indizeak lortzen ditugu. Jarraian $\widetilde{\mathbf{p}}_t$ normalizatzen dugu *softmax* normalizazio batekin, $\widehat{\mathbf{p}}_t$ lortuz. Azkenik, $\mathbf{u}_t$ kalkula dezakegu $\mathbf{k}_t$ indizeei dagozkien hitzen errepresentazio bektorialen arteko bataz besteko aritmetiko haztatua eginez, pisuak $\widehat{\mathbf{p}}_t$ direlarik.

Goi mailako optimizazio algoritmoaren deskribapena

Erabiliko diren hiru galera-funtzioak deskribatu ondoren, hauek iteratiboki minimizatzeko prozedura zehaztuko dugu. Hasteko, sare sortzailearen parametroak ez ditugu ausaz hasieratuko. Horren ordez, hainbat iteraziotan zehar doituko ditugu hasieran, 1 ekuazioko egiantz handieneko galera-funtzioa minimizatuz. Behin sare sortzaileak kalitate onargarriko esaldiak sortzen dituela, $\mathcal{C}_D$ corpusa bere erantzunekin eta $\mathcal{C}$-ko pertsonen erantzunekin hasieratuko dugu, eta sare diskriminatzailea lehenengo aldiz entrenatuko dugu.

Ondoren algoritmoaren begizta nagusia hasten da. Horretan, sare sortzailea eta diskriminatzailea iteratiboki entrenatzen dira. Sare sortzailea entrenatzeko galera-funtzio aurkaria (3 ekuazioa) eta egiantz handieneko galera-funtzioak (1 ekuazioa) txandakatzen dira. Prozedura osoan zehar, sare sortzailearen entrenamendu prozesu bakoitza amaitu ostean, hainbat sarrera ausaz aukeratzen dira $\mathcal{C}$-tik eta sare sortzaileak sortutako erantzunak $\mathcal{C}_D$-ra gehitzen dira, eta diskriminatzailea entrenatzen da hainbat iteraziotan zehar. Prozeduraren konbergentzia bermatzeko, diskriminatzailea entrenatzerakoan probabilitate handiagoarekin hartzen dira $\mathcal{C}_D$-n sartutako erantzun berriagoak.

3.3. Euskararen aurreprozesamendua eta lematizazioa

Esan dugunez euskarazko corpus batekin entrenatuko ditugu sareak. Ingelesa ez bezala, euskara hizkuntza eranskaria da egitura morfologikoaren aldetik. Hau da, euskarak monema independenteak elkartuz sortzen ditu hitzak. Horrela, askotan euskaraz hitz batekin esan daitekeena ingelesez hainbat hitz erabiliz adierazi behar da. Adibidez, ingelesezko "*to the cinema*" euskaraz "zinemara" bezala itzuliko litzateke, edo "*because of the baby*" "haurrarengatik" bezala. Sareen ikuspegitik hitz bakoitza token independente bat denez, sareak ez ditu ikusten euskaraz gertatzen diren hitzen arteko erlazioak, euskararen prozesamendu automatikoa zailduz. Hasiera batean behintzat, sarearentzat "haurrarengatik" eta "haurraren" hitzak "haurrarengatik" eta "daitezke" bezain ezberdinak dira.

Honek hitzen errepresentazioa zailtzen du bi sareen sarreran, baita sare sortzailearen irteeran ere. Sareen sarreretan, hitzen egituran arreta jartzen duten errepresentazio bektorialak erabiliko ditugu hitzen arteko erlazio horiek sortzeko, *Fastext* (Bojanowski *et al.*, 2016) esate baterako. Dena den, irteeran ezin da arazoa horrela konpondu, sare sortzaileak hitzen arteko probabilitate-banaketa bat sortzen duelako. Honi irteera ematen saiatzeko, hitzen lexemak kasu marketatik eta postposizioetatik banatzea proposatzen dugu. Zehazki, izen, izenordain, izenondo eta determinanteak bananduko ditugu lan honetan. Hitzen lexema eta kategoria gramatikala topatzeko, Agerri *et al.* autoreek (2014) eraikitako kode irekiko lematizadorea erabiliko dugu. Izen, izenordain, izenondo edo determinante baten lexema eta postposizioa banatuko diren ala ez erabakitzeko, baldintza sinple bat erabiliko da. Halako hitz baten bukaera postposizio baten berdina balitz, orduan hitza lexema eta postposizioan bananduko dugu. Adibidez, "zeruko" hitza "zeru" lexeman eta "-ko" postposizioan banatuko genuke. Hogeita bi postposizio hartu genituen kontuan: "-ri", "-ei", "-rekin", "-ekin", "-ren", "-en", "-n", "-tik", "-dik", "-rik", "-ra", "-tara", "-rengana", "-engana", "-rantz", "-raino", "-z", "-rako", "-ko", "-entzat", "-tzat" eta "-gatik".

Lematizazioaz gain, izen propioak <izen> tokenera bihurtuko ditugu, normalean pertsonen izenak baitira, eta beraz, funtzio berdina dutelako esaldietan. Era berean, zenbakiak <zenbaki> tokenera bihurtuko dira.

3.4. Saiakuntzak eta emaitzak

Orain arte azaldutako sareak, ikasketa algoritmoa eta euskararen aurreprozesamendua balioztatzeko, OpenSubtitles (Lison eta Tiedemann, 2016) corpusaren euskarazko bertsioarekin entrenatuko dugu deskribatutako elkarrizketa sistema. Corpus horretatik milioi bat sarrera-irteera bikote atera daitezke, ingelesezko bertsioan baino 420 aldiz gutxiago. Corpusa 3.3 atalean azaldutako metodologiarekin aurreprozesatuko dugu, ondorioz hitz desberdinen kopurua berrehun milatik ehun milara jaitsiz. Normalean egiten den bezala, hitz horietako azpimultzo bat baino ez dugu kontuan hartuko saiakuntzetarako: maiztasun handieneko 15.000 hitzak. Gainontzekoak corpusetik kenduko dira. Aurreprozesamenduaren efektua erakusteko, corpus aurreprozesatua zein aurreprozesatu gabearekin entrenatuko ditugu sareak.

Kasu bietan, dena den, hiper-parametro berdinak erabiliko ditugu sareen arkitekturan eta baita ikasketa algoritmoan. Hiper-parametro horietako inportanteenak jarraian aipatzen ditugu. Sare errekurrente guztiak, hau da, sare sortzailearen kodetzailea, deskodetzailea, eta sare diskriminatzailearen bi kodetzaileak, bi LSTM geruzaz osatuta daude. Sare sortzailearen geruzak 1.028 zelda dituzte, eta diskriminatzailearenak 128. Adam optimizazio metodoa (Kingma eta Ba, 2014) erabiliko dugu 3.2 ataleko hiru galera-funtzioak minimizatzeko, 512 tamainako *batch*-ak erabiliz. Hitzen errepresentazio bektorialak *Fastext* metodologiarekin hasieratuko dira. Sare sortzailea 50.000 iteraziotan zehar entrenatuko dugu, ikasketa begizta hasi baino lehen. Hori 500 aldiz errepikatu dugu ondoren. Iterazio bakoitzean sare sortzailea zein diskriminatzailea 40 iterazioetan zehar entrenatuko da.

Liu *et al.* autoreek (2016) adierazten duten moduan, ebaluazio automatikoak ez dira komenigarriak elkarrizketa sistemen kalitatea neurtzeko. Hortaz, entrenatutako sistemen funtzionamendua erakusteko, zenbait sarrerako mezuen aurrean emandako erantzuna erakusten dugu emaitza modura. Erreferentzia bezala, metodologia berdinarekin [5] baina OpenSubtitles corpusaren ingelesezko bertsioarekin entrenatutako sareak emandako erantzunak ere erakusten ditugu. Erantzun guztiak 1 taulan ageri dira.

1. taula. Mezu ezberdinen aurrean elkarrizketa sistemek emandako erantzunak. A ingelesez entrenatutako sistema da, B euskaraz aurreprozesamenduarekin entrenatutakoa eta D aurreprozesamendurik gabekoa.

Sarrera mezua	Sistemaren erantzuna
A: hi !	hello , <person>.
B: kaixo !	kaixo , <izen> naiz , <izen>-ren semea naiz .
D: kaixo !	kaixo , kaixo.
A: I love you	you know , I think maybe we could talk ...
B: maite zaitut	nik ere maite zaitut .
D: maite zaitut	maite zaitut
A: who is she ?	she 's my girlfriend , <person> , and ...
B: nor da neska hori ?	<izen> da neska .
D: nor da neska hori ?	nor da neska ?
A: what's the color of the sky ?	it 's a beautiful blue sky
B: zein da zeru -ko kolorea ?	zeru urdin iluna da .
D: zein da zeruko kolorea ?	zein da zure kolore faboritoa ?
A: do n't worry	I do n't want to hear about it.
B: ez kezkatu	ez dut ulertzen zergatik egin duen hau zu -tzat , <izen> ?
D: ez kezkatu	ez dut ahaztuko
A: could you come with us there ?	I 'll come with you .
B: toki hori -tara lagunduko diguzu ?	bai , horixe uste dut nik ere .
D: toki horretara lagunduko diguzu ?	jakina baietz
A: I forgot it .	I did n't know you were coming .
B: ahaztu egin zait .	<izen> .
D: ahaztu egin zait .	zer ?
A: can you see anything ?	I see you in the mirror .
B: zerbait ikusten ?	ez , ez
D: zerbait ikusten ?	ez .
A: it was a plane crash .	I was in the hospital
B: hegazkin istripu bat izan zen .	<izen> larri zaurituta zegoen eta bere etxe -ra bueltatu zen .
D: hegazkin istripu bat izan zen .	sute bat izan zen .
A: did he wake up again ?	no he did not, but he did n't sleep .
B: jaiki al da berriz ?	zer esan nahi du horrek ?
D: jaiki al da berriz ?	astelehenean atera behar duzu .
A: yesterday I saw you in the mountains .	and now I see you
B: atzo mendi -tik ikusi zintudan .	ikusi zintudan <zenbaki> aldi -n .
D: atzo menditik ikusi zintudan .	zer ?
A: see you	see you later , <person> .
B: agur	agur , aita .
D: agur	agur , ene erregea

[5] Ingelesez corpusaren tamaina handiagoa denez, sareak ere handiagoak dira eta iterazio gehiagotan entrenatu dugu. Zehaztasunak (López Zorrilla *et al.*, 2019) erreferentzian ematen dira.

4. Ondorioak

1 taulan ikusi daitekeen moduan, sare neuronal sortzaile aurkarien bidez euskaraz era nahiko koherente eta zentzudunean hitz egiten duen elkarrizketa sistema automatikoa lortu dugu. Ingelesarekin konparatuz euskaraz dauden baliabideen tamaina askoz txikiagoa izan arren, sare neuronalen bidezko metodologiak erabiltzea badagoela frogatu dugu. Horretarako, euskararen morfologia kontuan hartzea inportantea dela erakutsi dugu ere. Izen, izenordain, izenondo edo determinanteak lexema eta postposizioetan banatzea komenigarria da, sareak era eraginkorrago batean prozesatzen baitu lengoaia. 1 taulari begira, aurreprozesu horrekin sareak esaldi konplexuagoak sortzeko joera duela esan dezakegu.

Amaitzeko, lan honekin proposatu berri dugun (López Zorrilla *et al.*, 2019) eta testuarekin era guztiz diferentzialean lan egin dezakeen sare sortzaile aurkarien arkitektura baliozkotzen dugu, elkarrizketa sistema automatikoak euskaraz eraikitzeko aproposa dela egiaztatuz.

5. Etorkizunerako planteatzen den norabidea

Dena den, lan honetan aurkeztutako metodologia eta ideiak asko garatu behar dira benetan pertsona baten moduan euskaraz hitz egiten duen sistema lortzeko. Izatez, ingelesez ere oraindik urrun gaude halako sistemak sortzetik. Oraingoz baliabide handiagoko eta txikiagoko lengoaiekin sortutako sistemak parekatzea da gure hurrengo helburua. 1 taulan ageri den moduan, ingelesez entrenatutako sare sortzaile aurkaria era zentzuduneagoan eta gizatiarreagoan hitz egiten du euskarazko sistemarekin konparatuz.

Diferentzia hauek murrizteko, ezagutzaren transferentzia (*transfer learning* ingelesez) egiteko teknikak erabiltzeko asmoa daukagu. Ezagutzaren transferentziaren ideia nagusia corpus handiagoekin baina eginkizun ezberdin baterako entrenatutako ereduak eredu berriak sortzeko erabiltzea da. Kasu honetan, beraz, ingelesez sortutako sarea euskarazko sistema hobetzeko erabiltzea izango da gure helburua.

6. Erreferentziak

Agerri, Rodrigo, Josu Bermudez, eta German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual NLP tools. In *LREC*, volume 2014, 3823–3828.

Bahdanau, Dzmitry, Kyunghyun Cho, eta Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Bengio, Yoshua, Nicholas Léonard, eta Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* .

Bojanowski, Piotr, Edouard Grave, Armand Joulin, eta Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Bordes, Antoine, eta Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683.

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, eta Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, eta Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hochreiter, Sepp, eta Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9.1735–1780.

Hori, Takaaki, Wen Wang, Yusuke Koji, Chiori Hori, Bret Harsham, eta John R Hershey. 2019. Adversarial training and decoding strategies for end-to-end neural conversation models. *Computer Speech & Language* 54.122–139.

Jang, Eric, Shixiang Gu, eta Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* .

Kannan, Anjuli, eta Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198* .

Kingma, Diederik P, eta Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, eta Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547* .

Lison, Pierre, eta Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, eta Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Lu, Jiasen, Anitha Kannan, Jianwei Yang, Devi Parikh, eta Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, 314–324.

Luong, Minh-Thang, Hieu Pham, eta Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

López Zorrilla, Asier, Mikel deVelasco Vázquez, eta M. Inés Torres. 2019. A differentiable generative adversarial network for open domain dialogue. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, eta Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Olaso Fernández, Javier Mikel, eta M. Inés Torres. 2017. User experience evaluation of a conversational bus information system in spanish. In *8th IEEE International Conference on Cognitive Infocommunications*.

Serban, Iulian V, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, eta Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Shetty, Rakshith, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, eta Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, eta Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses.

Sutskever, Ilya, Oriol Vinyals, eta Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Sutton, Richard S, Andrew G Barto, eta others. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Tuan, Yi-Lin, eta Hung-Yi Lee. 2019. Improving conditional sequence generative adversarial networks by stepwise evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* .

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* LIX.433–460.

Weizenbaum, Joseph. 1966. ELIZA— a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9.36–45.

Yu, Lantao, Weinan Zhang, Jun Wang, eta Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

7. Eskerrak eta oharrak

Lan honen egileok gure esker ona adierazi nahiko genioke Eusko Jaurlaritzari, Euskal Herriko Unibertsitateari eta baita Europar Batzordeari, PRE_2017_1_0357 eta PIF17/310 zenbakidun diru laguntzekin, eta H2020 SC1-PM15 programako RIA 7 deialdiko 769872 zenbakidun diru laguntzarekin, hurrenez hurren, ikerketa hau babesteagatik.