DIAL: Telegram bidezko galdera-erantzun sistema

Jon Ander Campos
EHU / Tutorea
jcampos004@ikasle.ehu.eus

Iñigo Ortega
EHU / Ikaslea
iortega045@ikasle.ehu.eus

Abstract

Artikulu hau Telegram bidezko bot baten bidez elkarrizketak izateko eredu bati buruz jarduten da. Hori egiteko Seq2Seq sare neuronalen eredua erabiltzen da GRU motako sare errekurrenteen bidez eta atentzio sistema bat erabilita.

Eraikitako ereduarekin, Telegrameko bot bat egin daiteke eta elkarrizketak telegrametik bertatik izan bere bot-en API-a dagokion moduan erabilita.

1 Sarrera

Elkarrizketarako makina automatikoak Turing-en garaitik hasi ziren dagokien garrantzia izaten, honek berak proposatutako Turing probaren ondorioz, elkarrizketa bateko ondoko partaidea gizakia den edo ez egiaztatzeko erabili daitekeena. Hain zuzen, 1950tik dago elkarrizketen partaideak gizakiak diren edo ez egiaztatzeko interesa, ez proba huts bat bezala, baizik eta gizaki bat eta makina bat desberdintzeko zailagoa izan daitezen sistemak sortzeko.

Helburu horretarako, sistema asko eraiki dira, horien artean ospetsuenak, IBM Watson™, Alexa™ edo Siri™. Hasieran, erregeletan oinarritutako ereduak erabiltzen ziren hauen inplementaziorako, ordea, gaur egun, eta lan honen eredurako, ikasketa automatikoa erabiltzen da haren moldakortasunagatik.

Hori horrela izanik, azken urteotan helburu antzekoak dituzten aplikazioak sortu dira: Meena (Googleko zientzialari batzuek sortutako elkarrizketa bot bat (1)), Facebook Messenger-eko bot txertatuak (2) edo bezeroaren arretarako zerbitzuak eskaintzeko botak (3) (milioka Euro aurrezten dituztenak).

Kasu honetan, baita ere, ikasketa automatiko bidezko elkarrizketen eredu bat erabili eta inplementatu da. Eredu horren, bere hobekuntzen eta emaitzei buruzko jarduna da artikulu hau. Honen egiazkotasuna eta lekukoa hurrengoa inplementazioa da: Git biltegia.

2 Erlazionatutako lanak

Lan honetan Seq2Seq sare neuronalen arkitektura erabiltzen da. Hau, lehen aldiz, Googlen hasi zen erabiltzen. Hain zuzen, Sutskever et al. 2014ean ikertzaileek hasi ziren erabiltzen (4). Lehen eredu honek kodetzaile eta dekodetzaile bat lotzea proposatu zuen azken honetan egoerako irteera sarrera bezala erabiliz (Figure 1).

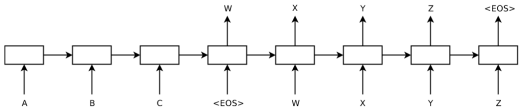

Figure 1: Sutskever et al. (2014): Lehen Seq2Seq eredua.

Eredu honen erabilpen nagusienetako bat itzulpen automatikoa da (5). Ordea, badirudi azken urteotan elkarrizketa sistemak inplementatzeko ere erabili dela nahiko emaitza onekin (1).

3 Sistema

Sistema honek aipatutako Seq2Seq eredua erabiltzen du. Honetan hitzen sekuentzia bat, beste batean bihurtzen da sare errekurrenteetan oinarritutako prozedura bat erabiliz. Hain zuzen, 2 zati nagusiz osatutako metodo bat erabiltzen da: Kodetzaile eta Dekodetzaile bat; biak GRU motako sare errekurrente bikoitzak izanik (BiGRU). Gainera, atentzio sistema bat erabiltzen da dekodetzailea, uneoro, kodetzaileak sortutako errepresentazio ezkutuez baliatu ahal izateko (6).

Sistemaren funtzionamendua honelakoa litzateke: Sarrera bezala esaldi bat emanda, GRU kodetzaileek jasotzen dute. Batek hasieratik bukaerara aztertzen du esaldia *aurreranzko* egoera ezkutuak sortuz eta besteak, alderantzizko

zentzuan berdina egiten du *atzeranzko* egoera ezkutuak sortuz. Sarreraren hitz bakoitzaren *embedding*-a *aurreranzko* eta *atzeranzko embedding*-en konkatenazioak definitzen du. Honela, dekodetzeko garaian, atentzio mekanismoari esker, unean interesgarrien diren *embedding*-ak hartzen dira gehien kontuan. Kasu honetan ere, bi GRU erabiltzen dira eta, kodetzailean egin den moduan, batek sarrera hasieratik aztertzen du eta besteak bukaeratik, baina, kasu honetan, sarrera atentzio mekanismoak itzulitako irteera da. Honek lenguaia naturaleko hitzak itzuliko lituzke (Figure 2).

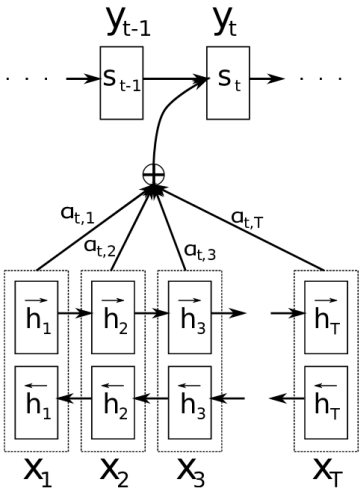

Figure 2: Erabilitako ereduaren adierazpen grafikoa. X_i sarrerako hitzak izanik.

Azaldutako sistema hau, Jon Ander Campos tutoreak eskainitako inplementazioan oinarrituta erabili da. Ordea, hurrengo azpi-ataletan aurkeztuko diren zenbait hobekuntza ere egin zaizkio.

3.1 Euskarazko tokenizazioa

Sistemak, jatorriz, ingeleseko tokenizatzaile bat erabiltzen zuen ikasketan, ingeleseko elkarrizketak izateko prestatuta baitzegoen. Ordea, kasu honetan, euskarazko elkarrizketak eduki nahi dira. Horretarako, *SpaCy* liburutegiak eskainitako euskarazko oinarrizko tokenizatzaile bat erabili da.

3.2 Erantzun motzen aurkako neurriak

Aipatu den bezala, ereduak erantzun oso motzak emateko joera dauka, hau da, "Bai" edo "Ez" erantzunak ematen ditu gehienbat, izan ere, elkarrizketetan gehien agertzen diren hitzak dira. Ondorioz, hitz horiek hain beste alditan ez agertzeko zenbait neurri hartu behar dira.

Kasu honetan, 2 neurri hartu dira, "Brevity Penalization" (BP) izeneko penalizazioa eta honen gainean aplikatutako Rayleigh distribuzio probabilitiskoa. Bi hauek ikasketan aplikatu dira, hain zuzen, galera funtzioan.

BP itzulpen automatikoan erabiltzen den penalizazio mota bat da (normalean *BLEU* izeneko metodo baten pean (7)). Honen helburua, kasu horretan, ereduak egindako itzulpen baten luzera errealarenarekin konparatzea da, ereduak itzultzen duena motzagoa balitz penalizazio bat aplikatuz.

Honelakoa litzateke bere ekuazioa:

$$BP(c,r) = \begin{cases} 1, & \text{baldin } c > r \\ e^{1-\frac{r}{c}}, & \text{bestela} \end{cases} \quad (1)$$

non c sistemak sortutako itzulpenaren luzera den eta r itzulpen errealaren luzera.

Ordea, honek arazo bat eragiten du: Ereduak bakarrik erantzun luzeenak itzultzen ikasten du. Hau, beti erantzun erreala baino luzera handiagoa edo berdineko itzulpen bat sortu nahi duelako gertatzen da. Orduan, luzera besterik ez optimizatzea eragiten du, hain zuzen, luzera handitzeko.

Horrek, printzipioz, ez luke arazorik eragin beharko, erantzun luzeak ere onargarriak izan daitezkeelako. Ordea, ereduak pauso bat gehiago ematen du: Hitzen konbinazio hoberenak erabiltzen ditu erantzun luzeenak sortzeko, zentzua izan edo ez. Honela, beti erantzun zentzugabeak itzultzen ditu, hitz luzeenak besterik ez baititu erabiltzen.

Arazo hori konpontzeko, Rayleigh distribuzioa erabili da. Hain zuzen, Rayleigh distribuzioaren ekuazioa hau bada:

$$Rayleigh(x,\sigma) = \frac{x}{\sigma^2} \cdot e^{-\frac{x^2}{2\sigma^2}} \quad (2)$$

Parametroak horrela definitu dira:

$$\sigma = 20 \quad (3)$$

$$0 <= y <= 1, \quad x = y \cdot 5.8 + 0.2 \quad (4)$$

Lortutako kurba, honen balio maximo posibleakrekin zatitzen da 0 eta 1-en artean mantentzeko (Figure 3).

Distribuzio hau BP-ren irteerari aplikatu zaio, honela, ereduak ez du erantzunen luzerarekiko optimizatzen. Orain BP-k 0.75 aldera dagoen balio bat itzultzen duenean ematen baita balio maximoa (1), beti erantzun luzeagoak lortzea eragotziz eta

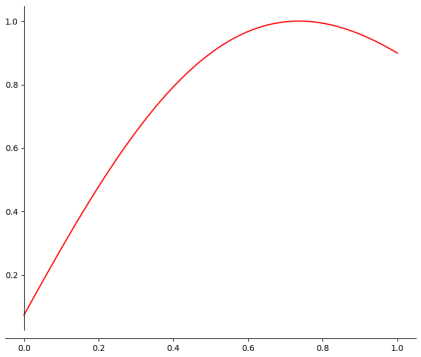

Figure 3: Rayleigh distribuzioaren kurba 4 ekuazioan adierazitako parametroak erabiliz.

motzagoak (probabilitate txikiagoarekin izan arren) onartuz.

4 Telegram zerbitzuara esleipena

Lan honen helburua elkarrizketarako sistema bat sortzea da, baina horretarako interfaze bat behar da. Hori Telegram-ek eskaintzen duen *Bot*en API a erabiliz egin da. Hain zuzen, Telegram-eko *Bot* bat inplementatu da sistemak emandako irteera eredu bat erabiliz bidalitako mezueei erantzunak emateko.

Bot honen inplementazioa *python-telegram-bot* Telegrameko API-aren inplementazioaren bidez egin da (github).

5 Datuak

Sare neuronalak ikasteko datuak behar ditu. Horretarako, filmen azpitituluak erabili nahi izan dira, baina, euskaraz, ez dago hauen datubase formalik, ondorioz, normalizazio prozesu bat egin behar izan zaie azpituluei erabili ahal izateko.

5.1 Azpituluen iturria

Azpituluak *OpenSubtitles* webgunetik hartu dira. Bertan hizkuntza eta film askotako azpituluak daude eskuragai.

Kasu honetan, euskarazko azpituluen datubasea erabili da, galdera-erantzun sistema euskarazkoa egin nahi delako.

Datu base honetan azpituluak lerroetan daude banatuta filmetan agertzen diren ordena berdinean.

5.2 Azpituluen normalizazioa

Film azpituluak ezin dira nolanahi erabili ikasketarako, izan ere, ez dakigu bertan zein esaldi izango diren galderak eta zeintzuk erantzunak. Gainera, azpituluetan karaktere asko soberan daude elkarriezketei dagokionez, adibidez, lerroen hasieran agertzen diren "-" motako karaktereak.

Ondorioz, zer nolako normalizazio egin behar den jakiteko, ikerketa bat egin da azpituluetan zehar. Honen emaitzak 5.2.1, 5.2.2 eta 5.2.3 ataletan aurkezten dira.

5.2.1 Karaktereen garbiketa

Aipatu bezala, karaktere asko elkarrizketetan ez dira esanguratsuak. Ondorioz, karaktere horiek ezabatu beharra daude. Ordea, ezabatu ahal izateko karaktere horiek zeintzuk diren jakin beharra dago.

Ikerketa egin ondoren, garbi gelditu da lerroen hasieran erabilgarriak ez diren zenbait karaktere daudela. Horiek ondorengoak dira: "-" (marra), "_" (azpi-marra) eta " " (espazioa). Ohiko marra karakterea erabiliena izan da, baina antzeko UTF-8 motako beste zenbait karaktere ere daude ("–", "—" eta "—").

Karaktere horiek zeintzuk diren jakinda, horien ezabaketa egin da.

5.2.2 Zenbait lerroko esaldien identifikazioa

Azpituluak ez daude galdera eta erantzun modura ordenatuta, izan ere, filmean esaldiak datozten moduan agertzen dira datubasean. Hau da, lerro bakoitiak ez dute zertan galderak izan eta bikoitiak erantzunak. Erabateko kaosa dago alde horretan, film batean pertsonaia batek zerbait esaten badu, beste pertsonaia bat ez dago erantzutera behartuta. Ez hori bakarrik, esaldi askok ez dute erantzunik behar edo esaldiak luzeegiak direnean, azpituluen egileak zenbait lerrotan banatzea erabaki dezake.

Egoera hori kontuan izanda, horri aurre egiteko zenbait neurri hartu dira:

- Zenbait egilek pertsonaia bakarrak esaten dituen esaldi oso luzeak banatzeko, " karakterea erabiltzen dute lerroen hasieran eta batzuetan bukaeran ere. Ondorioz, karaktere hori duten lerroak batu behar dira.

- " karakterearekin banatuta ez dauden zenbait lerro ere badaude, ordea. Horiek "," (koma) batekin daude zatituta, adieraziz, koma ondorengo lerroa esaldi berberekoa dela. Komekin banatuta dauden lerroak, orduan, baita ere, batu egin behar dira.

5.2.3 Galdera eta erantzunen sailkapena

Aurreko garbiketak egin ondoren, azpituluak sistemak ikasketarako onartzen duen sarrera formatuan

jartzea da egin behar dena. Hain zuzen, sistemak lerro bakoitzean bi elementu dituen fitxategi bat jasotzen du, lehenengo elementua galdera bat izanik eta bigarrengoa erantzuna, biak tabulazio batekin bananduta. Honela, lerro bakoitiak galderak bezala erabiltzen dira (lerroen lehen zatia) eta bikoitiak erantzun bezala (lerroen bigarren zatia).

Sailkapen hau, ordea, ez da guztiz zuzena, izan ere, oso zaila da tribialki lerro bakoitza galdera bat edo erantzuna den jakitea, ezin baitira kontrolatu azpitituluetan dauden salbuespen guztiak. Gainera, errore bat gertatzen denean, hurrengo galdera-erantzun guztiak gaizki sailkatzen dira beste akats bat dagoen arte.

6 Emaitzak

Erabilitako sistemak emaitza onak eta txarrak izan ditu. Horien aurkezpena 6.1 eta 6.2 azpi-ataletan egingo da.

6.1 Onak

```
Galdera:   Kaixo
Erantzuna: Kaixo ,

Galdera:   Zer moduz zaude?
Erantzuna: Ondo .

Galdera:   Egun on!
Erantzuna: Egun on , jauna .

Galdera:   Gabon
Erantzuna: Gabon .

Galdera:   Zorionak!
Erantzuna: Eskerrik asko .

Galdera:   Zorte on
Erantzuna: Eskerrik asko .
```

Baita ere, onak bezala onartu daitezkeen emaitzak daude, baina ez dira oso esanguratsuak elkarrizketei dagokionez.

```
Galdera:   Nor zara?
Erantzuna: Ez dakit .

Galdera:   Nor da zure aita?
Erantzuna: Ez dakit .
```

6.2 Txarrak

```
Galdera:   Arratsalde on!
Erantzuna: Egun on , jauna .

Galdera:   Non egin duzu mina?
Erantzuna: Ez .

Galdera:   Zer erosiko duzu?
Erantzuna: Ez .

Galdera:   Zertan zabiltza?
Erantzuna: Ez .

Galdera:   Arratsalde on!
Erantzuna: <unk> .

Galdera:   Zer egin duzu?
Erantzuna: Ez egin duzu? .
```

7 Analisia

Ikusi den bezala, nahiko emaitza onak lortzen dira ohiko esaldiei erantzuna emateko. Adibidez, agurrei normalean ondo erantzuten die, hauek izaten dituzten emaitzak, orokorrean, antzekoak izaten dira eta. Berdina gertatzen da zori ona opatzen denean, orokorrean, eskerrak ematen baitira erantzun bezala.

Ordea, erantzun oso desberdinak izaten dituzten galderei edo nahikoa ohikoak ez diren galderei ez zaie erantzun oso egokiak ematen. Izan ere, galdera bat asko agertzen den arren (adibidez, “Nor zara?”), erantzunak kasu bakoitzean oso desberdinak direnez, ez da gai emaitza garbi bat emateko eta nahikoa agertzen ez diren galderek ere ezin dute erantzun garbirik jaso ez baitago modurik erantzun egokiak ikasteko.

8 Ondorioak

Garbi dago emaitza hauek ez direla onak eta sistema hau, dagoen moduan, ezin dela produkzioan erabili. Itzultzen diren erantzunak sinpleegiak dira askotan eta multzo txiki bateko galdera bat egin ezean erantzun arraroak ematen dira. Hori gertatzearen arrazoia, zati batean sistemarekin dago erlazionatuta eta beste batean erabilitako datubasearekin, izan ere, erantzun motzak saihesteko erabilitako metodoa ez da erabili zitekeen hoberena. Gainera, datubasearen izaera kaotikoak arazo gehiegi ematen ditu ikasketa garaian.

Seguruenik, galera funtzioa aldatu beharko litzateke eta *Perplexity*-an oinarritutako beste bat erabili,

dagoeneko badaudelako hori erabiltzen duten ereduak (1). Horrez gain, gaur egun, mota honetako galdera erantzun sistema bat inplementatzeko ideia hoberena *Transformer* motako eredu bat erabiltzea litzateke, emaitza oso onak ematen ari baitira horrelakoak erabiltzen dituzten sistemek.

References

[1] D. Adiwardana and T. Luong, "Towards a conversational agent that can chat about...anything," 2020.

[2] J. Constine, "Facebook launches messenger platform with chatbots," 2016.

[3] "How chatbots are transforming wall street and main street banks? - marutitech.com," 2019.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," vol. arXiv:1409.3215, 2014.

[5] M. O. Anchordoqui, "Itzulpen automatikoa," 2020.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[7] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "Bleu: a method for automatic evaluation of machine translation," pp. 311–318, 2002.