Muturretik muturrerako solasaldi sistema

Maitane Martinez Eguiluz

Abstract

Proiektu honetan, dialogoa mantentzeko muturretik muturrerako bi sistema aurkezten dira. Lehenengoa, ingelesez garatu da, bigarrena, berriz, euskaraz. Ingelesezko sistema Telegram Bot batera egokitu da.

1 Sarrera

Proiektu honen ataza elkarrizketa mantentzen duen sistema bat sortzea da. Horretako, muturretik muturrerako sare errekurrentea da erabiltzen den sistema. Entrenatzeko datuak filmetako azpitituluak dira. Bi hizkuntzetan garatu da sistema: inglesez eta euskaraz. Gainera, Telegramen Bot bezala egokitu da inglesez entrenatua izan den sistema. [1]

2 Erlazionatutako lanak

Banadanu-k eta beste batzuek neurona sareetan oinarritutako itzulpen automatikoa proposatu zuten. Itzulpen tradizionalak ez bezala, kodetzaile-deskodetzaile motako modeloa proposatzen dute ingelesetik frantseserako itzulpena egiteko. Kodetzaile sare neuronalak iturriko esaldia irakurri eta luzera finkoko bektore batean kodetzen du. Dekodetzaileak kodetutako bektorearen itzulpena egiten du. (Bahdanau, 2014)

Lison-ek eta beste batzuek, artikulu horretan, azpitituluen aurreprozesazioan eta lerrokatzean hobekuntza ugari proposatzen dituzte, hala nola OCR akatsak automatikoki zuzentzea. Gainera, metadatuak erabiltzen dituzte azpitituluen kalitatea estimatzeko eta azpititulu bikoteak ebaluatzeko. (Lison, 2016)

[1] Lanaren GitHub helbidea: https://github.com/MaitaneMartinez/DIAL-

3 Sistema

Atal honetan muturretik muturrerako solsaldi sistema azaltzen da. Sistema hau Bahdanau-k eta beste batzuek proposatutako lanean oinarritzen da (Bahdanau, 2014) . Horretarako, dialogoa itzulpen ataza bat bezala proposatzen da. Esate baterako:

- Itzulpen automatikoa :
  - Sistemaren sarrera: Egun on guztioi
  - Sistemaren irteera: Buenos días a todos
- Dialogoa:
  - Sistemaren sarrera: Egun on guztioi
  - Sistemaren irteera: Baita zuri ere

Sistema kodetzaile-deskodetzaile motako modeloa da. Kodetzaileak iturriko esaldia irakurtzen du bi norabideko sare errekurentearen bidez eta luzera finkoko bektore batean kodetzen du. Deskodetzaileak, kodetutako bektorea erabiliz, esaldi berria sortzen du.

3.1 Kodetzailea

Kodetzaileak, sarrerako esaldia irakurtzen du tokenen bektorea bezala, x $= (x_1, \cdots, x_{T_x})$. Horretako, bi norabideko sare errekurrentea erabiltzen da. BiRNN sarea aurreranzko eta atzerako RNN-z osatutako dago. Aurreranzko RNN-a $\overrightarrow{f}$, ezkerrerik eskuinera irakurtzen du esaldia eta aurreranzko geruza ezkutua kalkulatzen du $(\overrightarrow{h_1}, \cdots, \overrightarrow{h_{T_x}})$. Atzerazko RNN-ak $\overleftarrow{f}$, eskuinetik ezkerrera irakurtzen du esaldia eta atzerazko geruza ezkutua kalkulatzen du $(\overleftarrow{h_1}, \cdots, \overleftarrow{h_{T_x}})$.

Hitz bakoitzeko ohar bat lortzen da aurrerazko geruza eskutua $\overrightarrow{h_j}$ eta atzeranzko geruza ezkutua $\overleftarrow{h_j}$ konkatenatuz: $h_j = \left[\overrightarrow{h_j}^T ; \overleftarrow{h_j}^T\right]^T$. Horrela, oharrak h_j aurreko eta ondorengo hitzen informazioa edukiko du. oharrek testuinguru bektorea osatzeko balio dute.

3.2 Atentzio geruza

Testuinguru bektorea c_i oharren h_i batuketaren bidez kalkulatzen da:

$$C_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \quad (1)$$

ohar bakoitzaren h_j pisua α_{ij} honela kalkulatzen da:

$$\alpha ij = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})}, \quad (2)$$

non

$$e_{ij} = a(s_{i-1}, h_j) \quad (3)$$

s_{i-1} deskodetzailearen RNN eskutuzko geruza da (5) eta j sarrerako esaldiaren oharra h_j da. a lerrokatze-eredua entrenatua izan da beste osagaiekin batera.

3.3 Deskodetzailea

Deskodetzailea, testuinguru bektorea c eta aurretik iragarritako hitza y_{i-1} emanik, hurrengo hitza $y_{t'}$ iragartzeko entrenatzen da. Horretarako, probabilitate bektorea sortzen da:

$$p(y_i|y_1, \cdots, y_{i-1}) = g(y_{i-1}, s_i, c_i) \quad (4)$$

non s_i RNN-ko eskutuzko geruza den, i denboran:

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \quad (5)$$

f eta g funtzio ez linealak dira. Hiru dekodetze estrategia desbedin erabiltzen dira: top1, topk eta multinomial. top1 estrategiak ikasitako token posibleena itzultzen du. topk estrategiak k erantzun posibleenetatik bat itzultzen du. Sistema honetan k=3 definituta dago. Azkenik, multinomial estrategia, banatze nominala erabiltzen du tokenak aukeratzeko.

Telegram Bot-a sortu da GitHub dauden jarraibideetan oinarrituta [2]. Ingelesezko sistema erabiltzen da eta aldagai global baten bitartez deskodetze estrategia alda daiteke, aldagai honi balio desberdinak esleituz.

[2]Your first Bot: https://github.com/python-telegram-bot/python-telegram-bot/wiki/Extensions-%E2%80%93-Your-first-Bot

4 Datuak

Atal honetan, erabilitako corpusak eta datuak azaltzen dira. OpenSubtitles corpusa erabiltzen da bi sistemak sortzeko [3]. Corpus honetan 92 hizkuntzatako azpitituluak daude. Gure kasuan, ingelesezko eta euskarazko azpitituluak erabili dira soilik.

Lehenengo sisteman, inglesezko filmen esaldiak erabiltzen dira. Esaldiak binaka elkartzen dira tabulazio baten bidez banatuta: bigarren esaldia lehenengoaren erantzuna litzateke. Guztira, milio bat esaldi pare daude. Baina, batzuetan galdera eta erantzuna ez datoz bat, adibidez:

- Galdera: Friends.
- Erantzuna: Good.

Bigarren sisteman [4], euskarazko filmen esaldiak erabiltzen dira. Esaldi horiek binaka jartzen dira, ingelesezko egiturari jarraituz. Gainera, garbiketa bat aplikatzen da esaldietan. Garbiketan, dialogo "-" ikurra eta puntu finala "." kentzen dira.

5 Emaitzak

Atal honetan, esperimentuak eta horien emaitzak azaltzen dira. Horretarako, ingeleserako entrenatu izan den muturretik muturrerako solasaldi sistema probatzen da. Hiru froga desberdin egin dira, dekodetze estrategia desberdinak erabiliz.

Elkarrizketa sinple bat erabili da probak egiteko:

- Hello
- How are you?
- So, what are your plans for this weekend?
- Bye

Lehenengo sarreratik hirugarrenera zailtasuna handitzen dira. Azken sarrera erabiltzen da sistema amaitzeko. Honako erantzun desberdinak hauek lortu dira:

[3]OpenSubtitles web gunea: http://opus.nlpl.eu/OpenSubtitles-v2018.php

[4]Sitema deskargatzeko helbidea: https://drive.google.com/drive/folders/1iO9cLB1rsKBMUbPrQzJTbw2J308WjaEW?usp=sharing

top1

- Hello ,
- I ' m fine .
- I ' m not .
- Bye ,

Beti, erantzun berdinak itzultzen dira. Erantzunak laburrak dira eta, hirugarren galderan, berriz, sistemak ez du sortzen erantzun egokirik. Gainera, puntuazio marka ez zuzenak erabiltzen ditu.

topk

- I , hello .
- Fine , fine , thanks , man ! .
- I ' s . to go to bed ? ?
- I , you

Erantzun desberdinak itzultzen ditu. Aurreko estrategiarekin konparatuz, erantzunak luzeagoak dira, baina zentzu gabeko puntuazio markak itzultzen ditu.

multinomial

- Hello you .
- I ' m fine
- Yes .
- Good bye

top1 estrategiaren antzeko erantzunak itzultzen ditu. Kasu honetan, puntuazio markak egokiagoak dira.

Ingelesez erabilitako elkarrizketa euskaratu da bigarren sistema testeatzeko:

- Kaixo
- Zer moduz zaude?
- Zer egingo duzu asteburuan?
- Agur

Honako erantzun hauek bueltatzen ditu sistemak:

top1

- Zer moduz?
- Ondo
- $< unk >$
- Agur

Hirugarren galderarentzat ez dauka erantzun posiblerik.

topk

- Kaixo,
- Oso ondo,
- Ez dakit zer
- Agur esan

Erantzun posibleenetatik aproposenak eta top1 erantzunekin desberdinak direnak aukeratu dira.

multinomial

- Kaixo
- Oso ongi
- Ez dut
- Agur

topk bezala, erantzun guztietatik aproposenak aukeratu dira. Hiruretan antzeko erantzunak itzultzen ditu sistemak.

Azkenik, 1. irudian ikusi daiteke Telegram Bot-a erabiliz mantendu den elkarrizketa.

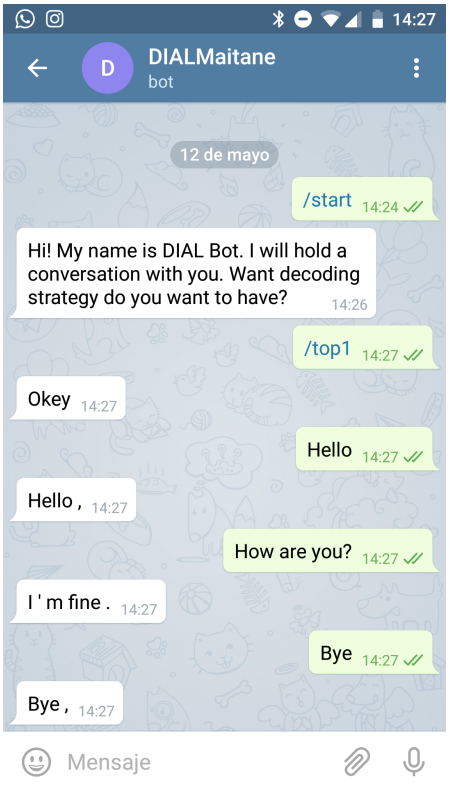

Figure 1: Telegram Bot-arekin izandako elkarrizketa

/star bidez dialogoa hasiko da eta dekodetze entrategia desberdinak erabili daitezke elkarrizketan zehar "/" aurretik jarrita. Ingelesezko sistemak dituen arazo berdinak edukiko ditu Telegrameko Bot-ek.

6 Analisia

Lehenengo sisteman lortutako emaitzak ikusita, esan daiteke hirugarren galderaren erantzunak ez direla zentzuzkoak. Hau da, sistema ez da gai galdera honi erantzuna emateko. Gainera, multinomial estrategian izan ezik, beste bietan puntuazio marka desegokiak erabiltzen ditu. Hori gertatzen da garbiketa sakon bat falta delako puntuazio marketan. Gainera, esaldi laburrak itzultzen ditu sistemak; izan ere, token ohikoenek pisu handia edukiko dute eta horiek bakarrik itzultzeko joera edukiko du sistemak. Horien artean, puntuazio markak daude.

Bigarren sisteman, puntuazio marken arazoak konpondu egin dira erantzun gehienetan. Gainera, hirugarren galderarentzat topk estrategia erabiliz erantzun zuzen bat itzultzen du. Baina, gutxi agertzen diren hitzak ez ditu ikasten, adibidez, izen propioak. Horren ordez, $< unk >$ motako tokenak itzultzen ditu. Aproposa izango litzateke euskaraz entrenatutako tokenizatzailea erabiltzea, baita normalizatzailea sistema hobeto funtzionatzeko.

Sistemak ez du dialogo bat mantentzen, galdetutakoari baino ez dio erantzuten, aurreko elkarrizketa kontuan hartu gabe.

7 Ondorioak

Proiektu honetan, bi sistema aurkezten dira. Biak muturretik muturrerako solasaldi sistemak dira, baina lehenengoa ingelesezko filmen azpitituluekin entrenatua eta bigarrena euskarazkoekin. Gainera, Telegram Bot-a sortu da ingelezko sistema erabiliz. Sistema horiek hiru deskodetze entrategia erabiltzen dituzte: top1, topk eta multinomiala.

Etorkizuneko lanei begira, interesgarria litzateke euskarazko sistema entrenatzea elkarrizketen bidez bakarrik. Datu horietan, galdera-erantzun motako egiturak errespetatzen dira. Beraz, nahiz eta datu kopurua txikiagoa izan, gerta daiteke erantzun aproposagoak itzultzea.

Bi hizkuntzatako sistema sortzea izango litzateke beste proposamen bat. Entrenatzeko datuetan errespetatuko beharko litzateke galderaren berdina izatea erantzunaren hizkuntza. Beraz, testeatzeko orduan, sistemari hitz egiten diozun hizkuntza errespetatu beharko luke eta ez nahastu.

References

Cho K. Bengio Y. Bahdanau, D. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Tiedemann J. Lison, P. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.