eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

Virtual Patient

Intentsioen Karakterizazioa

Egilea: *Julen Fuentes Aguirre*

IXA Ikerketa Taldea

2022ko abuztuaren 11

Gaien Aurkibidea

Portada **2**

1 Sarrera **2**

2 Eginkizunak **3**
2.1 Eskainitako bibliografiaren azterketa 3
2.2 Datuen aurreprozesamendua 3
2.3 “intent”-en hierarkia sortu 4
2.4 “LabForSim” korpuseko galderak “intent”-en arabera sailkatu 5
2.5 Sailkatzaile bat sortu eta ebaluatu 5

3 Ondorioak **6**

4 Bibliografia **7**

1 Sarrera

COVID-ak eragindako ezohiko egoera dela eta, elkarrizketa sistema automatikoen teknologiarenganako interesa dezente ugaritu da, medikuntzaren alorrean besteak beste. Hori dela eta, proiektu honen helburua honako alorrak batzen dituen teknologia garatzea izango da: Hizkuntza naturala, gizakiok darabilguna, alegia; laguntzaile birtualak; medikuntza alorra eta "machine learning" aurreratua.

Aipatu berri diren alorrak lantzeko elkarrizketa sistema bat osatuko da gaixo eta mediku baten artean, non gaixoaren papera RASA tresnaren bitartez osatutako "chatbot" batek hartuko baitu. "Chatbot"-ak elkarrizketak arazorik gabe aurrera eraman ahal izateko, medikuek egin ditzaketen galderak haien "intent" edo asmoaren arabera sailkatu beharko dira eta galdera horietatik informazio esanguratsuena erauzi beharko da eskuratutako informazioaren araberako zentzuzko erantzun bat hizkuntza naturalean emateko.

Proiektua osatzeko 150 orduko epea eman da (behin lan epea osatuta, eta betiere hala tutoreak nola ni neu adostasun batera heltzekotan, epea luzatu egin daiteke) eta orduak 6 astetan zehar banatu dira (25h aste bakoitzeko eta 5h egunero). Izandako denbora apur bat eskasa izan denez, ezin izan da proiektu osoa osatu. Hala eta guztiz ere, lortutako emaitzak onargarriak izan dira.

2 Eginkizunak

2.1 Eskainitako bibliografiaren azterketa

1. astea bereziki eskaini didaten informazio guztia aztertzeko erabili da:

- RASA tresnarekin egiten dena hobeto ulertzeko *tutorial generikoa* [1]
- *RASAren dokumentazioa* [2]
- *Rasa Masterclasses - YT* [3]
- *Espresio Erregularrak eta Sinonimoak* [4]
- *Detecting Similarity in Questions - GitHub* [5]
- Proiektu Adibideak - RASA
  - *Conversational AI with RASA* [6]
  - *Using Furhat and Rasa to Assist when You Forget a Word Mid-Sentence: A Student Group Project* [7]
  - *RoboCafe* [8]

Honetaz gain, *Rasa Masterclasses - YT* bideo-sorta ikusi den bitartean, *apunteak.txt* [10] (biltegiko **/Bestelako Dokumentu Lagungarriak** atalean topa daiteke) fitxategia betetzen joan da, RASAren instalazioa osatu da (python bertsioaren arabera arazoak ematen ditu; horregatik Python 3.7 edo 3.8 bertsioak erabiltzea gomendatzen da) eta bideo-sortarekin eskuratutako gaitasunak apur bat trebatu dira ("intent", "action", "story", "custom actions", "rules" ...) honako proiektu laburrak osatuz: *Chitchat* eta *Medicare Locator* [10] (biltegiko **/Froga Proiektu Laburrak** atalean daude gordeta).

2.2 Datuen aurreprozesamendua

Atal honetan aipagarria da esatea mediku eta gaixoen arteko elkarrizketa ereduak aurkitzeko zailtasunak izan direla Datu Babesaren Lege Organikoa (DBLO) dela eta. Horregatik, aurkitu diren korpusak, *LabForSims* eta *Campillos*, alegia, frantsesez soilik aurkitu dira [10] (biltegiko **/Korpusak** atalean daude eskuragarri).

Hasteko, *LabForSims* korpusa lantzen hasi da eta eman den lehendabiziko urratsa korpuseko *doctor_fr.txt* fitxategiko esaldi guztiak itzultzea izan da (jatorrizko bertsioa ezabatu gabe), *DeepL* [9] itzultzaileaz baliatuta. Izan ere, fitxategi honek medikuak egin ditzakeen galdera ugari ditu eta gure "chatbot"-a prest egon behar da edonolako galderen aurrean erantzun zentzudun bat emateko. Esaldien itzulpenak *doctor_es.txt* fitxategian gorde dira.

Ondoren, korpus bereko *dialogues* katalogoa hartu eta honek dituen 41 elkarrizketa adibideak honako hiru multzo hauetan banatu dira: *train* (elkarrizketen %70a; 29 fitxategi), *dev* (elkarrizketen %15a; 6 fitxategi) eta *test* (elkarrizketen %15a; 6 fitxategi). Ataza hau burutzeko *train_test_dev_osatu.py* "script"-a erabili da.

dialogues datu-sorta hiru multzotan banatu ostean, hurrengo ataza *doctor_fr.txt* fitxategiko esaldiak (itzuli gabeko bertsioa, *dialogues* datu-sortako esaldiak itzuli gabe baitaude) zein multzotan dauden bilatzea izan da. Proba arin batzuk eginez esaldi kopuru esanguratsu bat ez dela aurkitzen ikusi da. Gainera, esaldi bat elkarrizketaren batean dagoen bilatzerako orduan, amaierako galdera ikurrak arazoak ematen ditu eta esaldia ez da aurkitzen nahiz eta berdin idatzi. Horregatik, aipatutako arazoak konpontzeko eta denbora aurrezteko "bashscript" bat osatzea erabaki da, *train_test_dev_sailkapena.sh* deiturikoa. Egindako sailkapena *emaitza.txt* fitxategian gorde da.

Behin *LabForSims* korpuseko datuen aurreprozesamendua amaituta, *Campillos* korpusa lantzeari ekin zaio. Lehenik eta behin, korpus honetako elkarrizketetan hizki bitxi batzuk daudela ikusi da eta hori fitxategien kodeketagatik gertatzen da. Hori dela eta, elkarrizketa guztien kodeketa era automatikoan aldatzeko *kodeketa_aldatu.sh* "bashscript"-a erabili da. Korpus honetako aurreprozesamenduari amaiera emateko, elkarrizketa guztiak frantzesetik gaztelaniara itzuli dira.

2.3 "intent"-en hierarkia sortu

"intent"-en hierarkia osatzeko lehenengo urratsa era honetako atazetan, eta medikuntza arloan bereziki, sailkapena egiteko erabiltzen diren irizipideen azterketa bat egitea izan da. "intent"-en erauzketari buruzko informazio apur bat bildu ostean, proiektuko "intent"-en behin-behineko lehenengo hierarkia proposamenari ekin zaio, *intent.txt* fitxategian ikus daitekeen moduan [10] (biltegiko **/Intent-en Sailkapena** atalean dago ikusgai). Hurrengo urratsa *LabForSims* korpuseko *doctor_es.txt* fitxategiko esaldiak irakurtzea izan da,

esaldi horien arabera ager daitezkeen "intent"-ak antzemateko eta horri esker sailkapen hobea lortu da. Sailkapen hau osatu ondoren, tutoreekin bilera bat izan da eta han osatutakoa proposatu eta agertutako zalantzak argitu dira. Behin zalantzak argituta eta beharrezko moldaketak eginda, tutoreen oniritzia lortu da. Beraz, honako hau izango da behin-betiko "intent"-en sailkapena: *nlu.yml* [10] (biltegiko **/Intent-en Sailkapena** atalean aurki daiteke)

2.4 "LabForSim" korpuseko galderak "intent"-en arabera sailkatu

LabForSims korpuseko *doctor_es.txt* fitxategiko galderak sailkatzeko orduan, lehendabiziko esaldiak *nlu.yml* fitxategiko "intent"-en hierarkia osatu den hein berean sailkatu dira ager daitezkeen esaldi motak antzemateko. Behin hierarkia sendo bat osatuta, gainontzeko esaldiak banan-banan sailkatzen joan dira, bakoitzak duen asmoaren arabera.

2.5 Sailkatzaile bat sortu eta ebaluatu

Azkenengo fase honetan, *LabForSims* korpuseko *dialogues* datu-sorta hiru multzotan banatzerakoan lortutako *emaitza.txt* fitxategia oinarri hartuta RASArekin "chatbot"-aren lehenengo hurbilketa osatzen hasi da "train" multzoan dauden esaldiekin.

Honen ostean, "chatbot"-ak antzeman dezakeen "intent" bakoitzeko eman beharko lukeen erantzuna zehaztu da "rules" batzuen bitartez (elkarrizketan zehar aurretik aipatutakoa kontutan hartu gabe "intent" jakin bat antzematen denean, horri dagokion erantzuna itzuliko da). Kasu honetan, hau "chatbot"-aren oinarria besterik ez denez, "intent" bakoitza antzematen denean "intent"-a bera itzultzea erabaki da, lortutako emaitzen baliagarritasuna aztertzeko eta RASA tresnarekin trebetasuna hartzen joateko.

Azkenik, froga azkar bat egin nahi izan denez, sailkatzaile bat sortu da ezarpen lehenetsiekin eta honen zehaztasuna neurtu da "chatbot"-arekin elkarrizketa bat izanez, ezen "dev" multzoko esaldiak erabili baitira. Honekin lortu nahi dena sailkatzaileak nola iragartzen duen aztertzea da eta egiten den ebaluaketa bakoitzaren ondoren, beharrezkoak diren aldaketak egitea da, baina betiere "dev" multzoko esaldietara gehiegi ez doituz edo "overfitting" deritzona saihestuz.

3 Ondorioak

Ikus daitekeenez, sortu den sailkatzailearekin lortutako emaitzak ez dira oso onak izan. Hala eta guztiz ere, izandako denbora apur bat eskasa izanik eta proiektu honen bitartez landu nahi ziren gaitasunetan trebetasuna lortu dela ikusita, proiektu honekin lortutako emaitzak onargarriak direla esan daiteke. Halaber, aipatu beharra dago tutoreekin proiektuarekin jarraitzea adosteko aukera dagoela eta hala adosten bada, "intent"-en hierakia findu eta sailkatzaile hobeagoa lor daiteke, hala beharrezko ukituak emanez nola froga gehiago eginez.

4 Bibliografia

[1] Ivan Vulić (PolyAI & University of Cambridge) Paweł Budzianowski, Inigo Casanueva. Data Collection and End-to-End Learning for Conversational AI. URL http://poly.ai/wp-content/uploads/2019/11/EMNLP19-PolyAI-Tutorial.pdf.

[2] Rasa Technologies GmbH. Rasa Playground. URL https://rasa.com/docs/rasa/playground/.

[3] Rasa. Rasa Masterclass: Developing Contextual AI assistants with Rasa tools. URL https://www.youtube.com/watch?v=rlAQWbhwqLA&list=PL75e0qA87dlHQny7z43NduZHPo6qd-cRc.

[4] Rasa Technologies GmbH. NLU Training Data. URL https://rasa.com/docs/rasa/nlu-training-data/.

[5] Arpan Mukherjee eta Prabhat Kumar. Detecting Semantic Similarity in Questions. URL https://github.com/arpanmukherjee/Detecting-Semantic-Similarity-in-Questions.

[6] Xiaoquan Kong. Conversational AI with Rasa. URL https://github.com/PacktPublishing/Conversational-AI-with-RASA.

[7] Angus Addlesee. Using Furhat and Rasa to Assist when You Forget a Word Mid-Sentence: A Student Group Project. URL https://cutt.ly/LZMmJ9k.

[8] @carlobee (GitHub). RoboCafe. URL https://github.com/carlobee/RoboCafe.

[9] DeepL. DeepL Translate - El mejor traductor del mundo. URL https://www.deepl.com/translator.

[10] Julen Fuentes Aguirre. Virtual-Patient–Intentsioen-Karakterizazioa-. URL https://github.com/955750/Virtual-Patient--Intentsioen-Karakterizazioa-.