CUSTOMER SUPPORT CHATBOT

Jelena Cupać sw13/2017, Aleksa Goljović sw14/2017, Fakultet Tehničkih Nauka, Novi Sad

Uvod

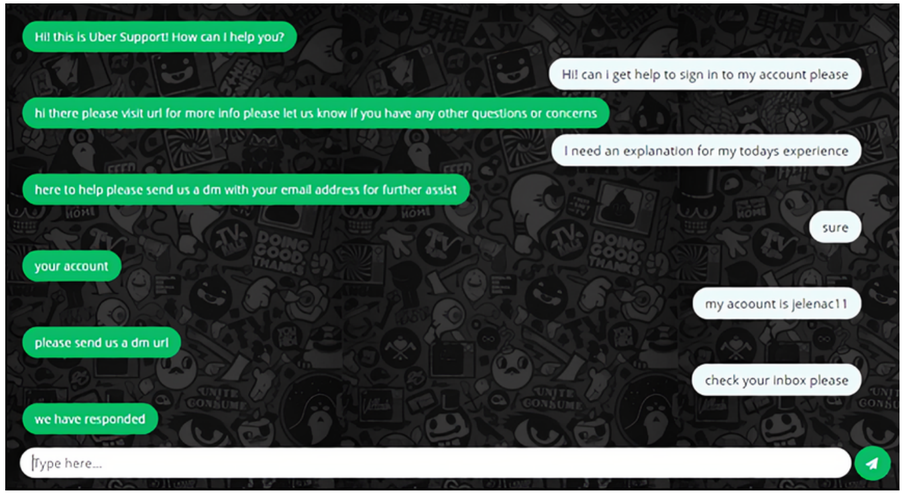
Cilj ovog projekta je davanje smislenih odgovora na korisnikove poruke u realnom vremenu. Korisnik ne treba da bude svestan da razgovara sa chatbot-om, već sa pravom osobom. Motivacija za ovaj projekat je bila to što je danas chat najučestaliji vid komunikacije. Ovaj vid komunikacije je dinamičniji od npr. email komunikacije. Takođe, korisnici će pre izabrati chat kao način da dođu do dodatnih informacija.

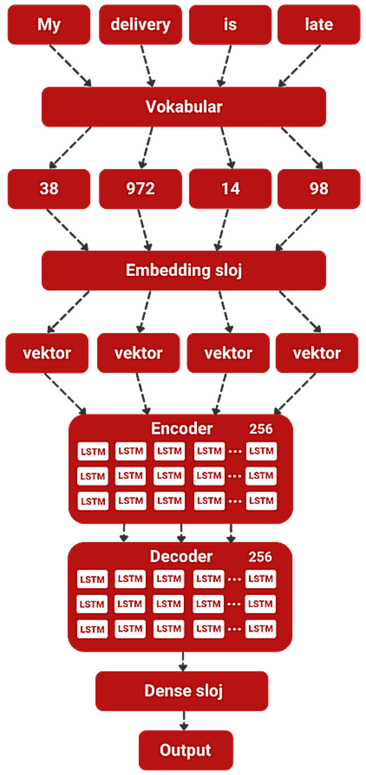
Metodologija

Za rešavanje ovog problema koristili smo enkoder-dekoder arhitekturu. Enkoder i dekoder se sastoje od po jednog RNN-a. Mrežu smo obučavali kroz 40 epoha. Koristili smo Adam optimizer sa learning rate-om 0.001. Veličina vokabulara iznosi 20000.
Reči u rečenici su predstavljene brojem koji se koristi u modelu prilikom obrade rečenice. Reči zajedno sa brojevima koji ih određuju su smešteni u vokabular. Vokabular je kreiran pomoću keras biblioteke. Za reči je određeno 1024 reči koje se koriste zajedno sa datom rečju u rečenici (kontekst). Obe mreže imaju po tri sloja od 256 LSTM ćelija. Izlaz dekodera je ulaz u dense sloj, gde imamo različit broj ćelija u zavisnosti od dataset-a koji je povezan sa 20000 ćelija. Ovaj sloj formira output. Mreže su implementirane pomoću TensorFlow i TensorLayer biblioteka. Slika arhitekture naše mreže može se pogledati na slici 3.

Rezultati

Kako na ista pitanja može biti odgovoreno na različite načine, za validaciju je korišćena samostavlna validacija i loss. Loss koji smo dobili na kraju je 0.3 za Uber, 0.5 za Apple i 0.6 za Amazon.





Skup podataka

Skup podataka koji je korišćen u projektu je Customer Service Support. Ovaj skup podataka broji preko 3 miliona tvitova i odgovara na iste u vezi sa korisničkom podrškom najvećih brendova. U skupu podataka se nalaze podaci za različite brendove poput Apple-a, Amazon-a, Uber-a, Delte, SpotifyCares-a itd. Podatke koje smo odabrali za treniranje naše mreže jesu podaci koji se odnose na korisničke podrške Apple-a, Amazon-a i Uber-a. Podatake za prethodna dva navedena brenda ima najviše, pa smo se zato odlučili baš za njih.

Pretprocesiranje podataka

U okviru pripreme podataka za treniranje neuronske mrežu, uradili smo sledeće korake:

- Prebacivanje reči u lowercase
- Zamena slenga i skraćenih oblika reči (poput faq, btw, gn, don't...)
- Uklanjanje tagovanja korisnika tzv. mention-a
- Uklanjanje heštagova
- Uklanjanje emodžija (😀, 🇷🇸, 🇺🇸) i emotikona (xD, :-D, :-P)
- Uklanjanje brojeva
- Uklanjanje reči koje ne pripadaju engleskom jeziku i kako bi mreža lakše učila,
- neispravno napisane reči su zamenjene korektno napisanim

Biblioteke koje su korišćene u ovom delu pretprocesiranja podataka su nltk, spellchecker, pyspellchecker, emot i keras.

Skup podataka je podeljen na 90:10, gde je 90% podataka iskorišteno za trening, a 10% podataka za test.

Zaključak

Mreža uglavnom smisleno odgovara na postavljena pitanja vezana za korisničku podršku. Problem koji postoji jeste što duge konverzacije nisu moguće i što mreža nije sposobna da odgovara na pitanja koja nisu usko povezana sa korisničkom podrškom.

Dalja unapređenja ovog projekta se mogu realizovati obučavanjem mreže i za davanje odgovara vezanih za druge korisničku podršku drugih brendova. Takođe, moguće je još optimizovati korišćene hiperparametre i koristiti neki attention mehanizam kao što je npr. Luong Style Attention Mechanism.