CHATBOT

Soft Computing 2019

Fakultet tehničkih nauka, Novi Sad,
Leona Nedeljković SW14-2016, Sonja Jošanov SW72-2016

Šta je Chatbot?

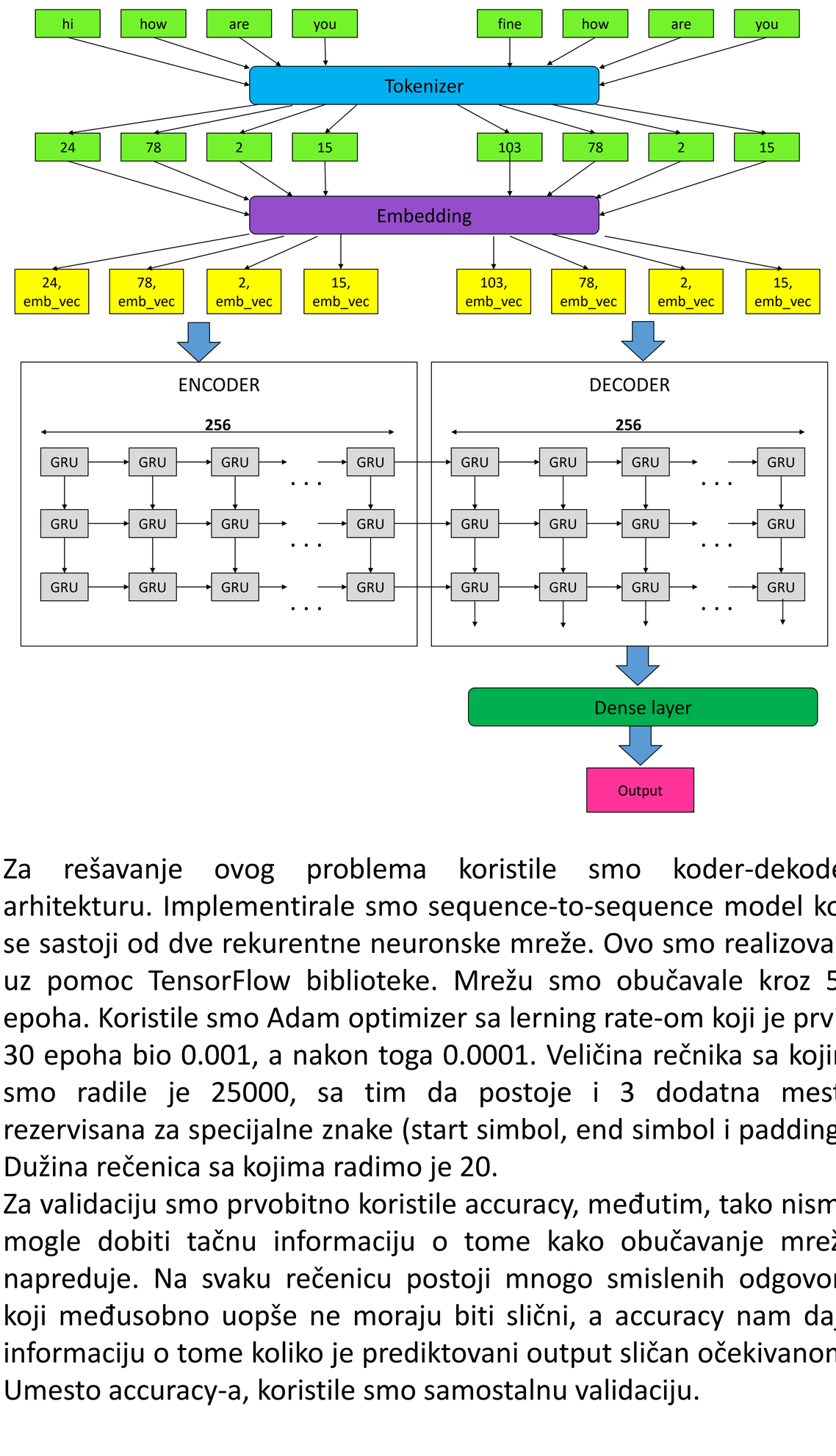
Chatbot je aplikacija koja je imala za cilj da daje smislene odgovore na rečenice koje zadaje korisnik. Naš cilj je bio da kreiramo i obučimo model koji će smisleno komunicirati sa korisnikom, tako da korisnik ima osećaj kao da razgovara sa čovekom.

Dataset

Kao dataset koristile smo komentare sa sajta Reddit, koji su objavljeni juna 2013. godine. Imale smo za ideju da koristimo postove koji imaju dobar "score" i najbolje odgovore na te postove.

Obzirom da zbog rečnika koji se koristi na internetu ovaj dataset nije savršen, mi smo izdvojile samo rečenice i reči koje će nam koristiti za obučavanje mreže. Izbacile smo rečenice u kojima se pominju linkovi, kao i znakove interpunkcije, brojeve i simbole koji nisu ASCII. Takođe smo sva slova svele na mala radi lakšeg obučavanja.
Kako bismo dodatno olakšale obučavanje mreže, neke nepravilno napisane reči smo pokušale da korigujemo pomoću NLTK biblioteke.

Neuronska mreža



Za rešavanje ovog problema koristile smo koder-dekoder arhitekturu. Implementirale smo sequence-to-sequence model koji se sastoji od dve rekurentne neuronske mreže. Ovo smo realizovale uz pomoc TensorFlow biblioteke. Mrežu smo obučavale kroz 50 epoha. Koristile smo Adam optimizer sa lerning rate-om koji je prvih 30 epoha bio 0.001, a nakon toga 0.0001. Veličina rečnika sa kojim smo radile je 25000, sa tim da postoje i 3 dodatna mesta rezervisana za specijalne znake (start simbol, end simbol i padding). Dužina rečenica sa kojima radimo je 20.
Za validaciju smo prvobitno koristile accuracy, međutim, tako nismo mogle dobiti tačnu informaciju o tome kako obučavanje mreže napreduje. Na svaku rečenicu postoji mnogo smislenih odgovora koji međusobno uopše ne moraju biti slični, a accuracy nam daje informaciju o tome koliko je prediktovani output sličan očekivanom. Umesto accuracy-a, koristile smo samostalnu validaciju.

Svaka reč u rečniku je predstavljena jedinstvenim brojem i prilikom obrade rečenica u modelu se koriste ovi brojevi. Tokenizer ima ulogu da za svaku prosleđenu reč pronađe njenu brojčanu vrednost.
Za svaku reč iz rečnika je određen kontekst, odnosno, vektor velićine 1024 koji sadrži reči (susede) koje se koriste zajedno sa posmatranom reči u rečenici. Embedding sloj ima ulogu da za svaku prosleđenu reč prosledi i vektor koji sadrži informaciju (decimalan broj) o tome koliko je svaki sused posmatrane reči "udaljen" od nje.
Ove vrednosti se prosleđuju kao inputi u koder i dekoder. Obe mreže imaju po tri sloja sa po 256 GRU ćelija.
Output dekodera predstavlja input za dense sloj. Za svaku rečenicu koja se formira, u ovom sloju postoji 20 ćelija koje su u potpunosti povezane sa 25003 ćelije koje predstavljaju reči iz rečnika. Na osnovu izračunatih vrednosti, u ovom sloju se formira output.

Rezultati

Obučavanje mreže je trajalo 3 dana. Loss koji smo na kraju dobile je 0.025. Neke od rezultata koje smo dobile možete videti na slikama.