Automatsko prevođenje

Vladimir Vuksanović

Matematički fakultet

1 Uvod

Automatsko prevođenje podrazumeva prevođenje teksta sa jednog prirodnog jezika na drugi korišćenjem kompjutera. Dakle, potrebno je za ulaznu recenicu proizvoljne duzine na izvornom jeziku generisati izlaznu recenicu, ne nuzno iste duzine na ciljnom jeziku. Standardan pristup za rešavanje ovog problema je korišćenjem enkoder-dekoder arhitekture opisane u [3] i [2]. Ova arhitektura u praksi daje veoma dobre rezultate i koristi je jedan od najpoznatijih servisa ovog tipa, Google Translate [4]. Osnovna ideja je da enkoder

2 Implementacija

2.1 Pretprocesiranje

Podaci za treniranje se sastoje od parova ulaznih i izlaznih recenica koje je potrebno na razlicite nacine obraditi pre njihove upotrebe. Zajednicko za obe je eliminisanje specijalnih simbola jer oni nisu korisni prilikom prevodjenja.

Kako bi recenice mogle da se upotrebe kao ulaz neuronske mreze potrebno je predstaviti ih kao niz brojeva. To se radi jednostavnim pridruzivanjem broja svakoj reci iz ulaznog skupa ili nekog predefinisanog recnika u kom slucaju se dodaje i token za nepoznate (out-of-vocabulary) reci, ovaj proces se naziva tokenzacija. Recenice se potom dopunjavaju "praznim"tokenima dok ne postanu iste duzine.

Izlazne recenice imaju nekoliko koraka vise. Potrebno je napraviti dva tipa izlaznih recenica, jedan tip predstavlja ulaz u dekoder a drugi predstavlja izlaz. Prva rec koju dekoder treba da dobije je pocetni (start of sequence) token i on se dodaje na pocetak svake recenice koja ulazi u dekoder. Analogno, na recenice koje predstavljaju izlaz iz dekodera potrebno je dodati zavrsni (end of sequence) token. Ostatak postupka je isti kao za ulazne recenice.

2.2 Enkoder

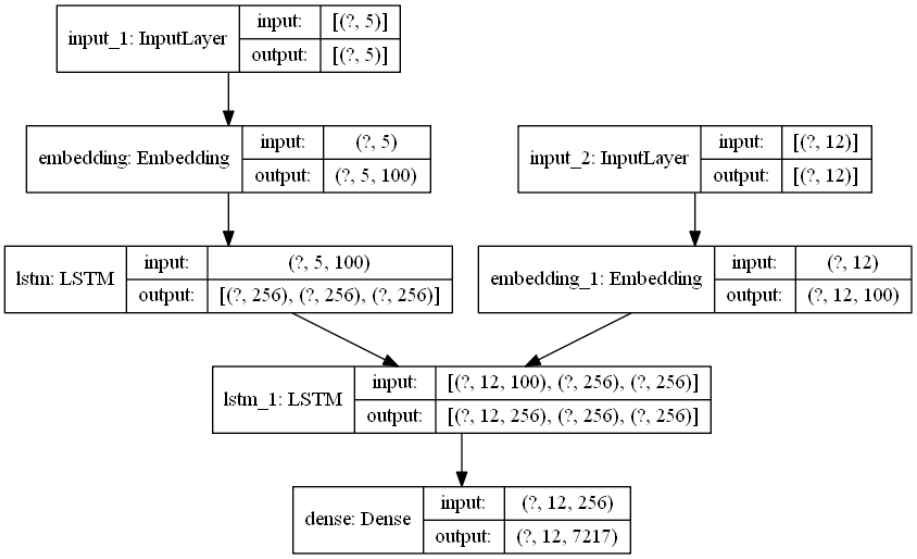
Enkoder je zaduzen da na osnovu dobijene sekvence generise reprezentaciju fiksne duzine koja bi trebalo da sadzi sve informacije potrebne da se konstruise prevod na ciljnom jeziku.

2.3 Dekoder

Dekoder koristi svoje unutrasnje stanje i prethodno generisanu rec kako bi odredio sledecu rec u recenici. Inicijalno se kao unutrasnje stanje koristi izlaz iz enkodera, a kao prethodno generisanu rec uzimamo pocetni (<sos >) token koji je uveden prilikom pretprocesiranja. Dekodiranje se prekida ili dok se ne dobije zavrsni (¡eos¿) token ili dok se ne dostigne zadata maksimalna duzina recenice.

2.4 Model

Kombinovanjem prethodno opisanih enkodera i dekodera tako da dekoder inicijalno dobije stanja iz enkodera, dobija se konacni model.



Slika 1: Konacan izgled modela

3 Rezultati

Opisani model je treniran na 20.000 parova recenica na engleskom i francuskom jeziku iz skupa preuzetog sa [1] koji je deo Tatoeba projekta.

3.1 Primeri prevoda

3.2 Analiza medjureprezentacije

3.3 BLEU ocena

4 Zaključak

Cak i sa veoma malim skupom od svega nekoliko hiljada recenica ovaj model daje veoma dobre rezultate.

Literatura

[1] URL: http://www.manythings.org/anki/.

[2] Kyunghyun Cho i dr. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL].

[3] Ilya Sutskever, Oriol Vinyals i Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215 [cs.CL].

[4] Yonghui Wu i dr. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL].