Ispravljanje gramatičkih grešaka na srpskom jeziku

Autori:
Vasilije Marković SV15/2021
Nataša Radmilović RA20/2021

Mentor: Branislav Anđelić
Predmet: Osnovi računarske inteligencije

Fakultet tehničkih nauka,
Univerzitet u Novom Sadu

Apstrakt

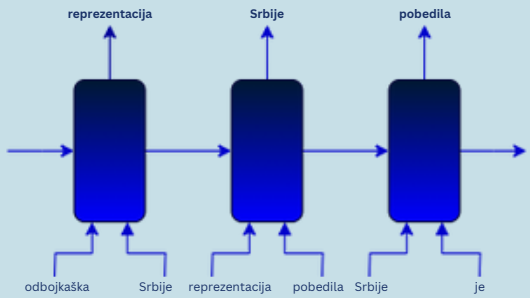
U današnjem društvu česte su gramatičke i pravopisne greške u pisanim tekstovima. Javljaju se u dokumentima različitih sfera, bilo da je to prosveta, zakon, trgovina itd.
U odsustvu alata za rešavanje ovog problema, cilj je napraviti program koji će za date ulazne rečenice ispraviti štamparske, pravopisne i gramatičke greške i, na samom kraju, evaluacija i analiza dobijenih rezultata, kao i diskusija o eventualnom poboljšanju modela.

Metodologija

Korišćen skup podataka predstavlja obiman korpus rečenica na srpskom, prikupljenih iz različitih novinskih članaka.
Pretprocesiranje podataka obuhvata izbacivanje specijalnih znakova, podelu rečenica na reči, i zatim proveru štamparskih ili jednostavnih grešaka upotrebom *Spellcheckera* i srpskog rečnika.
Sledeći korak je *inverzno* enkodiranje skupa reči, tako da akcenat bude na poslednjem slogu. Ovo je korisno za ispravljanje grešaka u padežima, rodovima i drugim gramatičkim kategorijama. Rezon omogućuje da reči sličnog nastavka nose sličnu brojčanu vrednost koju mreža koristi za treniranje.

LSTM

Za treniranje, korišćena je *Long short-term memory (LSTM)* rekurentna neuronska mreža, koja, za razliku od obične *RNN*, ima mogućnost pamćenja prethodnih ulaza i razlikovanje bitnih podataka za korišćenje u budućnosti od nebitnih.
Kao ulaz, korišćene su dve reči, gde je izlaz ona reč koja bi mogla da se nađe između njih.



Implementacija

Model se trenira na osnovu uređenih parova reči na osnovu kojih se predviđa reč koja stoji između njih. Iz skupa podataka od preko 1400 rečenica izdvojena je svaka uzastopna trojka reči i enkodovana u input, odnosno output vrednost.
Algoritam za enkodovanje pretvara jednu reč u broj, krećući se od poslednjeg slova ka prvom i svaki put dodajući UTF-8 vrednost slova, dignutu na određeni stepen. Stepeni se smanjuju kako iteracije idu do početnog slova, čime se postiže to da slova na kraju reči imaju veći značaj za mrežu. Cilj rezona je imitacija sličnih gramatičkih kategorija.
Model se trenira kroz 20 epoha i *loss* vrednost se konvergentno smanjuje. Ukoliko ga pronađe, može i da pročita već sačuvani model iz foldera *Models*, radi uštede vremena.
Istreniran ili preuzet model se koristi za iterativno računanje dva potrebna „magična broja", *probability threshold* i *bucket range*, nad izabranih 80 rečenica. U daljem tekstu objašnjena je njihova funkcija.
Nakon celokupne pripreme dolazi do predviđanja rezultata nad test-rečenicama. Svaka rečenica se tokenizira u skup reči, te se reči provuku kroz *Spellchecker*, odnosno proveri se ispravnost reči. Reč je ispravna ako je program pronađe u datom rečniku, u suprotnom, uzme onu koja je najbliža originalnoj po *Levenštajnovoj udaljenosti*.
Korigovane reči se dalje enkoduju i prosleđuju mreži. Za svake dve reči iz test-rečenica između kojih se nalazi tačno jedna treća reč, mreža predviđa brojčanu vrednost koja se može nalaziti između njih. Verovatnoća sa kojom se izvrši predviđanja mora biti veća od *probability threshold*-a da bi se zamena posmatrala kao ispravna.
Na osnovu dobijene vrednosti se, pomoću istog rečnika, izdvajaju sve one reči koje su, kada enkodirane, bliske originalnoj reči iz test-rečenice. Opseg koji se posmatra kao blizak određen je vrednošću *bucket_range*. Iz dobijenog skupa kao ispravljena reč uzima se ona koja je opet po *Levenštajnovoj udaljenosti* najbliža originalnoj reči iz test-rečenice.
Nakon što program prođe kroz moguće zamene, ispisuje promenjenu rečenicu i prosečnu verovatnoću svake promene koju je primenio.

Rezultati

Za testiranje iskorišćeno je 170 rečenica. Rezultati nisu savršeni zbog kompleksne prirode srpskog jezika i ograničene sintakse u podacima za trening. Međutim zapaženi su pozitivni rezultati u nekolicini slučajeva, kao što su prepoznavanje genitiva, menjanje prideva iz muškog u ženski rod, kao i menjanje množine u jedninu:

...bezbednosti Ujedinjenih **nacije** počela...
→ ...bezbednosti Ujedinjenih **nacija** počela...

...od početka **sledećih** godine...
→ ...od početka **sledeće** godine...

Na grafiku desno prikazana je zavisnost tačnosti predviđanja od ukupne verovatnoće kojom je predviđanje realizovano. Tačnost dobijenih ispravljenih rečenica u poređenju sa originalnim rečenicama izmerena je *Levenštajnovom udaljenošću*.