CV skill extraction using natural language processing

Sadržaj

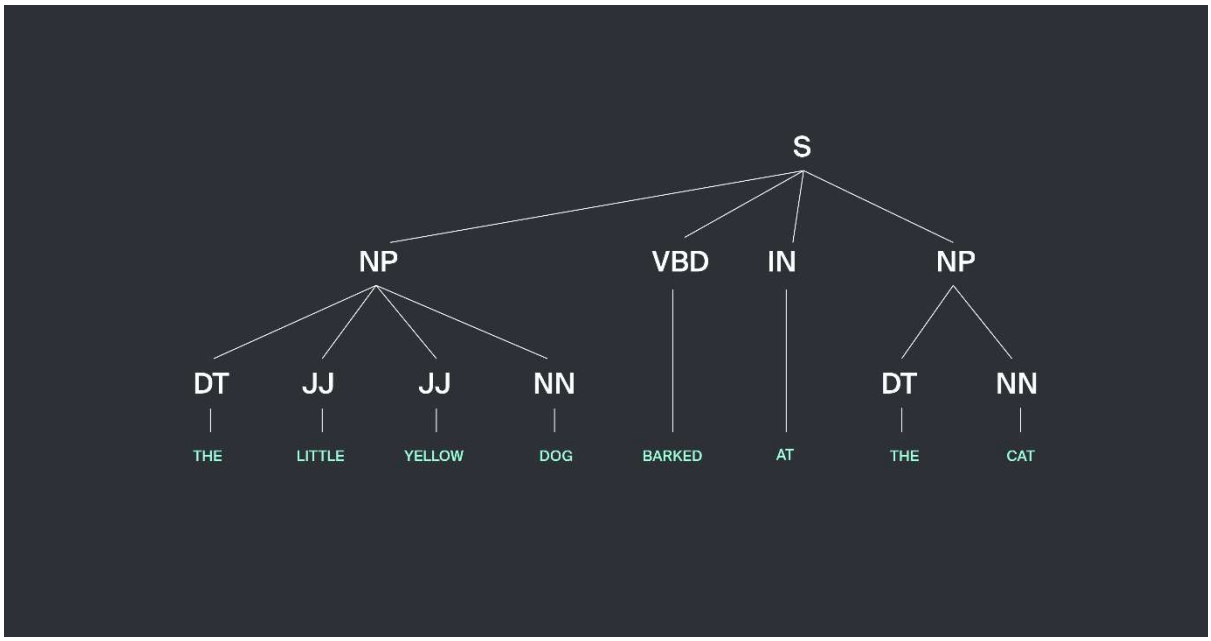
Opis zadatka

Korištenjem obrade prirodnog jezika (NLP) potrebno je izvuci vještine iz životopisa proizvoljne strukture. Ideja zadatka proizlazi iz članka https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada.

Ideja rješenja

Part of speech tagging

Prvi korak je izvući korisne informacije iz nestrukturiranog teksta. To ostvarujemo prepoznavanjem entiteta korištenjem ugrađenih metoda knjižnica kao što su NLTK. Part of speech tagging (POS tagging) metoda koristimo za označavanje vrste riječi i pravljenje stabla rečenične konstrukcije.



Prikaz stabla POS oznaka: S -sentence; NP - noun phrase; DT - determiner; JJ -adjective; VBD - verb, past tense; IN - preposition or conjuction; NN - noun

Vještine se uglavnom prikazuju kao imeničke fraze (eng. Noun phrases; NP). Koristeći POS tagger možemo iz teksta izdvojiti sve NP koje će naposljetku služiti kao kandidati za ulaze neuronske mreže. NP nisu standardni dio POS oznaka stoga će biti potrebno koristiti se tehnikom komadanja (eng. Chunking; više na https://www.nltk.org/book/ch07.html poglavlje 2). Ukratko to je tehnika spajanja riječi u veće komade uz pomoć regularnih izraza i oznaka entiteta.

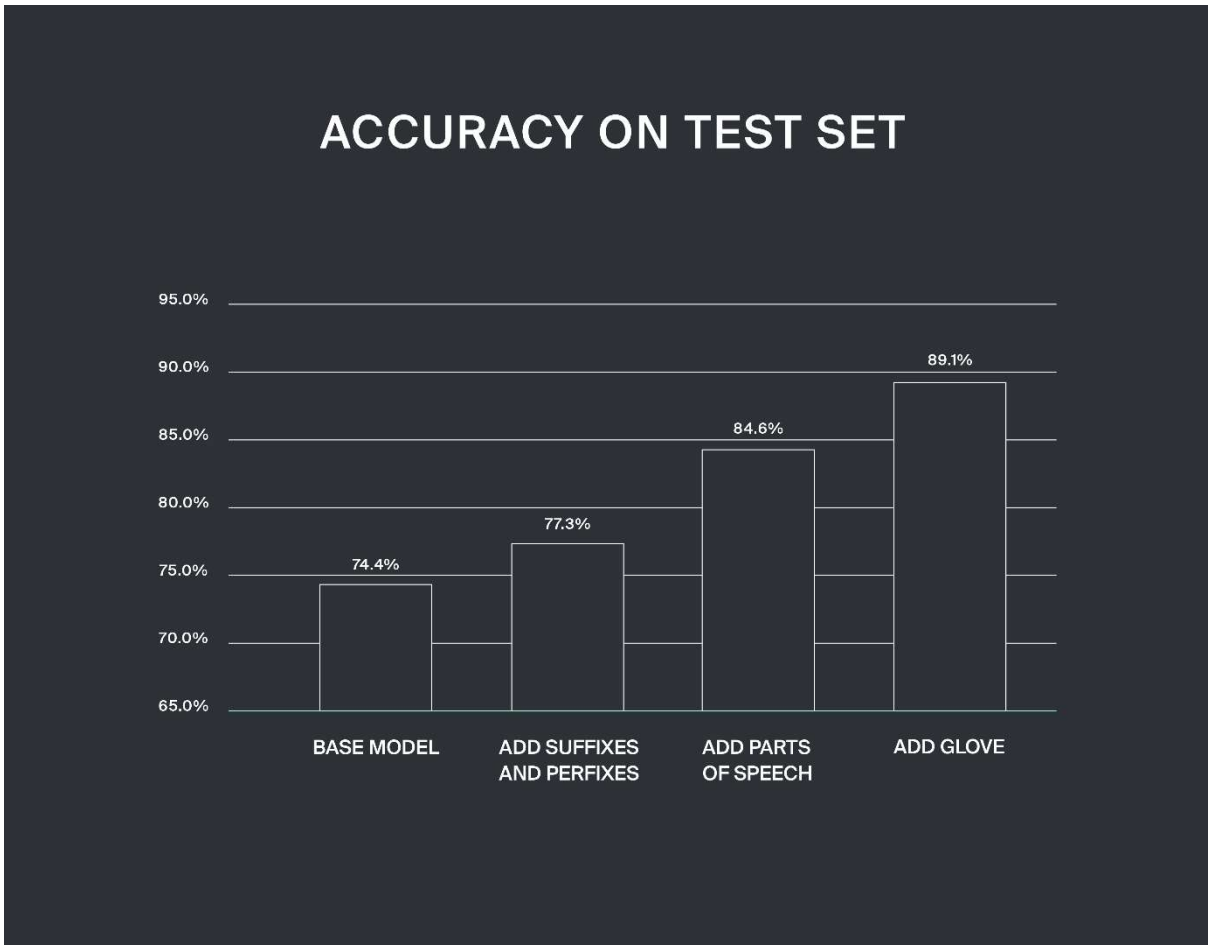
Ovim korakom dobivamo sve imeničke fraze izvučene iz teksta. Dio fraza su ciljane vještine, no dio nije jer fraze mogu predstavljati mjesta, osobe, objekte i ostalo.

Vekotrizacija riječi

Neuronske mreže dizajnirane su da uče iz brojčanih podataka te je zbog toga potrebno vektorizirati riječi. Svaki vektor riječi sastavljen je od binarnih značajki (0 – nije prisutna značajka, 1 – prisutna značajka). Neke od značajki su: pojava brojeva, veliko početno slovo, sve riječi napisane

velikim slovom, pojava simbola i slično. Provjerava se nalazi li se riječ u engleskom vokabularu i u tematskim listama kao što su imena, geografska imena i slično. Ovako vektorizirane riječi predstavljaju bazni model vektorizacije riječi. Na bazni model dodajemo binarne značajke koje prikazuju prisutnost popularnih engleskih prefiksa i sufiksa. Dodaje se i one-hot encodan (više: https://www.youtube.com/watch?v=v_4KWmkwmsU) prikaz POS oznaka. Naposljetku dodaje se i word embedding (vektorski prikaz riječi u n-dimenzionalnom sustavu; više: https://www.youtube.com/watch?v=gQddtTdmG_8).

Kako je za pouzdan word embedding model potreban jako velik skup podataka, a skup životopisa je uzak i mal skup podataka, koristit ćemo prethodno trenirane 50-dimenzionalne GloVe vektore (https://nlp.stanford.edu/projects/glove/).



Pretpostavljena preciznost modela s obzirom na dodane značajke

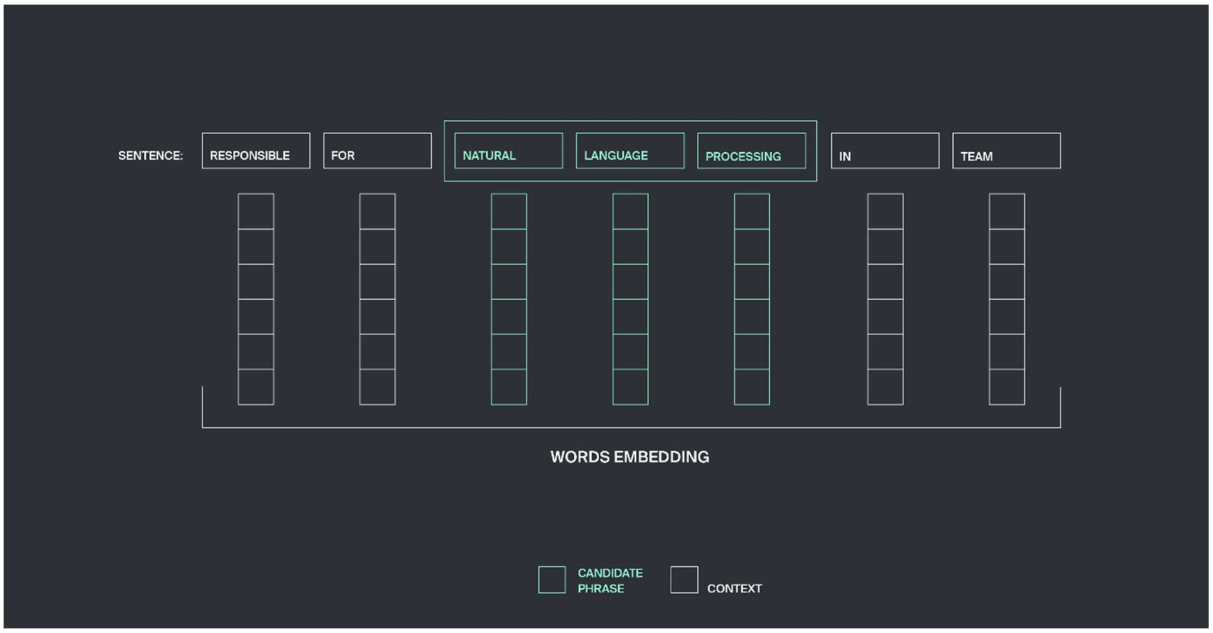
Možda će biti potrebno napraviti vlastiti POS tagger, ako bude problema oko oznaka pogledati članaka.

Neuronska mreža: SkillExtractor

Klasifikacija je ostvarena Keras neuronskom mrežom s tri ulazna sloja koji svaki prima specijaliziranu vrstu podataka.

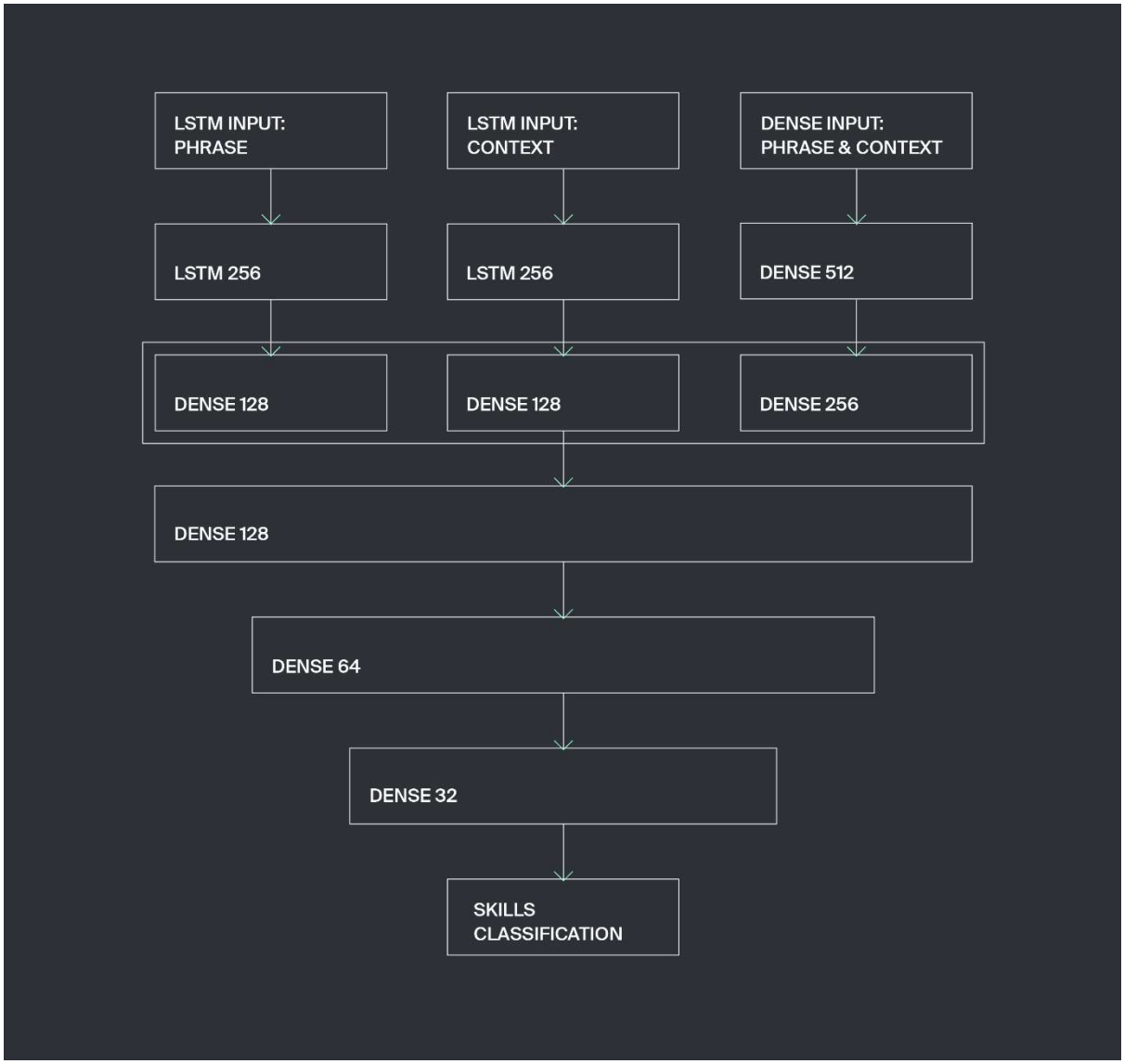
Prvi ulazni sloj prima fraze tj. varijabilno dugačke vektore (nemaju sve fraze jednak broj riječi). Vektori su sastavljeni od prethodno vektoriziranih riječi koje čine frazu. Ulaz se obrađuje LSTM slojem.

Drugi ulazni sloj uvodi kontekst zadane fraze. Uzimaju se n riječi s lijeve i desne strane fraze fraze te se sve konkateniraju u varijabilno dugačak vektor. Prema članku najoptimalniji n je jednak 3.



Prikaz ulaza drugog sloja (kontekst) za n=2

Treći ulazni sloj prima vektor fiksne dužine koji sadrži generalne informacije o kandidatskim frazama i kontekstu s obzirom na koordinate tj. maksimalne i minimalne vrijednosti vektora. Time se dobiva velik spektar informacija: u kojem prostoru se nalazi vektor, postojanje ili odsustvo binarnih značajki u cijeloj frazi itd.



Arhitektura neuronske mreže SkillExtractor