Primož Belej
Vpisna številka: 63150398
Brstnik 4, 3270 Laško
Slovenija

Komisija za študijske zadeve
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

Vloga za prijavo teme magistrskega dela

Kandidat: Primož Belej

Podpisani/-a študent/-ka magistrskega programa na Fakulteti za računalništvo in informatiko, zaprošam Komisijo za študijske zadeve, da odobri temo dela, podrobno opisanega v nadaljnjem predlogu teme magistrskega dela.

Okvirni naslov magistrskega dela:

slovensko: **Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami**
angleško: **Part of speech tagging of slovene language using deep neural networks**

Za mentorja/mentorico predlagam:

Ime in priimek, naziv: izr. prof. dr. Marko Robnik Šikonja
Ustanova: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Elektronski naslov: marko.robnik@fri.uni-lj.si

Za somentorja/somentorico predlagam:

Ime in priimek, naziv: dr. Simon Krek
Ustanova: Univerza v Ljubljani, Filozofska fakulteta in Institut Jožef Stefan
Elektronski naslov: simon.krek@guest.arnes.si

V Ljubljani, dne .....

Podpis mentorja: Podpis kandidata/kandidatke:

PREDLOG TEME MAGISTRSKEGA DELA

1 Področje magistrskega dela

slovensko: obdelava naravnega jezika, strojno učenje
angleško: natural language processing, machine learning

2 Ključne besede

slovensko: oblikoskladenjsko označevanje, globoko učenje, konvolucijske nevronske mreže
angleško: part-of-speech tagging, deep learning, convolutional neural network

3 Opis teme magistrskega dela

3.1 Uvod in opis problema

Oblikoskladenjsko označevanje besedil (v nadaljevanju označevanje) je pomemben del priprave besedila, torej nestrukturiranih podatkov, za nadaljne naloge s področja obdelave in razumevanja naravnega jezika. Pri tem postopku besedam določamo oznake, ki nosijo informacijo o besedni vrsti in drugih morfoloških in skladenjskih lastnostih.
Za označevanje besedil se uporabljajo algoritmi strojnega učenja, še posebej za slovenščino in druge morfološko bogate jezike.
Globoko učenje je skupina algoritmov strojnega učenja. Uporablja več slojev nelinearnih procesnih enot, ki izvajajo preoblikovanja podatkov. Vsak sloj uporablja podatke iz predhodnega sloja. Tekom učenja, se enote v vmesnih slojih prilagajajo podatkom in iščejo konfiguracijo, ki bi najbolje opisovala podatke iz učne množice.
V magistrskem delu bomo besedila v slovenskem jeziku označevali s pomočjo globokega učenja, saj se je to izkazalo uspešno pri podobnih nalogah [1, 2, 3].

3.2 Pregled sorodnih del

Grčar, Krek, Dobrovoljc (2012) [4] V okviru projekta Sporazumevanje v slovenskem jeziku je skupina raziskovalcev razvila orodje imenovano Obeliks. Za označevanje besed so uporabili klasifikacijski algoritem, ki temelji na principu maksimalne entropije. Za učne podatke so uporabili korpus ssj500k [5]. Za učenje in napovedovanje

so značilke iz besedila dopolnili z drevesom končnic. Pri napovedovanju besedne vrste je označevalnik dosegel 98,3 % točnost, celotne oznake pa je določal z 91,34 % točnostjo.

Ljubešić, Erjavec (2016) [6] Avtorji tega dela so za osnovo uporabili označevalnik Obeliks. Osredotočili so se na optimizacijo označevanja neznanih besed. Poročajo o 98,94 % točnosti pri napovedovanju besedne vrste, za celotne oznake pa 94,27 %. Za napovedovanje oznak so uporabili sekvenčni algoritem pogojno naključno polje Markova (CRF, conditional random field) ter uvedli dodatne spremenljivke. Za učenje napovednega modela so uporabili korpus ssj500k različice 1.3 [5].

Dos Santos, Zadrozny (2014) [1] V tem delu sta avtorja predstavila model, ki za označevanje uporablja globoko učenje. Njun model uporablja predstavitev besed na nivoju znakov. S predstavljenim modelom globokega učenja sta implementirala označevalnika za angleški in portugalski jezik. Angleški je dosegel 97,32 % točnost, portugalski pa 97,42 %.

3.3 Predvideni prispevki magistrske naloge

Pričakujemo, da nam bo v sklopu magistrskega dela uspelo razviti označevalnik za slovenski jezik z napovedno točnostjo primerljivo ali boljšo od obstoječih označevalnikov. Predvidevamo, da bomo to dosegli brez tvorjenja dodatnih značilk.
Z uporabo ansambla obstoječih napovednih modelov in našega, bi lahko napovedno točnost še dodatno dvignili.

3.4 Metodologija

Za implementacijo naše rešitve bomo uporabili knjižnico Keras za programski jezik Python. Ta knjižnica omogoča delo z nevronskimi mrežami in ponuja enoten vmesnik za knjižnice kot sta Tensorflow in Theano.
Pri imlementaciji se bomo zgledovali po modelu predstavljenem v [3]. Ta model deluje na nivoju znakov, a napoveduje na nivoju besed. Vhode pošlje skozi konvolucijsko nevronsko mrežo (CNN, convolutional neural network), katere izhod uporabi na jezikovnem modelu rekurenčnih nevronskih mrež (RNN-LM, recurrent neural network language model).
Za razvoj napovednega modela bomo uporabljali korpus ssj500k 2.0 [7], ki vsebuje več kot 500.000 ročno označenih besed. Primere iz korpusa bomo razdelili v učno, validacijsko ter testno množico. Na testni množici bomo primerjali uspešnost naše rešitve ter sorodnih rešitev za slovenski jezik. Napake bomo analizirali in skušali razložiti, kar bomo uporabili za uspešnejši ansambel večih metod.

3.5 Literatura in viri

[1] C. Dos Santos, B. Zadrozny, Learning Character-level Representations for Part-of-Speech Tagging, Proceedings of the 31st International Conference on Machine Learning ICML-14 (2011) (2014) 1818–1826.

[2] M. Labeau, A. Allauzen, Non-lexical neural architecture for fine-grained POS tagging, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015) (September) (2015) 232–237.

[3] Y. Kim, Y. Jernite, D. Sontag, A. M. Rush, Character-aware neural language models, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16) (2016) 2741–2749.

[4] M. Grčar, S. Krek, K. Dobrovoljc, Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik, Proceedings of the 8th Language Technologies Conference (2012) 89–94.

[5] S. Krek, T. Erjavec, K. Dobrovoljc, S. Može, N. Ledinek, N. Holz, Training corpus ssj500k 1.3 (2013) [navedeno 29.11.2017].
URL `http://hdl.handle.net/11356/1029`

[6] N. Ljubešić, T. Erjavec, Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene, LREC 2016 (2016) 1527–1531.

[7] S. Krek, K. Dobrovoljc, T. Erjavec, S. Može, N. Ledinek, N. Holz, K. Zupan, P. Gantar, T. Kuzman, Training corpus ssj500k 2.0 [navedeno 2017-11-29].
URL `http://hdl.handle.net/11356/1165`

Ljubljana, 29. november 2017.