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**
angleško: **Part of speech tagging of slovene language**

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:
Ustanova:
Elektronski naslov:

V Ljubljani, dne .....

Podpis mentorja: Podpis kandidata/kandidatke:

PREDLOG TEME MAGISTRSKEGA DELA

1 Področje magistrskega dela

slovensko: obdelava naravnega jezika, globoko učenje
angleško: natural language processing, deep learning

2 Ključne besede

slovensko: oblikoskladenjsko označevanje, oblikoslovno označevanje, nadzorovano učenje, konvolucijske nevronske mreže
angleško: part-of-speech tagging, supervised 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, na nadaljne naloge s področja obdelave naravnega jezika.
Pri tem postopku besedam določamo oznake, ki nosijo informacijo o besedni vrsti ter morebitnih dodatnih lastnostih.
Za označevanje besedil se uporabljajo različne tehnike, ki ponavadi uporabljajo algoritme s področja strojnega učenja. Takšne tehnike so se izkazale za uporabne pri označevanje slovenščine in drugih morfološko bogatih jezikov, a do zdaj niso bile sposobne označevati z uspešnostjo primerljivo s človeškim strokovnjakom.
V našem magistrskem delu bomo besedila v slovenskem jeziku označevali s pomočjo globokega učenja, saj se je to v preteklosti že izkazalo uspešno pri takšnih in podobnih nalogah [1, 2].

3.2 Pregled sorodnih del

Grčar, Krek, Ljubešič (2012) [3] V okviru projekta Sporazumevanje v slovenskem jeziku je skupina raziskovalcev z Instituta Jožef Stefan in zavoda Trojina razvila orodje poimenovano Obeliks. Za označevanje besed so uporabili klasifikacijski algoritem,

ki temelji na principu maksimalne entropije. Za učne podatke so uporabili korpus ssj500k. Za učenje in napovedovanje so značilke iz besedila dopolnili z drevesom končnic. Pri napovedovanju besedne vrste je njihov označevalnik dosegel 98,3 % točnost, celotne oznake pa je označeval z 91,34 % točnostjo.

Ljubešić, Erjavec (2016) [4] Avtorji tega dela so za osnovo uporabili označevalnik Obeliks. Osredotočili so se na optimizacijo označevanja neznanih besed. Poročajo o 25 % izboljšavi tako pri znanih kot neznanih besedah v primerjavi z označevalnikom Obeliks. Za napovedovanje oznak so uporabili sekvenčni algoritem CRF ter uvedli dodatne spremenljivke. Za učenje napovednega modela so uporabili korpus ssj500k s številko različice 1.3.

Dos Santos, Zadrozny (2014) [1] V tem delu sta avtorja predstavila svoj 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 tega magistrskega dela uspelo razviti označevalnik za slovenski jezik z napovedno točnostjo primerljivo z obstoječimi označevalniki. Predvidevamo, da bomo to dosegli brez tvorjenja dodatnih značilk.
Z uporabo metode zlaganja obstoječih napovednih modelov z našim, 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.
Za razvoj napovednega modela bomo uporabljali korpus ssj500k 2.0 [5], 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.

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] 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.

[4] N. Ljubešić, T. Erjavec, Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene, Lrec 2016 (2016) 1527–1531.

[5] S. Krek, K. Dobrovoljc, T. Erjavec, S. Može, N. Ledinek, N. Holz, K. Zupan, P. Gantar, T. Kuzman, Training corpus ssj500k 2.0, slovenian language resource repository CLARIN.SI (2017).

Ljubljana, 28. november 2017.