Razpoznavanje imenskih entitet na podatkovni množici ssj500k

Jakob Makovac
UL FRI, Večna pot 113
jm6531@student.uni-lj.si

Abstract

This document contains the formatting requirements for TACL final versions. These formatting rules take effect for all final versions received from September 2, 2018 onwards.

1 Uvod

Prepoznavanje imenskih entitet je pomembno področje obdelave naravnega jezika. Cilj je pridobivanje informacij iz teksta tako, da prepoznamo in ustrezno označimo različne imenske entitete, na primer: osebe, lokacije, organizacije ipd.

Za ocenjevanje rezultatov se navadno uporabljajo natančnost, priklic in ocena F.

Trenutno so najboljši rezultati doseženi s pomočjo modelov, ki se poslužujejo nevronskih mrež. Ocena F, ki jo dosegajo taki algoritmi so okrog 90% za angleščino in nekaj manj, okrog 85% za težje jezike, na primer španščino, nizozemščino (Yadav and Bethard, 2019).

Namen našega dela je, da najprej poskušamo implementirati razpoznavanje imenskih entitet s pomočjo pogojnih naključnih polj, nato dodati oziroma spremeniti značilke za izboljšanje rezultata, nato pa vključiti dodatne postopke za pridobivanje značilk, verjetno s pomočjo nevronskih mrež.

2 Sorodna dela

Najprej smo poskušali ponoviti rezultate članka Štajnerja (Štajner et al., 2013). V članku so uporabili podatke iz označenega korpusa ssj500k. Implementirali so postopek nadzorovanega učenja z algoritmom pogojnih naključnih polj. Najboljši rezultati, ki so jih dosegli dosegajo 74% natančnost in 72% priklic. Rezultati so se precej razlikovali po razredih entitet. Pri osebnih imenih na primer je bila končna ocena F 88%, pri stvarnih na primer pa samo 33%. Rezultate so poskušali izboljšati z vključevanjem dodatnih podatkov. Uporabili so na primer tako oblikoskladenjske oznake, kot tudi leksikone.

Model pogojnih naključnih polj se pogosto nadgrajuje z dodatnimi algoritmi, ki izboljšajo njegovo učinkvitost. Model LSTM-CRF na primer, izboljša napovedi modela CRF za okoli 5%, predvsem z izboljšanjem priklica (Habibi et al., 2017).

3 Podatki

Pri svojem delu smo uporabili slovenski korpus ssj500k, ki vsebuje okoli 500 tisoč besed. Te so ročno označene na nivoju tokenizacije, stavčne segmentacije, morfoloških oznak in lematizacije. Od tega je polovica besed označenih tudi na nivoju imenskih entitet, kar je pomembno za naše delo (Krek et al., 2018).

4 Metode

4.1 Pogojna naključna polja

Model pogojnih naključnih polj (conditional random fields ali krajše CRF) se pogosto uporablja pri procesiranju naravnega jezika ali določenih aplikacijah računalniškega vida. Uporaben je pri modeliranju sekvenčnih podatkov. Če stavek predstavimo kot zaporedje besed, je vsaka beseda označena z najbolj verjetnim stanjem. Označevanje je podobno Markovskim modelom odvisno le od lastnosti trenutne in predhodne besede.

Model CRF si lahko predstavljamo kot graf v katerem stanja predstavlja množica razredov entitet. Definiramo lahko dve naključni spremenljivki X in Y, kjer je X naključna premenljivka prek zaporedja podatkov, ki jih želimo označiti, Y pa spremenljivka preko razredov entit. (X, Y) je torej naključno polje, če ima Y markovsko lastnost glede na graf: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, kjer $w \sim v$ predstavlja da sta w in v soseda v grafu

(Lafferty et al., 2001).

5 Rezultati

6 Diskusija

7 Zaključek

References

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2018. Training corpus ssj500k 2.1. Slovenian language resource repository CLARIN.SI.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vikas Yadav and Steven Bethard. 2019. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *arXiv e-prints*, page arXiv:1910.11470.

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Named entity recognition in slovene text. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81.