Named Entity Recognition

Jakob Božič
jakob.bozic@gmail.com

Abstract

Named entity recognition (NER) is used to extract the named entities from the text into the pre-defined categories. Most commonly these categories are at least names, locations, organizations and other, however in various branches of NER different, branch-specific categories are added. State-of-the-art (SOTA) results are, similarly as in other fields of Machine Learning, given by deep (recurrent) neural networks, which have in last years completed the transition from hand-crafted to deep features. In our work we first give a brief overview of recent development of the field with more focus on Slovene language and then use of the current SOTA (RNN) approaches on ssj500k (Krek et al., 2019) dataset. We will also try to take advantage of transfer learning, which could prove very useful given the rather small amount of resources in Slovene. (We propose few improvements of the method, which are suitable for Slovene language and perform extensive evaluation. <- To je stvar prihodnosti)

1 Related work

We focus on more recent, deep learning based approaches, using RNNs. A brief overview of history of NER is given in (Yadav and Bethard, 2019), which also contains results of different NER approaches for four different languages and shows that deep learning approaches outperform traditional methods across the board. We will implement the method proposed in (Yang et al., 2017), which uses transfer learning to achieve then SOTA performance.

2 Data

We evaluate our approach on ssj500k (Krek et al., 2019) dataset. It contains approximately 500.000 words, manually annotated on different levels. Roughly half of the dataset is also annotated with named entities, on which we mainly focus.

3 Methodology

Detaled description of the model we selected as a starting point is given in (Yang et al., 2016). The selected model uses Gated Recurrent Units (GRUs) (Cho et al., 2014) and Conditional Random Field (CRF) to output labels. It takes advantage of both word- and character-level embeddings and GRUs. Three various forms of transfer learning are evaluated, we will focus on cross-lingual transfer, however they also studied cross-domain and cross-application transfer.

4 Results

We compare our results on ssj500k dataset with those from (Štajner et al., 2013).

5 Discussion

TODO

References

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches.

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. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks.

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.