NLP - First defense

Anonymous TACL submission

1 Introduction

Named entity recognition (NER) is a subfield of natural language processing, where we seek to process text so that we locate and classify named entities into pre-defined categories. This has been done through hand crafted grammar-based techniques and through machine learning models. While state-of-the-art solutions for English are reaching near-human performance, we will attempt to implement a NER system for Slovenian language using the ssj500k dataset (Krek et al., 2019). To do this, we have researched existing approaches in terms of how they are implemented and evaluated. This report summarizes our findings and approaches in the field of NER.

2 Existing solutions

Since many approaches have been developed to solve the problem of named entity recognition, we started looking at surveys that were done on NER systems. One such survey covers fifteen years of research in the NER field (Nadeau and Sekine, 2007), where the authors describe the developed machine learning techniques and also review features and model evaluation.
Another survey was done on deep learning models, which have only recently been studied and developed (Yadav and Bethard, 2019). In this survey they present architectures for NER and compare them to previous approaches.

From these surveys we found out that most of the popular and recent approaches use some variant of recurrent neural networks (RNN). An efficient and lightweight architecture is presented in the paper (Shen et al., 2017), where they have drastically reduced the amount of training data, while still achieving close to state-of-the-art results.

Besides neural networks we have also looked at other papers, which describe conditional random fields or Hidden Markov Models as described in (Zhou and Su, 2002), where they also achieved great performance. We also reviewed an article published on NER in Slovene (Štajner et al., 2013), where they used conditional random fields and the same ssj500k database as we were planning to use.

3 Initial ideas

When reviewing our options, we decided our first approach would be an LSTM neural network. The TensorFlow library allows us to easily create, train and evaluate such networks in Python. For parsing the dataset we intend to use existing tools, such as the tei-reader (Python 3 Library for Reading the Text Content and Metadata of TEI P5 Files) and Natural Language Toolkit (NLTK).

References

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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 473–480, USA. Association for Computational Linguistics.

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.