Named Entity Recognition

NLP - Second defense

Jure Bevc, Anže Dežman, Robert Modic

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 [3]. 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 Related work

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 [4], 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 [7]. 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 [5], 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 [8], where they also achieved great performance. We also reviewed an article published on NER in Slovene [9], where they used conditional random fields and the same ssj500k database as we were planning to use.

3 Implemented baselines

For the task of Named entity recognition of the Slovene language we decided to evaluate the *Stanford Named Named Entity Recognizer model* [2] and the *DeepPavlov BERT based model* [1]. We trained them using the ssj500k [3] dataset.

3.1 Dataset preprocessing

We used the *minidom* [6] Python library to write an *XMLParser* class, with functions to allow extraction of all words with their named entity subtype.

In ssj500k datset the named entities are marked with the <*seg*> tag, subtype provided in the "subtype" argument. This tag can contain one or more

<*w*> tags, denoting words, that are part of a named entity. However, not every named entity in the dataset was marked with the <*seg*> tag. We only used the part of the dataset that had the named entities marked.

We wrote two named entity extraction methods, one for the *Standord model* and one for *DeepPavlov BERT based model*. They both return the same elements, just in different format.

3.2 DeepPavlov BERT based model

The preprocessed dataset output of newline separated sentences was first randomly split into a trainig set, validation set and test set with the ratios of 0.8, 0.1 and 0.1, as instruced in the documentation[1].

After about 12 hours of training the model was evaluated. The training set had the following amount of named entities:

NER subtype	number
LOC	222
MISC	38
ORG	116
PERSON	301

The following table shows the results of the evaluation.

NER tag	precission	recall	F1 score
LOC	90.54%	91.78%	91.16
MISC	56.25%	50.00%	52.94
ORG	79.31%	76.67%	77.97
PERSON	87.71%	93.62%	90.57

From the initial results we can see, that the currently trained model seems to be very good at detecting persons or locations. Low performance for MISC subtype can be attributed to low number of such samples in the ssj500k dataset.

3.3 Stanford model

The testing and training datasets that were used for the DeepPavlov BERT model were also used to evaluate the Stanford Named Named Entity Recognizer model [2]. This model is based on a Conditional Random Field (CRF) sequence model. Table 3.3 shows the results of the evaluation.

NER tag	precission	recall	F1 score
LOC	84.36%	81.35%	82.82
MISC	60.29%	32.28%	42.05
ORG	76.07%	57.40%	65.43
PERSON	75.95%	88.24%	81.63

We can observe that although the model achieved decent results in some categories, it generally performed worse than the DeepPavlov model.

4 Future directions

Now that we have built two working models, we have several options for future testing. One possible direction is finding even more new and interesting models to build and evaluate, or we could simply work on improve the existing ones. We could also test these models using data sets in different languages, to see what kind impact the Slovenian grammar has on the models performance.

References

[1] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, A. Litinsky, V. Logacheva, A. Lymar, V. Malykh, M. Petrov, V. Polulyakh, L. Pugachev, A. Sorokin, M. Vikhreva, and M. Zaynutdinov. Deeppavlov: Open-source library for dialogue systems. 07 2018.

[1] http://docs.deeppavlov.ai/en/master/features/models/ner.htmltraining-data

[2] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[3] S. Krek, K. Dobrovoljc, T. Erjavec, S. Može, N. Ledinek, N. Holz, K. Zupan, P. Gantar, T. Kuzman, J. Čibej, Š. Arhar Holdt, T. Kavčič, I. Škrjanec, D. Marko, L. Jezeršek, and A. Zajc. Training corpus ssj500k 2.2, 2019. Slovenian language resource repository CLARIN.SI.

[4] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[5] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[6] xml.dom.minidom. 20.7. xml.dom.minidom — minimal dom implementation — python 3.6.10 documentation.

[7] V. Yadav and S. Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.

[8] G. Zhou and J. Su. 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, 2002. Association for Computational Linguistics.

[9] T. Štajner, T. Erjavec, and S. Krek. Named entity recognition in slovene text. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81, Dec. 2013.