FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Ing. Miloslav Konopík, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Technická 8, CZ-30614 Plzeň
Phone: (+420) 377 63 2418

Review of the Doctoral Thesis

Jana Straková: Neural Network Based Named Entity Recognition

Thesis Topic

Jana Straková's Ph.D. thesis deals with the research field of named entity recognition (NER). NER systems look for expressions of a special meaning such as locations, persons, organizations, numbers, or calendar data. These expressions often hold a key information for understanding the meaning of a document.

Discussion of the Thesis Content

Chapter 1 explains the importance of NLP processing. I feel that the NER task should be introduced already in this chapter, including open issues and challenges the author would like to focus on.

Chapter 2 nicely introduces the NER research field. Since the chapter is targeted on starting researchers, I consider the example presented on page 8 misleading. Beginners often mistakenly believe that the NER task could be easily solved by Gazetteers (dictionaries) and this example might support this mistake. I would prefer some examples showing the complexity of the task such as “Dan je mistrovský stupeň v Judu”. It would clearly show that named entities must be classified in the context.

Chapter 3 introduces the Czech NE corpora CNEC 1.0 and 2.0. I miss some details about the annotation process. How many annotators participated in creation of the corpora? What was the inter-annotators agreement? How exactly were the sentences for annotation selected? I agree that the NLP community has benefited from the corpus since it was used at University of West Bohemia as well.

Chapter 4 describes two machine learning approaches that can be used for the NER task: log-linear modeling and softmax neural networks. The author tackles the problem of unseen weights in large models and explains the need to use the softmax layer at the output of the neural network. Figure 4.3 shows the effect of setting certain values to unseen weights on POS tagging and NER task performance. It seems that certain range of values can improve the classification accuracy. However, I see no experiment showing that an optimal weight established on development data improves also the performance on test data.

Chapter 5 describes creation of a NER system. The chosen approach uses several techniques such as softmax function, two-stage prediction and *Brown* clusters to obtain the state of the art in Czech NER. The lead in the state of the art at the time of creation was impressive.

Chapter 6 describes *NameTag* : an open-source tool for named entity recognition. The published on-line demo is very attractive. I have instructed a small student team to integrate *NameTag* into our simple question answering system in order to test it properly. The results exceeded my expectations. The tool provides very robust results in our system.

In chapter 7, the author presents a featureless and language agnostic NER system. However, experiments only for Czech and English are presented. Why not evaluate the system on other available NER corpora – such as Spanish, Dutch, and others? On page 72, in footnote 4, the author shows that the *word2vec* tool was trained with one epoch only, however, it is generally recommended to use 3 – 5 epochs. In section 7.2, the author states a hypothesis that character embeddings should be especially helpful for morphologically rich languages. However, I do not see any confirmation of such statement in Table 7.1. Character embeddings contribute both to Czech and English results. This might be included in the discussion section. Despite all, the architecture of the proposed neural network is innovative and the achieved results are promising.

Conclusion

I believe that the author achieved three main contributions:

- She constantly pushed the state of the art in the Czech NER research field.
- She participated in creation of NER corpora.
- She participated in building the open source *NameTag* tool.

All these achievements significantly contributed to the NER research field. Clearly, the work has a high research value.

I found no significant formal problems in the thesis. It is clearly written and well understandable.

I conclude that the author demonstrated the ability to carry out independent research and therefore I recommend Jana Straková's thesis for defense.

Plzeň, May 11, 2017

Ing. Miloslav Konop´ık, Ph.D.
(reviewer)