Univerzita Karlova v Praze
Matematicko fyzikální fakulta
Studijní oddělení – doktorské studium

Review of doctoral thesis

Neural network based named entity recognition

submitted by **Jana Straková**

The thesis deals with names entity recognition (NER) that is an important part of natural language processing (NLP). It concentrates on the neural-network based techniques for NER applied on Czech language, with careful evaluation on standard and publicly available Czech data. The results are also available in the form of public research software. The thesis has 8 chapters, one appendix, and 108 pages in total. This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents overall conclusion and recommendation to the PhD committee.

Technical content of the thesis and remarks to chapters

Chapter 1 introduces the reader into the thesis and gives basic motivations. I appreciate the thoroughness of section 1.1 with the description of author's contribution in shared works and publications, I think any thesis should contain such precise information. For foreign readers not familiar with the Czech system of husband's name adoption, it would be nice to make an explicit link "Strakova - Kravalova". The schematic instructions for reading the thesis (Fig. 1.1) are really nice.

Chapter 2 gives an overview of NER within the broad area of NLP. After the definition of the task and introduction to machine learning (that could be skipped in my opinion), it contains a part (sec. 2.2.2) on NER encoding that is directly linked to its evaluation. Here, I would appreciate schematic illustration of BIO and BILOU encodings. The chapter mentions also unlabeled data and semi-supervised training and then deals with popular features used for NER. Finally, it contains a state-of-the-art description of Czech and English NER, the Czech one turning obviously around the Czech named entity corpus (CNEC) of which the candidate is a co-author. Although its parameters are discussed in the next chapter, I would like to have at least basic data included already in section 2.4.

Chapter 3 presents CNEC in detail. I appreciate including data description at this point in the thesis, as (similarly to other machine learning tasks), data is the very core, and in some cases, the data itself gets modified by the thesis experimental work. I was missing any information of the *sources of texts* for CNEC (mentioning Czech National Corpus is not enough) – the source of data is quite important for experimental work and also for practical applications. The chapter discusses in detail also the granularity of labels and I appreciate that the experimental chapters address these different granularities. CNEC seems to be quite biased towards sentences with high percentage of NE, but this is understandable, as all the annotation was manual and later, the authors of CNEC addressed it in the 2nd version of the corpus.

In section 3.2, CNEC is used for developing NER systems and one might consider the results given in section 3.2.2 as baseline ones. I would however find it very useful to present the reader the technical description of this system – or saying clearly that this is the system from Chapter 5. I find encouraging, that the results generated on CNEC 2.0 are not too different from 1.0, so that the high proportion of NE in the data does not seem to hurt. I would however like to see a comparative table of the two versions showing the counts of sentences, words and NE instances (extension of Table 3.1). In the presentation of NER errors (table 3.6), I think confusion matrices for coarse and fine NE classes would be more suitable. Finally, this chapter (and several other sections of the

thesis) mention that for full usage of larger context for NER, it would be necessary to crawl back to CNK – here, I would like to know, if this work has ever been attempted and what were the results.

Chapter 4 describes the development of neural network (NN) based infrastructure for several NLP tasks. Although not fully related to NLP on Czech, it shows the intellectual and software development of these tools and I find it quite relevant. After an introduction (sometimes rather lengthy), algorithmic and application issues of NNs are discussed. The results in Table 4.1 indicate that the feature vector sizes were huge, and I wonder if any feature dimensionality reduction techniques (PCA, LDA, HLDA) were tried and with what results. In addressing the optimization of global (sequence) criteria, I would also recommend to have a look at the works of Dan Povey and Karel Vesely in automatic speech recognition, maybe they would contain interesting ideas also for NLP/NER. Finally, the result tables (4.2-4) should be completed with the best world's results for individual tasks obtained prior to 2013 (this work) and in 2016/17, and the results should be commented, as they are in the later chapters.

Chapter 5 deals with NER using NNs with standard NLP-motivated features. The system is described with reference to the previous chapters, but I was missing description of some of the blocks – for example the post-editing. The decoding part is using the Viterbi algorithm, but again, more precise description is missing – did it use just rules or were some transition probabilities or costs defined, and if so, how ? In the two-stage system, it seems that a window of 500 predictions of the candidate word was taken into account, I wonder how this context size was found (experimental tuning ?) and how it goes together with the "broken context" property of CNEC. The results show that NNs on Czech NER work better than the previous published works by a large margin.

Chapter 6 is on a publicly available tool for NER "NameTag". I appreciate the efforts to make the results of thesis (both corpora and SW) available, this greatly contributes to the credibility of the results. In table 6.1, I would again appreciate not mentioning just Strakova et al. 2013, but saying clearly "system described in Chapter 5". Also, it seems that the released systems are a bit less accurate than the experimental one (F-measure 80.3, resp. 81 versus the magical 82.82), it would be nice to explain why – simpler architecture ? Less features ?

Chapter 7 investigates into NER systems based on low-dimensional word-embedding features. It presents the related work, but rather than seeing word/character-embeddings as a "black box", their short description should be included. Also, it would be nice to add details on their training (English Gigaword and Czech SYN) – at least the sizes of data. On the other hand, I really appreciate the architecture suggested in Fig. 7.2 that is nicely adaptable to other languages (one can omit the "classical" NLP features in case we don't have tools/models to derive them) and I liked the character-based features. The results are excellent even in presence of word- and character-features only, and get even better with classical features. I also like critical comparison and discussion of the results on English. Finally, I had to grin when the author complained about training time of the full system – ½ day on one CPU is something ASR people would dream of…

The final chapter contains the conclusions, with a summary of research results and pointing possible future directions. It would be nice to disclose which industrial company made use of the results (if not forbidden by an NDA) and to elaborate a bit on the future work in separate "low hanging fruit" and "pipe dreams" sections.

Summary on the technical content of the thesis

I have some critical comments on the thesis (mainly in the rigor of describing the systems, data, and experimental parameters) but as a whole, the thesis clearly demonstrates the qualities of the candidate – capability to study non-trivial literature from several fields, suggest own novel solutions, implement them, carefully test and discuss the results. I highly appreciate the quantity and quality of experiments done on different data-sets and the fact that most of the data and software is publicly available – this contributes to the *trust* one can have in the experimental results. The author also gives an impression of a great team worker (judged from numerous collaborations and joint papers).

The suggested NN approaches for NER perform excellently and are on par with the state of the art in English, and are leading in the Czech NER, with suggested ways to extend them to other languages. The contribution of the candidate to the Czech and international research community is therefore without any question.

Comments on the formal aspects

The thesis is written in a nice, almost error-free English and its structure is logical and easy to follow, with some isolated exceptions mentioned above. The mathematical writing is correct, sometimes, the candidate prefers verbal expression of what is done – this is fine for literature surveys, but in the technical sections describing own development, I would sometimes appreciate a bit more mathematical "flesh". The figures and schemes are well executed, people with engineering background would appreciate more schemes. There is a limited number of typos and grammatical errors, the candidate will receive a commented version of the document to help her fix these problems, in case corrections are allowed for the final publication of the thesis.

Summary and recommendation

I have carefully examined the doctoral thesis of Ms. Jana Strakova. Despite the criticism raised above (many points are rather recommendations than critique), in my opinion, it is a solid work that contributes to progress in the NLP research field.

To conclude, I do recommend accepting the Thesis as a partial requirement for granting Mr. Jana Strakova the Doctoral degree at the Charles University in Prague.

For the defense, I have the following questions:

1. At several places in the thesis, you mention that going back to CNK would allow for reconstructing the missing long contexts. Has this been tried ? If so, by whom and what were the results ?
2. Were any feature dimensionality reduction techniques (PCA, LDA, HLDA) tried with the pre-word-embedding NN NER systems ?
3. Explain, how the size of window of 500 predictions was determined and how it goes together with the "broken context" property of CNEC.
4. NameTag seems to be a bit less precise than your experimental systems. Why ?

In Brno, May 29th 2017

Dr. Jan "Honza" Cernocky, Associate Professor
Head of Department of Computer Graphics and Multimedia
Responsible of BUT Speech@FIT group
Faculty of Information Technology, Brno University of Technology
Bozetechova 2, 612 66 Brno, Czech Republic
Tel: +420 5 41141284 Fax: +420 5 41141290,
mailto:cernocky@fit.vutbr.cz, http://www.fit.vutbr.cz/~cernocky
http://www.fit.vutbr.cz/ https://cs-cz.facebook.com/FIT.VUT
http://speech.fit.vutbr.cz https://www.facebook.com/BUT-Speech/