REVIEW

of PhD thesis by Le The Anh

"Deep Neural Network Models for Sequence Labeling and Coreference Tasks"
submitted for the degree of candidate of technical sciences
specialty 05.13.01. - "System analysis, control theory, and information processing"

Review author

Full name: Natalia Loukachevitch

Academic degree: Doctor of Technical Sciences

The year of awarding the degree and the scientific specialty in which the degree is awarded: 2016, specialty - 05.13.17 Theoretical foundations of computer science

Academic title: Dr

Place of work: Lomonosov Moscow State University, Leninskie Gory 1, 4, Moscow, Russia

Position: Leading Researcher

Contact information: phone: +7 (926) 144 6163 louk_nat@mail.ru

Currently, deep neural networks methods allow achieving state-of-the art results in many applications of natural language processing (NLP). Therefore, further studies of neural network architectures corresponding to specific NLP tasks are significant. The thesis written by Le The Anh is devoted to application of deep neural networks to tasks of sequence processing and co-reference resolution.

In Chapter 1 the author considers history of deep learning, deep learning models and brief overview of natural language processing tasks. Main contributions of the work are indicated. Chapter 2 provides important concepты of deep learning for natural language processing. Section 2.1 describes details of various vector representations and the history of their appearance. In the next section (2.2), the author considers the most known current deep learning approaches for natural language processing such as convolutional and recurrent neural networks, including LSTM architectures. The section 2.3 is devoted to recently-appeared pre-trained language models, such as Elmo, Transformers, GPT, BERT.

Chapter 3 is devoted to the description of the proposed approach in sequence labeling tasks, including named entity recognition, POS tagging, and sentence boundary detection. Here also the related work on the sequence labeling tasks is presented, including Vietnamese and Russian-based related work. The proposed approach includes the following components: word, character, and capitalization embeddings, chunks; bidirectional LSTM, feed-forward network, and a CRF layer. The CNN architecture is used as a method for generating character embeddings. For further improvement of the results BERT and ELMO contextualized embeddings are used.

The approach is evaluated on six datasets labeled with named entities: three Russian datasets, Vietnamese, Chinese, and English. The proposed model achieved results compared with the state-of-the art results on the English CONLL-2003 dataset, and the best results on Russian and Vietnamese datasets.

The same model is applied to the task of sentence boundary detection, which is treated as a sequence labeling task. The approach is directed on identification of statements and questions for chat-bots.

Chapter 4 is devoted to the co-reference resolution task, which is very important for information extraction applications. A short survey of approaches in co-reference resolution is given. A new model named Sentence-level Coreferential Relation-based model (SCRb) was proposed. The model obtained the results comparable to state-of-the-art results on the English OntoNotes dataset and the best results on two Russian datasets.

The main results of the work are:

1. An original hybrid model for sequence labeling task was proposed and studied. This model extended existed Bi-LSTM CRF architectures with (1) trainable CNN for generation of character-level representation of an input sequence, and (2) Bi-LSTM network for encoding capitalization features. The model achieved state of the art performance on Russian and Vietnamese datasets.

2. Extensions of the original architecture with encoders based on language models ELMo and BERT were evaluated on Russian and English datasets. They obtained state of the art performance on Russian datasets, and the performance comparable with the state of the art results on CoNLL-2003.

3. Application of proposed sequence labeling model to the sentence boundary detection task produced solid results of 89.99% F1 and 95.88% F1 on the Cornell Movie-Dialog and DailyDialog datasets.

4. An original model for learning sentence-level coreferential relationships was introduced. Incorporation of this model in the baseline coreference architecture improved its performance for English. Application of the model with sentence coreference module and language model extensions for the Russian language allowed achieving the state of the art performance.

The achieve results seem significant from the theoretical and practical points of view.

It worth noting some remarks about this work:

1. The representation of chunks in sequence labeling architectures is not explained clearly.

2. There is no comparison of the results achieved in the sentence boundary task with other approaches.

These comments do not decrease the significance of work.

I believe that the topic of the dissertation is of both theoretical and practical interest. It corresponds to the basic directions of research in priority areas of the Artificial Intelligence domain. The work is fully consistent with the requirements, and Le The Anh deserves the degree of candidate of technical sciences in the specialty 05.13.01. – System analysis, control theory, and information processing.

«10 » 06 2020

Natalia Loukachevitch,
Doctor of sciences,
Leading researcher of
Research Computing Center of
Lomonosov Moscow State University

Подпись Лукашевич Н.В. заверяю

Директор НИВЦ МГУ имени
М.В. Ломоносова

член-корреспондент РАН,
проф. Воеводин В.В.