REVIEW

of the 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"

Reviewer

Full name: Artem Olegovich Shelmanov

Academic degree: Candidate of Technical Sciences (PhD)

The year of awarding the degree and the scientific specialty, in which the degree is awarded: 2015, specialty 05.13.17 — "Theoretical Foundations of Computer Science"

Academic title: -

Place of work: Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, Moscow, Russia 121205

Position: Research Scientist

Phone: +7 (495) 280 14 81

Email: a.shelmanov@skoltech.ru

The thesis of Le The Anh is devoted to the problems of the automatic processing of natural language texts. Natural language is a complex phenomenon that encodes information for effective communication between humans who can use all their intellectual power and knowledge about the world to correctly interpret the original meaning of textual messages. Implementing natural language understanding in an automated system is a challenging task, since describing an algorithm of text interpretation via a limited number of formal rules is usually ineffective due to ambiguity of textual elements of different levels (graphemes, words, sentences, etc.) To deal with this ambiguity, recent research works in this area resort to data-driven approaches based on statistical and machine learning models.

Le The Anh, in his work, considers two topics of natural language processing, namely sequence tagging and coreference resolution. Sequence tagging is a

backbone of information extraction systems. It is primarily used for named entity recognition – an essential task of recognizing in texts person, location, and organization names, as well as many other objects of different nature. Coreference resolution is considered as one of the most challenging tasks in text processing, which provides an ability for a computer system to determine whether two mentions of objects in a text refer to the same entity or not. Both sequence tagging and coreference resolution are very important for streamline NLP applications such as AI assistants, dialog systems, information search engines.

Le The Anh addresses these problems with deep neural network models. Neural models have become ubiquitous in many NLP research directions. Due to the ability to learn deep latent feature interactions effectively, neural models have brought a breakthrough in the quality of natural language understanding compared to classical machine learning methods.

Le The Anh proposes a **novel** neural architecture for sequence tagging based on the CNN-BiLSTM-CRF model. Modifications proposed by Le The Anh include multiple convolutional layers with different filter sizes for character encoding and a recurrent subnetwork for processing categorical features such as capitalization. This architecture is also combined with recently proposed deep distributional semantic models that produce contextualized word representations (ELMo and BERT). The model is tested on a named entity recognition task and a sentence boundary detection task. The good side of the work is that Le The Anh conduct multilingual experiments for named entity recognition and show the effectiveness of his method across Russian, Chinese, Vietnamese, and English languages. For Russian and Vietnamese languages, Le The Anh achieves state-of-the-art results.

For coreference resolution, Le The Anh develops a **novel** modification of a previously proposed feedforward neural model with attention and iterative refinement. He introduces an additional term for the scoring function that takes into account the relatedness of sentences, in which the referring and referred objects appear. The term is calculated with a neural subnetwork that is optimized jointly with the rest of the model. Experiments on corpora in English and Russian show that this modification helps to improve the baseline performance in some scenarios. Le The Anh also combines the baseline coreference resolution model with the deep pre-trained model BERT. With the help of this combination, the aspirant has won the coreference and anaphora resolution shared task on the Dialog evaluation seminar. Hence, his model shows state-of-the-art results among data-driven methods for coreference and anaphora resolution for the Russian language, which is a remarkable achievement.

There are few remarks on the considered thesis.

1. The results of experiments with sentence boundary detection are not compared to any baselines, which makes it difficult to judge the relevance of the contribution in this subtask.
2. In the anaphora and coreference resolution shared task, the winning system does not use the first modification proposed by the aspirant. Thus, it shows that the application of this modification is limited.
3. The discussion of the Dialog evaluation shared task results in Subsection 4.7.2 is too brief. It lacks the comparison of different types of implemented models and the motivation behind the chosen option for the final submission.

However, these shortcomings detract from the merits of the thesis. The work has been implemented at a high scientific level, its results have been reported at 3 international conferences and published in an international journal. There are 4 articles in total published on the subject of the thesis.

The work of Le The Anh has become a part of the widely used open-source framework for processing of Russian texts based on deep neural network models, namely DeepPavlov. Publishing not only results but also the source code of the models makes it possible for researchers and developers from academia and industry to take advantage of Le The Anh's work, which is an important advantage.

The thesis summary reflects the main aspects of the work. It discusses the relevance of the topic, thesis goals and objectives, methods of research, scientific novelty, practical value, implementation, approbation, contribution, the volume of the work and achieved experimental results.

The work by Le The Anh demonstrates new significant scientific results that help to push further state of the art in natural language processing, especially for the Russian language. I believe that the topic of the thesis is of both scientific and practical interest; it corresponds to the basic directions of research in priority areas of science. I believe that 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."

Date: 15.06.2020

Artem Shelmanov

Подпись А.О. Шелманова подтверждаю
Hereby I confirm the signature of Artem Shelmanov

МЕНЕДЖЕР ПО ПЕРСОНАЛУ ПОЧЕПЦОВА

Skoltech