Report on the content of the dissertation

Oseledets Ivan Valerievich

(name of the committee member)

Candidate's full name: Le The Anh

Dissertation title: "Deep Neural Network Models for Sequence Labeling and Coreference Tasks"

Specialty: 05.13.01 - System analysis, control theory, and information processing (information and technical systems)

Scientific degree for which the dissertation is submitted: Candidate of technical sciences

Date of the defense: 18.11.2020

The evaluation of the dissertation in accordance with the Regulations on the award of scientific degrees of candidates and doctors of sciences at MIPT (hereinafter referred to as Regulations):

1. **Relevance of the dissertation topic:**

   The topic is devoted to the development of new models in natural language processing using deep learning method. Specifically, the author considers sequence Labeling and coreference resolution, which are very important.

   Sequence Labeling plays an important role in the field of NLP since many NLP tasks can be formed as Sequence Labeling. Typical tasks include Named Entity Recognition (NER), Part of Speech (POS), Word Segmentation, or Chunking. Coreference Resolution is also an important but hard task because of (1) its huge space search and (2) the need for global clustering decisions. Prediction results of both of these tasks are useful for NLP systems such as Information Extraction, Question Answering, or Chatbots. Any improvement of state-of-the art results is very relevant

Scientific novelty of the results:

The main novelty of the dissertation is the combination (fusion) of different techniques in order to obtain new models that achieve state-of-the art (SOTA) results in the abovementioned tasks. This is done by combine three components:

- Multi-level word vector representation comprised of (1) a context-free word word embedding, (2) character-level word embedding generated by a deep convolutional neural network, and (3) additional word features like POS, Chunk;
- Capitalization embedding of words that capture capitalization features of both left and right context of the current word. This is done by utilizing a Bi-directional Long Short-Term Memory network;
- Context-based word embedding produced by modern language models such as ELMo and BERT.

A lot of experiments were done to evaluate the proposed models on two tasks: NER and Sentence Boundary Detection (SBD) in four languages including Russian, Vietnamese, English, and Chinese. The obtained results point out that the proposed models achieve state-of-the-art performance on both Russian and Vietnamese NER datasets and a solid performance on SBD datasets.

For Coreference Resolution, Le The Anh proposed a new approach which differs from previous ones in:

- Powering the mention detection phase by modern language models such as BERT, ELMo.
- Improving both mention detection and mention clustering phases by learning sentence-level coreferential relations.

The proposed model to extract sentence relations in the coreference context, namely Sentence-level Coreferential Relation-based (SCRb) model helps to improve the baseline model on both Russian and English datasets.

2. **Theoretical and practical importance of the dissertation:**

- The proposed models for the Sequence Labeling task can be used for several NLP tasks like POS, Chunk, NER, Word Segmentation. These models obtain

cutting-edge performance on both Russian and Vietnamese datasets and a comparative performance on English datasets. The trained models are already published as a part of the popular open-source NLP framework DeepPavlov (https://deeppavlov.ai). This makes it easier to be used for both researchers and developers.

- The SBD model trained on conversational DailyDialog dataset is implemented as a part of DeepPavlov framework. This can be integrated into chatbot systems to split unpunctuated input text into sentences. By this way, the chatbot can be deal with the long and complex user utterances.
- The SCRb model can be used to extract the sentence relationship in a document-level context. The author experimentally shows that if the quality of this relationship is good enough, the model performance could be significantly boosted. Moreover, the SCRb model can be used to accelerate not only coreference models but also Question Answering, Text Summarization.

3. **Completeness of publication of the main results of dissertation in peer-reviewed scientific journals, according to the Regulations:**

   The proposed models and obtained results have been fully presented at four international conferences, and published in five papers. Four of them were already indexed by Scopus.

4. **Questions and remarks (according to the part 4.13 of the Regulations, the candidate addresses the questions and remarks formulated below during the defense):**

   1) Although in my opinion it is not always required to have a single "common" result for all of the chapters, it would be nice to have a general methodology that underlies all of the contibutions of the thesis. I.e., can one present a single result that contains certain "sub-results"?
   2) For the combined models, it is often a good idea to do "ablation studies", i.e. determine which part contributes most to the improvement of the final metrics. Have these studies been conducted?

3) Reproducability of deep learning model is very important, and also the initialization. The results presented in the thesis are not always presented with the variance. For example, Figure 3.8 has the variance over 5 runs (which may be not enough), but Table 3.11 does not (is it a mean value? Maximum value?).
4) Again, for reproducibility it is important to have all the hyperparameters of the training (batch size, learning rate, type of the optimizer used) available. The selection of these parameters may have an influence on the final result.

5. **General evaluation of the dissertation (excluding the introductory part):**
The results of the disserations are very good: it is enough to compare the quality of the obtained models to the ones that existed previously. This is a very serious improvement. The methods are implemented in a well-known large-scale software package DeepPavlov. The dissertation satisfies all requirements for the degree of candidate of technical sciences, specialty 05.13.01 – System analysis, control theory, and information processing (information and technical systems).

Date 22.10.2020

Signature _____ / Oseledets Ivan Valerievich

I hereby confirm the signature of Oseledets Ivan
Head of HR administration Gut O.S.

Skoltech