Report on the content of the dissertation

Проталинский Олег Мирославович

(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:**

Nowadays, deep neural network models comprehensively change the way we address NLP problems and became the main approach in the field of NLP. The topic of Le The Anh's thesis is applying Deep Learning for Natural Language Processing (NLP) tasks.

In this work, the author proposes and analyses a number of hybrid deep neural networks for Sequence Labeling task. These proposed models can be used to address a series of NLP tasks such as Named Entity Recognition (NER), Part of Speech (POS), Chunking. Besides, the task of Sentence Boundary Detection (SBD) is also solved by, firstly, reformulating as a Sequence Labeling task.

For the task of Coreference Resolution, the author proposes a new method to explore the sentence-level coreferential relation in the document context. This approach can be used to enhance the model performance of not only Coreference

Resolution task but also any other NLP task that requires the information about the sentence relationship like Question Aswering or Text Summarization.

2. Scientific novelty of the results:

- A novel hybrid model is introduced for the Sequence Labeling task. Generally, in this model, the word embedding is comprised of:
  - Pre-trained word embedding
  - Character-level feature-based word embedding
  - Capitalization feature-based word embedding
  - Contextualized word embedding, generating by fine-tuning a modern language model such as BERT or ELMo.
- The proposed model is evaluated on NER and SBD tasks. The conducted experiments show that the model obtains cutting-edge results on Vietnamese and Russian NER datasets, and good results on SBD datasets.
- A new method to address the Coreference Resolution task is proposed which focuses on exploring the information about sentence-level coreferential relation. The author conducted some experiments and confirmed that this information is very useful for Coreference Resolution task. The Sentence-level Coreferential Relation-based (SCRb) model is built and obtains better results on both Russian and English datasets compared with the baseline model.

3. Theoretical and practical importance of the dissertation:

- In the dissertation, several deep neural network models for Sequence Labeling task are presented and analyzed. These models can be used for many NLP tasks like NER, POS, Word Segmentation, or Chunking. The conducted experiments show that the proposed models achieved 99.17%, 94.43% on Russian NE3 and Vietnamese VLSP-2016 datasets, respectively. The implementations of these models and the trained models are integrated as a component of the open-source conversational AI framework, namely DeepPavlov (https://deeppavlov.ai), that is easy to download and use.

- The implementation of SBD model along with the trained model on DailyDialog dataset are also available to free download on the github page of DeepPavlov framework (https://github.com/deepmipt/DeepPavlov). This model can be used as a first preprocessing step in a chatbot system.
- The SCRb model can be utilized to extract the sentence relationship in a long text and can be applied to several high-level NLP tasks like Question Answering or Text Summarization.

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

   The main content of the dissertation has been fully presented at four international conferences, and timely published in five papers, out of which four papers were indexed by Scopus.

5. **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. What method was used for training of neural networks?
   2. How much sample size was used for training of neural networks ?

6. **General evaluation of the dissertation:**

   Le The Anh's dissertation addresses two important NLP tasks that are significant and useful for many other NLP tasks. The proposed models along with achieved results are presented and discussed at the prestigous conferences. The practical significance of the developed models is confirmed by their publication as a part of DeepPavlov framework. This makes them easy to use by both researchers or developers..

Date 30. 10. 2020

Signature / Проталинский Олег Мирославович

Подпись удостоверяю
Начальник управления по работе с персоналом
Н.Г. Савин