REVIEW

on the procedure for the defense of Doctoral Thesis:

„Intelligent Context-Aware Natural Language Dialogue Agent"

for the acquisition of the educational and scientific degree "Doctor"

of

Ph.D. Candidate: **Momchil Emilov Hardalov**

Scientific field: **4. Natural sciences, mathematics and informatics**

Professional field: **4.6 Informatics and Computer Science**

Doctoral program: "Software Technologies – Knowledge Discovery"

at **Department of Software Technologies**

Faculty of Mathematics and Informatics (FMI)

Sofia University "St. Kliment Ohridski" (SU)

The opinion is written by Prof. Olga Ilieva Geirgieva, Department of Software Technologies, FMI – SU as a member of the scientific jury according to order № РД 38-600/28.10.2022 of the Rector of Sofia University "St. Kliment Ohridski".

1. General characteristics of Doctoral Thesis and the presented materials

The presented Doctoral Thesis contains in total 163 pages, which include an introduction, four chapters on the essence of the dissertation research and a conclusion containing 18 figures and 39 tables. Parts of the main work are two applications in both in common of 8 pages. What is striking is the large number of titles in the bibliographic reference - a total of 350, all published after 2015 and presented in prestigious scientific publications - journals and proceedings of conferences in the topic of the dissertation. The scientific work is written in English, and the Abstract is presented in both English and Bulgarian.

The first chapter is introductory, in which the aim and tasks of the dissertation work are defined.

The second chapter presents a comprehensive analytical overview of existing systems with conversational agents, classified according to the character and goals of the dialogue, and the development techniques. The review pays particular attention to existing data sets to help create systems that maintain a conversation with users.

In the third chapter, a joint slot-value element language model for natural language understanding is presented, which builds on the well-known BERT model and its version RoBERTa. The validity and improvements of the proposed model are shown and compared with existing solutions and on different databases.

The fourth chapter focuses on methods that use additional information to answer to a user question. Approaches have been proposed to create conversational agents using context-aware response generation algorithms.

The fifth chapter analyzes and proposes solutions that implement an automated conversation with dialog managed by generating new responses different from the ones used. The task of multilingual models for cross-language transfer without additional training is realized.

The last –sixth chapter is concluding. It indicates the scientific contributions of the thesis work and presents directions for future research.

The PhD Candidate demonstrates in-depth knowledge and ability to use a wide range of the contemporary technologies in the field of dissertation work. The paper is written with great awareness of the research topic. The scientific theses and analyzes are defined and proven with the necessary thoroughness and accuracy. The material is carefully laid out, well-structured and illustrated.

These facts assure me that all requirements regarding the volume, contentt, structure and layout of the PhD Thesis been met.

In the competition documents, other documents were provided in support of Momchil Hardalov's work - an abstract in Bulgarian and English, a copy of all the scientific articles on the procedure, a report, a protocol and an opinion of similarity assessment, administrative documents assuring the implementation of the Law of Development of the Academic Staff in the Republic of Bulgaria.

2. Data and personal impressions about the candidate

Momchil Hardalov graduated a bachelor's degree in "Computer Systems and Technologies" at the Technical University - Sofia. He obtained his master's degree in "Information Extraction and Knowledge Discovery" at FMI SU with an excellent grade.

I have known the candidate since his enrollment in February 2017 as a PhD student at the Department of Software Technologies. My immediate impressions are related to his excellent performance in the annual attestations and presentations as a PhD student. As a member of the doctoral minimum committee, I had the opportunity to appreciate his

thoroughness and high level of knowledge in the subjects of the doctoral studies. Momchil Hardalov's work as a part-time assistant at the department was distinguished by responsibility, as well as by the ability to present topics in the field of machine learning in an attractive way for the students.

3. **Analysis of the scientific and the scientific-applied achievements of the candidate, contained in the presented doctoral thesis and the publications enclosed to it according to the procedure**

Scientific research in the thesis work is extremely relevant. The obtained scientific and scientific-applied results are of theoretical and practical value for building software systems with a communication channel with users. The existence of such a channel has already become a contemporary trend in implemented systems. This is an essential element that meets an existing need - automation work of the customer service. The dissertation provides original scientific results and innovative practical solutions in response to this need by addressing the challenge of creating a conversational agent with natural language understanding.

The contributions of the PhD Thesis are in the analysis, development and verification of methods necessary in the overall sequence of elements for building communication with a user - understanding natural language and generating an automated dialogue with the user. They can be summarized in the following scientific and scientific-applied contributions:

- **Scientific and scientific-applied contributions**

- **A new language model** named *Transformer-NLU* for natural language understanding has been developed, building on the well-known BERT and RoBERTa, by adding an additional layer for intent estimation, accounting for additional word features, and predicted distribution of the intents. The improvement of the proposed model over existing solutions is demonstrated through experiments on two standard data corpora. The application of the model is in building customer service chatbots to answer user requests.

- An **original methodology for retrieving relevant passages** from documents related to a specific context, for selecting an answer of a question with reading comprehension and **subsequent application to a target language without additional training** is proposed. The proposed solution addresses the problems resulting from a small training data resource and the lack of question contexts in the new target corpus.

- A **new method for learning from noisy data** is proposed for the purpose of building conversational agents. An iterative update of the labels based on predicted values of the loss function is applied.

- A **new methodology for ranking the candidate answers for the next replica** of conversational agents by using a question-answer quality assessment model with probabilistic assessment of the agent's answers is introduced. The proposed methodology implements an automated conversation with dialog managed by generating new responses.

- **Applied contributions**

It should be also highlighted the specific applied contributions of the thesis that are of a practical value for the experimental work of the research and for the benefit of future research and technological development in the field of conversational agents:

- A corpus of questions in Bulgarian has been collected for reading a question with comprehension with a selectable answer.

- A new cross-linguistic and multilingual benchmark corpus has been created for answering questions from high school exams (matriculation exams) in sixteen languages. A lot of data and code has been published to help the work on multilingual models.

- A corpus was created to discover already verified claims - pairs of tweets and corresponding articles for fact-checking.

In summary, the assessment of the contribution value of the PhD Thesis is notably high. The achieved results are significant and have both well-ground theoretical and clearly demonstrated practical value. It should be emphasized the large volume and quality of the work done.

4. Approbation of the results

The list of publications that reflect the results of the dissertation work contains six titles. Four of the publications are referenced in SCOPUS, and three of them have SJR. One of the publications is in a journal referred in Web of Science, and the other five are presented in the proceedings of prestigious conferences. The PhD Candidate is a first author in all publications. Declarations of equal participation of co-authors are presented for all titles. I believe that the presented publications sufficiently and indisputably prove the representativeness of the research and the results of the dissertation work.

The report on the fulfillment of the minimum national requirements shows that the publications cover 108 points from the group of indicators "Г" out of the required 30 points.

Thus, the candidate greatly exceeds the required minimum set by Art. 2b, para. 2 and 3 of the Law of Development of the Academic Staff in the Republic of Bulgaria and, accordingly, the additional requirements of SU for the acquisition of an educational and scientific degree "doctor" in the scientific and professional field of the procedure. It should be emphasized that the publications presented in support of the dissertation work have been evaluated by the specialists in the field. Those referenced in SCOPUS from the list of dissertation publications are cited in 28 publications reported up to the day the review was prepared.

There is no proven plagiarism in the submitted PhD Thesis and scientific works under this procedure.

5. Quality of the thesis abstract

The abstract of the PhD Thesis is presented in Bulgarian by 47 pages, and in English by 45. It summarizes precisely and in sufficient detail the theoretical studies, analysis and their practical implementation described in the dissertation. Its content follows that presented in the dissertation, describing the research and the contributions of each of the chapters. The abstract meets all the requirements for its writing.

6. Critical notes and recommendations

I have no critical remarks regarding the formulation of the tasks, analysis, accuracy and completeness of the obtained results.

7. Conclusion

Having familiarized myself with the doctoral thesis presented by the procedure and the scientific works accompanying it and based on the analysis of their significance and the scientific and scientific-applied contributions contained in them, **I confirm** that the presented PhD Thesis and the scientific publications to it, as well as the quality and originality of the results and achievements presented in them meet the requirements of the Law of Development of the Academic Staff in the Republic of Bulgaria, the Rules for its application and the relevant Rules of the SU "St. Kliment Ohridski" for the candidate's acquisition of the educational and scientific degree "Doctor" in the scientific field 4. Natural sciences, mathematics and informatics and professional field 4.6. Informatics and Computer Science. In particular, the candidate satisfies the minimum national requirements in the professional field and no plagiarism has been found in the scientific works submitted for the competition.

Based on the above, **I recommend the scientific jury to award Momchil Emilov Hardalov an educational and scientific degree "Doctor" in scientific field 4. Natural sciences, mathematics and informatics, professional field 4.6. Informatics and Computer Science.**

18.01.2023   Reviewer: ....

/ **Prof. Olga Georgieva /**