Evaluation of Dissertation Thesis
Opponent's review of the PhD. Work

Author: **Ing. Jiří Martínek**
Title: **Deep Learning Methods for Dialogue Act Recognition using visual information**

Affiliation: University of West Bohemia, Faculty of Applied Sciences

The proposed doctoral thesis is focussed on the problem of dialogue act recognition (DAR), which is important for the general understanding of the discourse or texts. The work proposed novel original approaches based on the combination of text processing, image recognition and data enrichment. The multi-modal and multi-lingual methods are promising research field, which is currently being studied very intensively. The application of these methods on the task of DAR is new, innovative, and interesting to the current needs of the scientific community. The results of the work can be directly applied not only in the context of DAR but also on the other tasks of text and image processing such as image extraction and named entity recognition.

Formally, the work is divided into three parts in which the author describes existing methods for DAR, design and testing of models for optical character recognition (OCR), and design and testing of models for multi-modal and multi-lingual DAR. The overview of the existing methods covers distributional language models for word and sentence representation and a description of the recurrent deep learning networks for text classification focusing on the dialogue act recognition.

The author has formulated the following thesis and goals of his doctoral thesis which can be summarized as:

- It is possible to use a small amount of annotated data for training of OCR systems and achieve very good results
- Cross-lingual transfer learning can improve quality of DAR methods
- Additional visual information in cross-modal transfer learning can improve quality of DAR methods

These theses were validated in following novel scientific results, which represent the original contributions of the work:

- The proposition of models for efficient OCR for historical documents based on the combination of convolutional and recurrent models
- The proposition of new combination of the methods for data enrichment for OCR models based on data transformation and generation of synthetic data for the given textual corpus
- The proposition of cross-lingual DA recognition models based on the combination of multi-lingual transfer learning and data enrichment with automatic translation between multiple languages

- The proposition of a new multi-modal model for DA recognition from image documents using text and image inputs to complement and enhanced each other results.

I have several notes about the presented work and achieved results, and some of them deserve discussion:

- For the multi-modal model, author evaluated systematically at first model with the textual input only, with the image input and finally the overall performance of the composed model (for variable quality of the input images). My question for the discussion is, if it is possible to more investigate and explain how the transfer learning is influencing the quality of the final model? How much information we can gain from the enrichment of the image data, and how much is transferred from the enhanced language model?
- When the proposed methods will be applied on the other tasks, such as named entity recognition or information extraction, do you expect similar benefits in multi-modal approach? Will it be possible to still reduce the requirements for the annotated data, without the compromises on performance and robustness of the methods?

The description of the research work is sound and correct. The doctoral thesis is well written, well conceptualized, and the style of the work is erudite and well organised. The work is intuitively well understandable with the balanced level of generality and specificity.

I think, the core of the PhD. thesis was published at the required level. The achieved results were published in 3 journal publications and 8 conference proceedings. The doctoral thesis satisfies conditions of a creative scientific work, and the list of publications confirms the author's scientific erudition.

Finally, I can state, that the author of the thesis Ing. Jiří Marínek proved to have an ability to perform research and to achieve scientific results. In my opinion, the thesis meets the generally accepted requirements for obtaining the academic degree of Ph.D. I RECOMEND the submitted dissertation thesis, based on the previous evaluation, for defence and after its successful defence, Ing. Jiří Marínek shall be awarded the academic degree "philosophiae doctor" PhD.

Košice, 23.2. 2022

assoc. prof. Ing. Peter Bednár, PhD.
Department of Cybernetics and Artificial Intelligence
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia

Report on Jiří Martínek doctoral thesis entitled

"Deep Learning Methods for Dialogue Act Recognition using Visual Information"

Jiří Martínek's doctoral thesis manuscript presents his work on optical character recognition, and on dialog act recognition. In the last part of the doctoral thesis manuscript, these two aspects are combined for dealing with dialog act recognition from image documents containing written dialogues. All the investigated and reported approaches are based on deep learning methods.

The main part of the doctoral thesis manuscript consists of 82 pages from introduction to conclusion, plus 9 pages of bibliography. Besides the introduction and conclusion, there are three main chapters. Chapter 2 presents a survey of deep learning approaches in text processing. Chapter 3 is devoted to optical character recognition, including state-of-the-art approaches, and contributions. Chapter 4 presents the contributions in dialogue act recognition; first from transcribed oral dialogues, and then from image documents containing written dialogues.

Chapter 1 introduces the topics of the doctoral thesis: dialogue act recognitions, as well as optical character recognition, which is a necessary step when the written dialogue is extracted from an image document. The introduction also emphasizes the doctoral thesis contributions.

Chapter 2 presents a survey of deep learning approaches in text processing. It starts with a detailed introduction of various approaches for word embeddings, which are representations of words as vectors. It then describes popular neural network architectures related to natural language processing, including recurrent neural networks, sequence-to-sequence models, attention mechanisms, and transformer-based architectures. The chapter ends by a short review of dialogue act recognition. It would have been interesting to add a discussion on the dialog act annotations: how are defined the dialog act tags? How many tags? Are they consistent across speech corpora of a same language? Across languages? In addition, some examples would have been appreciated.

Chapter 3 is devoted to optical character recognition. The chapter starts by a presentation of convolutional neural network architectures used in image classification and object detection, as well as their application in natural language processing tasks. The remaining part of the chapter is dedicated to optical character recognition. After a short review on that topic, Jiří

Martínek details his contributions with respect to optical character recognition applied to historical newspapers. This involve three steps: the segmentation of the image (corresponding to a page of a journal) into blocks of texts, the segmentation of each block into lines, and finally the recognition of the characters for each line. Deep neural network models are used in each of these three steps.

The models for recognizing the characters of a line are trained using the "connectionist temporal classification" loss that was initially proposed for speech recognition; the use of this loss avoids the need for pre-segmented training data (here, pre-segmentation of the image into characters). Nevertheless, some annotated data is necessary for training the model, i.e., pairs of items corresponding to the image of a line and the corresponding text. Getting such data for historical newspapers is costly and time consuming. Consequently, Jiří Martínek has investigated approaches for improving the training process with a limited amount of reference training data. The proposed approaches rely on data augmentation (geometrical modification of the text image and/or addition of noise) and on the use of synthetic data. Synthetic data can be created in large amounts. Among the approaches studied, the best character recognition performance is achieved with a two-step training process: first training on synthetic data, and then fine-tuning on the original annotated data.

This part of the research has been published in three international conferences and two journal papers.

Chapter 4 presents the contributions in dialogue act recognition. The first part of the chapter deals with dialogue act recognition from transcriptions of oral dialogues; and Jiří Martínek investigates multi-lingual and cross-lingual aspects. Experiments are conducted using both convolutional and recurrent neural networks, on English and on German dialogues from the Verbmobil corpus, for which utterances are annotated with dialogue act labels. Taking into account the previous utterance brings useful contextual information and leads to better results (than when ignoring it). He also proposed to apply transfer learning and fine-tuning for cross lingual dialogue act recognition: a model is first trained on English data; then for other languages, the utterances are translated into English, both for the training data used for fine-tuning the model, and for the test data used for evaluating the identification of dialogue acts. This leads to better performance than training from scratch. This is probably due to the limited amount of annotated data available for training; and it would be interesting to know if the proposed cross-lingual approach would still bring better performance when larger amounts of training data are available. In addition, when analyzing the cross-lingual results, it is mentioned that low F1 score occurred because of infrequent dialogue acts; to confirm this analysis it would have been interesting to also conduct a few experiments by using only the four most frequent dialogue acts.

The second part of the chapter concerns the recognition of dialogue acts from image documents containing written dialogues. This part relies on both optical character recognition and on dialogue act recognition applied on the recognized characters. Experiments are conducted using images of various quality (clean or noisy images). To compensate for errors in optical character recognition on bad quality images, Jiří Martínek has proposed a multimodal approach, which relies on the recognized sequence of characters and on an embedding of the image. This leads to improving results on noisy images.

The evaluation of dialogue act recognition in the first part of the chapter was conducted using perfectly transcribed dialogues; whereas in the second part the evaluation was conducted on text automatically extracted from images (and thus containing errors). It would be interesting to know if the impact of errors coming from optical character recognition is similar to the impact of errors coming from automatic speech recognition; unfortunately, automatic speech recognition was out of the scope of the doctoral thesis.

The research on dialogue act recognition has been published in three international conferences.

A short conclusion ends the doctoral thesis manuscript.

Overall, the doctoral thesis manuscript is well organized, well written and easy to read. Moreover, there are many figures to illustrate the various neural network based architectures. The doctoral thesis manuscript shows that Jiří Martínek has elaborated and conducted a large set of experiments for optical character recognition on the one side, and for dialogue act recognition on the other side. For all these experiments, he has used various types of neural network architectures. The contributions of Jiří Martínek are clearly detailed, and they have been published in several international journals and conferences, including top rank conferences.

To conclude, according to the work reported in the doctoral thesis manuscript and the associated scientific publications, I express a very favorable opinion for the PhD defense of Jiří Martínek.

Nancy, February 15th, 2022

Denis Jouvet
(*HDR, Université d'Avignon et des Pays du Vaucluse*)
INRIA Nancy – Grand-Est / LORIA