Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA18209 - NexusLinguarum

Grantee name: Atanas Hristov

Details of the STSM

Title: Deep learning for linguistic data analysis

Start and end date: 22/08/2022 to 03/09/2022

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

(max. 500 words)

The work carried out during the visit to Epoka University in Tirana generally followed the main topics of the Work Plan presented to the STSM grant application. During the STSM, I spent two weeks at Epoka University under the supervision of ass. prof. Arban Uka. During my visit, the work carried out has been mainly focused on developing a research environment for deploying deep learning techniques for linguistic data analysis. The STSM examine the effectiveness of deep learning in understanding the specificities of linguistic data analysis. We manage to implement a deep architecture and show that deep learning is an efficient and powerful supporting tool for linguistics. The proposed deep learning architecture was evaluated and the performance was tested at the computer system located at Epoka University.

The deep learning architecture that was established is for classification tasks since classification is one of the core types of tasks in linguistic data analysis. During the development, it was shown that the capabilities of deep learning prove that deep neural architectures are not only a trend but an efficient supporting tool for linguistics.

Additional values of the work were given by the discussion about the type of architecture that is most appropriate for knowledge extraction. Also, problems with data aggregators were discussed. Specially preprocessing of the text, vocabularies, batch size and data division was observed.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(max. 500 words)

The main results have been conducted mainly together with my host ass. prof. Arban Uka, and by the support of his team. We managed to achieve all activities attached to the proposed work plan including:

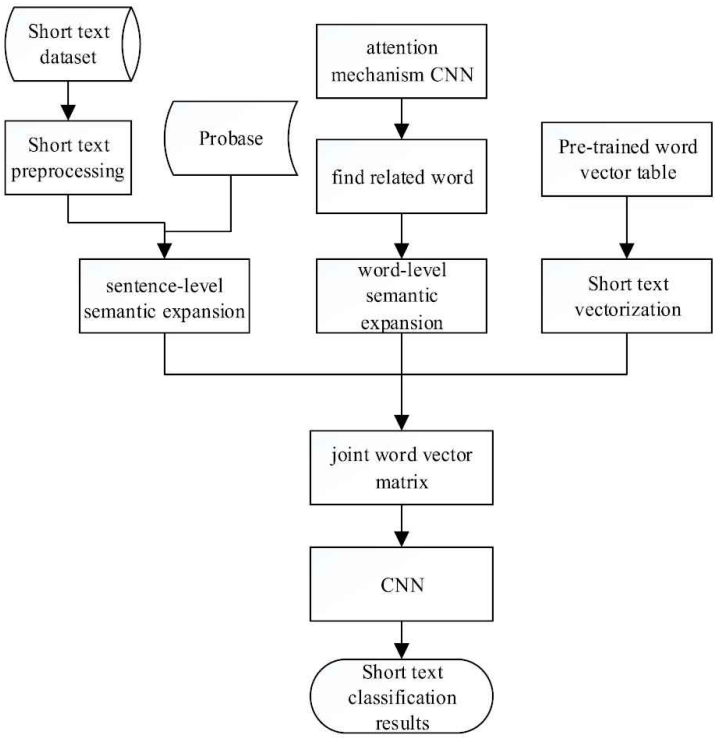
1. Review, identify and analyze the general problems and modern trends in the area of deep learning applications for supporting linguistics.

2. Design, study, and deployment of two different types of deep neural networks:

- Convolutional Neural Network for short-text categorization
- Recurrent Neural Network for word-level classification

Convolutional Neural Network (CNN) for short-text categorization

The first architecture is based on convolutions, since the need for massive calculations and filtering in text categorization using a convolutional matrix. The methodology implemented on CNN consists of computing features on batch, data generation, embedding and gathering layers. All word vectors are randomly generated and trained as model parameters. In order to be in line with the latest research as a benchmark, we use the work of Haitao Wang et al. The details are given in the figure below.



Recurrent Neural Network for word-level classification

The second architecture is recurrent because the words in text processing are usually repeated. This allows words to be used several times in different contexts, and deep architecture shows efficiency in this.

The text classification with RNN was done using TensorFlow. First, setup input pipeline was done, with buffer size 10000 and batch size 64. Next, a text encoder was created, with a vocabulary size of 1000 words. Creating the model is following. The overview of the model is as described: INPUT → Text Categorization → Embedding → Bidirectional → Dense → Classification. We train the model in 10 epochs and 30 validation steps.

We consider this STSM very successful, and we estimate that the project will take several weeks to be completed. Since I have opened doors for future collaborations with the group of ass. prof. Arban Uka we plan to collaborate with the host institution in the near future.
Since our project is still in the very early phase, we intend to disseminate our work by preparing scientific publications for international scientific journals and conferences, once our project is complete.