SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA16105
STSM title: Big Data and Scalable Crowdsourcing towards fostering the Language Skills of European Citizens (BiCSkill)
STSM start and end date: 18/08/2019 to 30/08/2019
Grantee name: Asst. Prof. Mahdi Bohlouli

PURPOSE OF THE STSM:

This STSM aims in networking and analyzing a potential solution towards the use of crowdsourcing and big data analysis in collecting and scalable analysis of the big data. In this regard, the main goal is to study and brainstorm on (1) how crowdsourcing data could be collected through big data technology as well as (2) how already collected big data can be analyzed through efficient and scalable solutions. This fits well to the expertise of Petanux as the guest institute who delivers technology and big data expertise as well as Web2Learn as a host institute and provides professional expertise in social science and e-learning. The final outcome of this STSM should be a sort of recommendation and roadmap report on efficient involvement of larger societies through crowdsourcing.

Furthermore, BiCSkill aims to critically discuss and realize feasibility study on frameworks for an understanding of aforementioned topics with researchers at the host institution. Set a basis for exploration of MapReduce, Hadoop, Spark, and further technologies and Computer Assisted Language Learning (CALL) paradigms around incentivized language learning, crowdsourcing mechanisms and the impact on users-learners and on content (user-generated and institutionally-provided).

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

In the following the tasks carried out during the stay at this STSM are listed and each one is described in details:

1- **Analysis and discussion on definition of basic terms and domain knowledge in Automation of language training:** With the rise of social media and e-commerce, big amounts of data, mainly texts such as messages, emails, comments, and articles are being generated. Accordingly, we need efficient methods to retrieve useful information from the texts. Supporting humans in analyzing such disparate amounts of data is important. This demands a need to automate classification of linguistic texts in various languages being generated in the web and produce meaningful and efficient knowledge. This can even be used in training of languages for professionals and all other users. Recently, there has been emerging technologies in Artificial Intelligence (AI) and in the field of Natural Language Processing (NLP) to target stated problem. NLP comprises of different functions such as search, clustering, classification, summarization, information monitoring and retrieval. Among these, classification and clustering have some similarities in the functional- ity. Both of them try to find which input belongs to which category. The difference here, is that in the former, classification, the set of labels of data is known and for the latter, in clustering, during processing, the classes are made and the labels are unknown.

2- **Discussion on what and how Machine Learning can support crowdsourcing for language learning:** We already discussed and analyzed how ML can support language learning by means of crowdsourcing? What algorithms from ML can be used? What could be the dataset from language learning for crowdsourcing of ML algorithms? Machine learning can efficiently support crowdsourcing for language learning activities. It should be stated that language learning can be efficiently benefited from the use of combined machine learning and crowdsourcing. This will be resulted mainly in the annotation of training datasets through ML algorithms. Accordingly, ML and unsupervised algorithms can very well train and generate enough datasets for language learning goals. Furthermore, generative algorithms such as Generative Adversarial Networks (GANs) can even generate materials for language goals. One simple example of this is like fake fact, which has been highlighted recently. Similarly, generative algorithms can generate materials for the goals of EnetCollect. This is more and more important for those less used languages in the Europe. These languages mainly miss enough datasets, tutors and annotators in crowdsourcing world. So, ML and generative algorithms can easily support their missing role.
   As a conclusion, machine learning can support EnetCollect in evaluating and improving the quality of results achieved from crowdsourcing. Currently, EnetCollect team has tested crowdsourcing with the help of some dozens of language speakers (see Pybossa experiments) so the results achieved so far cannot be scaled up. Also, in the crowdsourced experiments so far, the dimension of training the machine (as in Deep-L) by users in a crowdsourcing fashion has been totally ignored (the link between crowdsourcing and machine learning is absent).

3- **Discission and realization of usefull approach for automated generation methods on Learning of less-used Languages:** We also realized another interesting idea about less-used languages. The main point is about how generative algorithms such as GAN can help producing/enriching corpora in less-used languages?! GANs can generate enough and artificial datasets for less-used languages in order to have convincing datsets in terms of volume and quality for training of ML algorithms.

4- **Discussion on Potentials of ML for Question Answering in Language Learning:** Another important discussed and highlighted topic was about how ML can help in question-answering tasks in a language learning context.

5- **Discussion and realization of concepts for post-processing activities in language learning:** In addition to pre-processing activities already mentioned, there is also great opportunity for post-phase actions of using ML in the language training. For instance, using crowdsourcing users can assess ML algorithms, refine/ adjust results produced by ML for already prepared language training materials. In this regard and in the case of using supervised algorithms, users can label training data for supervised ML algorithms. In the frame of this meeting, we also created a proper and relevant bibliographic resources, which are under review of participant and is also recommended for further investigation by the cost action members. The list of the recommended bibliography is already included in the next section.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

After having multiple discussions and meeting, we already defined different tasks that have been resulted in or can be resulted in various outputs. These include, but not limited to the following:

1- **A result in conception of methodology for transferring lessons learned and expertise and outcomes from enriched languages to less-used languages using Transfer Learning (A type of ML method):** As an example of how ML and mainly NLP can support language training and transferring from one to another, The main task of [Bahdanau et al., 2015] is translating English corpus to French one. Machine Translation that are the member of encoder-decoder family. Encoder converts the input sentence to a fixed-length vector and Decoder prepares and generates the output (translation) from that vector. Both decoder and encoder networks learn to gather. The substantial problem here, is about the context vector with fixed length. So, in this work instead of using one fixed-length vector as a context vector, is proposed encoder makes a sequence of vectors and decoder chooses the required subset. For achieving the current target output, by considering the current input state and the previous outputs, the attention mechanism computes weights corresponding to each vector. Ergo, for each time step the value of vectors are distinct. Both encoder

and decoder take advantage of BRNN, so, this model can get variable size input and make variable output, too. For the next output this model just considers the informative and meaning relations.

2- **Carefull Review and analysis of literature in this area and shortlisting and recommending the most relevant ones for further consideration:** Further recommended literature is as following:

- [Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). NEURAL MACHINE T RANSLATION. pages 1–15.
- [Bowman and Potts,] Bowman, S. R. and Potts, C. A large annotated corpus for learning natural language inference.
- [Cao et al.,] Cao, Y., Fang, M., and Tao, D. BAG: Bi-directional Attention Entity Graph Convolutional Network for Multi-hop Reasoning Question Answering.
- [Chen and Durrett,] Chen, J. and Durrett, G. Understanding Dataset Design Choices for Multi-hop Reasoning.
- [Chen et al., 2019] Chen, J., Hu, Y., Liu, J., Xiao, Y., and Jiang, H. (2019). Deep Short Text Classification with Knowledge Powered Attention. (2017).
- [Chorowski, 2014] Chorowski, J. (2014). Attention-Based Models for Speech Recognition. pages 1–9.
- [Dong et al., 2016] Dong, F., Zhang, Y., and Yang, J. (2016). Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring.
- [For, 2017] For, L. O. W. (2017). B -d a f m c. pages 1–13.
- [Gong and Bowman, 2017] Gong, Y. and Bowman, S. R. (2017). Ruminating Reader: Reasoning with Gated Multi-Hop Attention.
- [Kim, 2011] Kim, Y. (2011). Convolutional Neural Networks for Sentence Classification.
- [Kumar and Rastogi,] Kumar, A. and Rastogi, R. Attentional Recurrent Neural Networks for Sentence Classification. Springer Singapore.
- [Lan et al., 2017] Lan, M., Wang, J., Wu, Y., Niu, Z.-y., Wang, H., and Engineering, S. (2017). Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification. pages 1299–1308.
- [Lewis et al., 2004] Lewis, D. D., Rose, T. G., and Li, F. (2004). RCV1 : A New Benchmark Collection for Text Categorization Research. 5:361–397.
- [Lstm-cnns crf, 2016] Lstm-cnns crf, E.-t.-e. S. L. B.-d. (2016). End-to-end Sequence Labeling via Bi- directional LSTM-CNNs-CRF.
- [Pfister, 2018] Pfister, T. (2018). Attention-Based Prototypical Learning. (1):1–19.
- [Ravichander et al.,] Ravichander, A., Naik, A., Rose, C., and Hovy, E. EQUATE : A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference.
- [Rush, 2017] Rush, A. M. (2017). Tructured ttention etworks. pages 1–21.
- [Socher et al., 2013] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.
- [Song and Wang,] Song, Y. and Wang, J. Attentional Encoder Network for Targeted Sentiment Clas- sification.
- [Tai et al.,] Tai, K. S., Socher, R., and Manning, C. D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks.
- [Tang et al., 2015] Tang, D., Qin, B., Feng, X., and Liu, T. (2015). Effective LSTMs for Target-Dependent Sentiment Classification.
- [Tran and Niederée, 2018] Tran, N. K. and Niederée, C. (2018). Multihop Attention Networks for Ques- tion Answer Matching. pages 325–334.
- [Trivedi et al., 2018] Trivedi, H., Kwon, H., Khot, T., Sabharwal, A., Balasubramanian, N., and Brook, S. (2018). Repurposing Entailment for Multi-Hop Question Answering Tasks.
- [Wang et al.,] Wang, Q., B, H. Y., Wang, W., Huang, Z., and Guo, G. Multi-hop Path Queries over Knowledge Graphs with Neural Memory Networks, volume 1. Springer International Publishing.
- [Wang and Manning, 2012] Wang, S. and Manning, C. D. (2012). Baselines and Bigrams : Simple , Good Sentiment and Topic Classification. (July):90–94.
- [Wang et al., 2018] Wang, Y., Di, X., Li, J., Yang, H., and Bi, L. (2018). Sentence Similarity Learning Method based on Attention Hybrid Model Sentence Similarity Learning Method based on Attention Hybrid Model.
- [Wang et al., 2016] Wang, Y., Huang, M., Zhao, L., and Zhu, X. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. pages 606–615.
- [Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. pages 1480–1489.

- [Yin et al.,] Yin, W., Sch, H., Xiang, B., and Zhou, B. ABCNN : Attention-Based Convolutional Neural Network for Modeling Sentence Pairs.
- [Yu et al., 2018] Yu, W.-c. C. H.-f., Dhillon, I. S., and Yang, Y. (2018). SeCSeq : Semantic Coding for Sequence-to-Sequence based Extreme Multi-label Classification. (2):2–7.
- [Zhou et al., 2015] Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). A C-LSTM Neural Network for Text Classification.
- [Zhou et al., 2016] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. pages 207–212.

FUTURE COLLABORATIONS (if applicable)

It is supposed that STSM partners plan further research projects and proposals in order to realize and implement crowdsourcing and machine learning techniques in the field of computer aided language learning and documentation. This will mainly benefit less used languages in Europe and will provide online and autonomous platforms for their usage and training. Furthermore, STSM partners aim to involve major partners of this cost action in their project proposals in the frame of Erasmus+ program for key action 2 towards facilitating mobility of professionals in Europe and overcoming the languae shortcoming in the mobility of professionals.
Furthermore, this STSM partners may aim in preparing a conference publication to conceptualize the machine learning utilizing crowdsourcing in the frame of EnecCollect cost action. The main goals is to target one of this action's scientific events such as workshops or conferences.