SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: CA16105
STSM title: Evaluation of reading comprehension questions using crowdsourcing techniques
STSM start and end date: 04/03/2018 – 10/03/2018
Grantee name: Itziar Aldabe

PURPOSE OF THE STSM/

The aim of this STSM has been to establish the first steps to define a general strategy to use crowdsourcing to evaluate Reading Comprehension. With this purpose, we defined the following tasks:

1. To summarize the state-of-the-art on question generation and the procedures to evaluate the quality of the systems.
2. To establish an evaluation procedure to measure the quality of automatically generated questions for reading comprehension.
3. To establish evaluation guidelines with experts in the area

During the one week STSM, we aimed at summarizing state-of-the-art question generation systems and their evaluation, so that we would have a general overview of the possible solutions and their applicability in our scenario. Tasks 2 and 3 were established as future collaborations.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

Question Generation (QG) and its evaluation is a challenging task that has attracted a number of researchers over the last years. In 2008, the first Workshop on Question Generation (Rus and Graesser 2009) was held and this was the starting point for this ever-increasing community. QG is defined (Rus and Graesser 2009) as the task of automatically generating questions from some form of input. Among the various tasks which have been proposed, we focus on the text-to-question task. As regards the evaluation measures, (Rus and Graesser 2009) mention that QG systems can be evaluated either manually or automatically. In the STSM, we explore both options with special emphasis on crowdsourcing. The use of crowdsourcing to evaluate QG systems is a relatively new approach that could be of interest to the cost action. It is important to design crowdsourcing strategies to evaluate the creation of language learning materials. Concretely, we find interesting to start working on Reading Comprehension as it is not only a specific research area for language learning but for any learning area in education.

During the STSM, the grantee carried out the following tasks in order to work on the evaluation of

Funded by the Horizon 2020 Framework Programme
of the European Union

reading comprehension questions using crowdsourcing techniques:

1. Virtual meeting with experts on questions generation and crowdsourcing.
2. Presentation of previous work on exercise generation to the members of Språkbanken (the Swedish Language Bank)
3. First steps on the description of the evaluation procedure to measure the quality of automatically generated questions using crowdsourcing.

As a result of these tasks, we detected and listed various factors to be defined so that the evaluation procedure using crowdsourcing is correctly built:

1. Evaluation metrics.
    1. Automatic metrics for evaluating question generation systems: BLEU (Papinei et al., 2002), METEOR (Denkowski and Lavie, 2014), NIST (Doddington, 2002)
    2. Comparison of automatically generated questions with manually generated ones (Chinkina and Meurers, 2017)
    3. Evaluation with X experts in questions and pedagogy and evaluation of questions generated by the system. Inter-rater reliability (Olney et al., 2012)
    4. Best Worst Scaling (Flynn and Marley, 2014)
2. Examples of the use of crowdsourcing and question generation (Labutov et al, 2015; Rajpurkar et al., 2016; Chinkina and Meurers, 2017)
3. Crowdsourcing scenario. Variables to be defined in advance:
    1. Target users: experts
    2. Questions to be answered in the crowdsourcing task.
    3. Size of the passage.
    4. Number of workers per question
    5. Item scale

References:

Chinkina, M., & Meurers, D. (2017). Question Generation for Language Learning: From ensuring texts are read to supporting learning. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Copenhagen, Denmark.

Denkowski M. and Alon Lavie A., 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language,Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.

Doddington G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. HLT '02 Proceedings of the second international conference on Human Language Technology ResearchPages 138-145

Flynn T.N., and Marley A.A.J., 2014. Best-worst scaling theory and methods. in: Handbook of Choice Modelling, chapter 8, pages 178-201 Edward Elgar Publishing

Labutov I., Basu S., Vanderwende L. 2015. Deep questions without deep understanding. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15)

Olney A., Graesser A., and Person N. 2012. “Question Generation from Concept Maps.”D&D 3 (2012): 75-99.

Papinei L, Roukos S., Ward T., and Zhu W. 2002. BLEU: a method for automatic evaluation of machine translation. In ACL 2002, pp 311-318

Rajpurkar P., Lopyrev K., Zhang J., and Liang P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing

Rus, V., and Graesser, A., eds. 2009. The Question Generation Shared Task and Evaluation Challenge. Sponsored by the National Science Foundation. ISBN 978-0-615-27428-7.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The main outcomes of the STSM can be summarized as follows:

1. A first version of the evaluation procedure to measure the quality of automatically generated questions for reading comprehension. We have decided to compare the quality of automatically generated question against manually generated ones by distributing both types of questions to crowdworkers. For that, we have detected different evaluation measures that will be taken into account when defining the final version of the evaluation procedure.
2. A possible collaboration network between four different researchers:
    - Itziar Aldabe, IXA research group, University of the Basque Country (UPV/EHU), Spain
    - Elena Volodina, University of Gothenburg, Sweden
    - Maria Chinkina, LEAD Graduate School, Universität Tübingen, Germany
    - Andrea Horbach, Language Technology Lab, University of Duisburg-Essen, Germany

FUTURE COLLABORATIONS (if applicable)

After the STSM, we have agreed to continue working on the following tasks:

1. To establish the evaluation procedure to measure the quality of automatically generated questions for reading comprehension.
2. To establish evaluation guidelines with experts in the area.
3. To write a report which contains: a) a summary of evaluation on question generation using crowdsourcing; b) the evaluation procedure to measure the quality of automatically generated questions for reading comprehension, and c) the evaluation guidelines.

In order to fulfill the tasks we have defined the following time-line:

1. To Introduce the idea in enetCollect in autumn 2018.
2. To finish the report before March 2019.
3. To present the report in enetCollect in spring 2019 and involve people from the COST action in

the evaluation of questions for Reading Comprehension.