cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

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: State-of-the-art review of implicit crowdsourcing approaches and experiments for collecting language-related data
STSM start and end date: 05/02/2020 to 21/02/2020
Grantee name: LAVINIA-NICOLETA APARASCHIVEI

PURPOSE OF THE STSM:

As the main objective of this STSM was aimed at designing a prototypical experiment,it mainly to the deliverable D2.4 ("Design plans, implemented prototypes and related evaluation reports"). In that context, this STSM allowed to speed up the on-going proof-of-concept efforts aiming at developing a prototypical implicit crowdsourcing vocabulary trainer by planning an experiment on the Romanian version of the language resource on which the vocabulary trainer relies (ConceptNet).

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

Recent efforts started during the Task 3 of the 2019 Crowdfest have allowed to design a prototypical vocabulary trainer whose content is automatically generated from ConceptNet and for which the answers of the students are used to both improve the ConceptNet data used to generate the vocabulary exercises and extend ConceptNet with new entries. ConceptNet is available for numerous languages (currently 304). For pragmatic reasons, the first experiments were performed with English as it is the language best covered. During this STSM, the conditions were prepared to reproduce in the second semester in spring 2020 the past experiments made for English for a lesser-resourced language from the Romance family: Romanian. On a scientific level, new issues related to the available ConceptNet data and the exercises that will be generated from them were observed.. On a practical level, this STSM fine-tuned the preparation to run the experiment and prepared new approaches to extend the Romanian ConceptNet.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The main results obtained are of two types.

- Organizing details for the experiment : it was decided to aim at running the experiment as early as possible in April with a special of foreigners students. With regards to the number of students to involve, it will depend on the amount of time each one of them will be involved. At present, no precise numbers have been established as first contacts must be first made with Romanian teachers. Nonetheless, we hope to gather some 10 answers for an amount of 2000 questions.

- Development of four new approaches to use for the vocabulary trainer, of which three were tested. First an approach to favour the inclusion of new words in ConceptNet. This first approach is based on frequency lists extracted from Sketchengine. Second and third, two approach to generate automatically with decent quality from ConceptNet automatically synonymy questions about synonyms which expected answers is YES or NO, with decent quality. For NO questions, the relations between words in Romanian that do not link synonyms were used in same languages that are not synonyms was used. By doing so, the same semantic

landscape of the pairs of words it was preserved. For YES questions, translations in other languages to jump from the target language to a different language and back to the target language were used. First prototypical implementations in both cases allowed to generate questions for which the expected answer was 85% correct for English (as a test case) and with a semantic relatedness between words that were considered as satisfying.

- Finally, a fourth approach that will be developed in the future, but which was debated during this STSM was to generate a list of words on which the questions should focus to consider for the questions based on a by reusing a first list of seed terms that would be provided by a teacher (for example (car, moto) => (wheel, road, highway, gasoline)).

FUTURE COLLABORATIONS (if applicable)

The following steps is, on the one hand, to run with the experiment on Romanian while, on the other hand, complete the proof of concepts for the four approaches mentioned above for a later integration in the vocabulary trainer.

The results obtained during this STSM and the results that will be obtained from the crowdsourcing experiment will be used in the Master dissertation work of the applicant.The applicant is also being invited to attend and contribute to the WG2 meeting organized in Naples in March of this year.

To whom it concerns, as the host of this STSM, I hereby approve this report on the day of 26/02/2020

Best regards,

Lionel Nicolas