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: 17/06/2019 to 31/07/2019
Grantee name: LAVINIA-NICOLETA APARASCHIVEI

PURPOSE OF THE STSM:

The main objective of this STSM was to develop a comprehensive literature review of the state of the art for implicit crowdsourcing approaches and experiments, with a special focus on the technique used for collecting language-related data such as NLP datasets (Crowdfest task 4, D2.2).

In the context of this STSM, next to the literature review, first steps were made to consider and devise approaches to combine language learning and implicit crowdsourcing, such as the ones that generate language learning exercise content from datasets and use the answers of learners to improve the datasets used (D2.3). Among the approaches considered and/or devised, one approach has been selected and a prototypical implementation has been planned (cf. Crowdfest task 3 and D2.4) to crowdsource Romanian NLP resources. (It is envisioned to implement this prototype in the context of a Master's thesis after the end of the STSM.).

Also, in the context of this STSM, the applicant has approached a practical programming task of implementing an enhanced visual display (based on the javascript library D3.js) of the interlinking of datasets and learning tasks allowing stakeholders interested in implementing prototypes to identify other stakeholders to collaborate with (c.f Crowdfest task 4 and D2.4).

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

The Zotero bibliography was populated step-by-step by first importing google search results for articles with specific keywords "GWAP" /" citizen"/"TURK" /" Crowdflower" published at the 2018 and 2016 editions of LREC and at the 2019 edition of ACL.

Such a first step allowed to collect a first set of references from the abovementioned conferences and a second set of references through the ones mentioned in the previously found articles A second, more fine grained, step was then implemented through a google query limited to specific parts of the conferences' websites (e.g."crowdsourcing OR amt OR amazon OR mechanical OR turk OR crowdflower OR crowd site:http://www.lrecconf.org/proceedings/lrec2016/summaries"). At the same time, some descriptions about different trends of crowdsourcing, namely *Human computation, Citizen Science, Explicit crowdsourcing, Game with a Purpose, Implicit crowdsourcing, Wisdom of the crowd,* were compiled. During the next steps, 17 articles were selected and studied, of which 4 were judged as non relevant to WG2 and 13 were summarized and labeled. While studying these articles, we also developed and refined a template and a methodology to help us formalize the review procedure, which in turn led to several revisions of the summaries compiled for these articles. At the same time the list of potentially relevant publications was regularly updated by briefly checking and evaluating the relevance of the references found in articles studied, and the articles referencing them.

In parallel, several tutorials were studied by the applicant on https://www.w3schools.com/js/default.asp to upgrade her programmatic skills. A D3.js visualization was also selected and partially adapted to the needs of the "galaxy" display foreseen by the Crowdfest task 4.

Also, recurrent discussions were held to plan the foreseen crowdsourcing experiment aiming at crowdsourcing Romanian NLP datasets during the first half of 2020.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

Literature review:
Until now, the number of the articles in the "Candidate List" reached **140 articles, 13 of which have been summarized and labelled and 4 of which were evaluated as not relevant to WG2**. By developing a methodology and a template that we followed, we improved the quality of the summaries of the articles that belong in the WG2 bibliography. Entries of all articles in the "Candidate List" have been added to the "enetCollect bibliography-WG2 review" Zotero group. The ones that were not yet evaluated in terms of relevance were added to the sub-collection "Relevance to be evaluated", while the non-relevant ones were added to the sub-collection "Non-relevant" and the summarized ones were added to the "Validated/To be transferred" sub-collection.

Visualization of literature data:
After improving the programmatic skills of the applicant, we also started to adapt an existing D3 visualization to display"galaxy" model of the data exported from the Zotero Library.

Planning an implicit crowdsourcing experiment
After involving the applicant in an online meeting regarding the Crowdfest Task 3, it has been decided to adapt the approaches and tools developed by this initiative to run an experiment for crowdsourcing Romanian datasets during the first half of 2020.A planning has been devised accordingly.

FUTURE COLLABORATIONS (if applicable)

The following steps will be taken to expand the WG2 bibliography. Every week, 2-3 summaries of interesting articles for WG2 bibliography will be added. The applicant will also keep on implementing the enhanced visual display (based on the javascript library D3.js) of the interlinking of datasets and learning tasks (cf. Crowdfest task 4) allowing stakeholders interested in implementing prototypes to identify other stakeholders to collaborate with (Task 4 and D2.4). The applicant will also be invited to attend and contribute to the WG2 meeting organized in Malta in November of this year. In the context of another STSM foreseen in February, the applicant and hosts intend to fine tune an experiment that will focus on existing resources of the Romanian NLP repertoire and for which the targeted crowds will consist of Romanian students with the aim of implementing the experiment during the first half of 2020.

To whom it concerns,

as the host of this STSM, I hereby approve this report on the day of 05/08/2019.

Best regards,

Lionel Nicolas