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

Action number: CA18209

Grantee name: Kirill Yankov

Details of the STSM

Title: Creation of Lexical Knowledge Bases based on Wikipedia Content

Start and end date: 02/05/2022 to 21/05/2022

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)

Week 1 (2 – 8 May). Design phase.

The first week was dedicated to getting familiar with the knowledge sources (Wikipedia/DBpedia NIF, labels, redirects and disambiguations datasets) and the OntoLex Lemon model. Design and modelling of mapping of the data from the knowledge sources to OntoLex Lemon were done. The architecture design of the framework is intended to be scalable and modular. We chose Apache Spark as the data processing engine, as it has the advantages of scalability and distributed processing: can be run on single server or on a cluster which allows it to easily scale for processing of large amounts of data.

Week 2 (9 – 15 May). Implementation phase.

The second week was dedicated to the implementation of the designed framework. The implementation included development of an application in the Scala programming language with the use of Apache Spark framework. The application consumes datasets' files in N-Triples format as input and outputs processing results for each of the datasets in a form of N-Triples files, and also merges all the results into a single N-Triples file so that it can be used as a single lexical data database. The files then can be consumed by any RDF-processing engine for further querying and analysis.

The framework extracts lexical data from the aforementioned datasets: concepts and links to their descriptions, synonyms and polysemantic words, links lexical senses of the

[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.

synonyms/polysemantic words. The process is implemented in a form of pipeline: the extracted data is fed to a mapping engine, which builds graphs of linked words and senses, maps it to the OntoLex Lemon model, and then outputs the result.

For the ease of executing the framework a Dockerfile build was implemented. It allows to run the framework without installation of build tools for programmers.

The framework is implemented in a modular and functional way, which enables easy transformation and addition of new features like support of other sources of the data or data filtering. The framework can be run for any language, the prerequisites are: availability of the datasets for that particular language.

Week 3 (16 – 21 May). Testing and documentation phase.

The third week was dedicated to the testing, generation of the dictionaries in the OntoLex Lemon model for five different slavic languages (macedonian, serbian, bulgarian, czech, slovak) and writing documentation. The testing was performed by running the framework on different input data and validation of the output results. There were discovered and fixed several problems concerning running the framework in different environments, in particular: out of memory for some data, errors during processing compressed data files, errors in different file systems. The generated data was tested with the rapper RDF validation tool for syntactical correctness.

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)

All the goals of the STSM were reached. The development outputs of the STSM are:

- General framework for extraction of linguistic information for various languages from Wikipedia knowledge base. As well as Dockerfile for building and running the software in a Docker container. The source code of the software is maintained on github: https://github.com/m1ci/dbpedia-lex. The examples of inputs and outputs are available on the repository README page.
- Novel lexical datasets for five selected languages (macedonian, serbian, bulgarian, czech, slovak). The datasets are temporarily published under CC BY-SA licence (https://drive.google.com/drive/folders/1mUBt1yT8x8ailnTuAINGxFlN3nbP_TpR?usp=sharing) with the plan to publish them in near future as part of the LLOD cloud.

The developed framework and the derived datasets from the STSM serve for a greater support for under-resourced languages, which is one the main Action objectives, and to the following tasks from Working Group 1: Task 1.1 (use of Ontolex-lemon), Task 1.2 (creation of new datasets for the LLOD cloud.) and Task 1.5 (addressing low-resourced languages).

Planned follow-up activities:

- Statistical analysis of the generated lexical data and its quality, comparison to other similar datasets.
- Research on methods of improvement of the quality of the generated data.
- Publication of a scientific paper based on the conducted work.
- Presentation of the results of the STSM on one of the follow-up WG meetings.