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

Action number: CA18209

Grantee name: Maxim Ionov

Details of the STSM

Title: Development of cross-lingual linking pipelines in a linked data context

Start and end date: 28/03/2022 to 27/04/2022

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.

My work during the stay was focused on creating reproducible pipelines for conversion and linking of lexical resources.

More specifically, we were preparing the release of a new version (v2.3) of the Apertium RDF dataset — a collection of 53 bilingual dictionaries (including many less-resourced languages, both spoken in the EU and outside of it, such as Galician and Asturian or Kazakh (https://www.apertium.org/). The conversion was done in a reproducible way through the Fintan framework (https://github.com/Pret-a-LLOD/Fintan), which allows recreating the dataset regularly, since the source data is being constantly updated by the Apertium community. An important part of the work carried out during the stay was the discussion and fixing of known issues in the Apertium RDF data, and their iterative validation/curation for the new version v2.3 which will be published as data dumps on Zenodo and GitHub initially, and later as a SPARQL endpoint.

Another work direction was to employ this component (Apertium RDF transformation) as a first component of a larger pipeline, compatible with the Teanga workflow manager (Teanga is an LD-aware framework to build NLP workflows, developed in the context of the Prêt-à-LLOD project, https://github.com/Pret-a-LLOD/teanga).

The second component in this pipeline was OTIC-Link, a translation inference tool previously developed in UNIZAR to infer translations between language pairs connected through a pivot language, and used to compute baselines in the TIAD campaign series (https://github.com/Pret-a-LLOD/OTIC-api). During

[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.

the stay, I collaborated with researchers of the host team in the adaptation of such a tool into a RESTful API service, pluggable into Teanga pipelines.

We have also considered using data validation mechanisms according to Teanga specification, but found it inapplicable to the OTIC workflow.

Another pipeline that I have participated in developing was a link discovery service that connects Apertium RDF and BabelNet (https://babelnet.org/) to enrich the Apertium RDF data with additional sense information that can be found in BabelNet. My main contribution has been its adaptation to make it Teanga-compatible, with BabelNet-linking component being transformed into a RESTful API service.

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.

Most of the tasks planned for the STSM has been accomplished:

- An improved pipeline for converting Apertium to RDF and linking it with LexInfo vocabulary. Most of the reported issues were fixed and the data of the new version (v2.3) has reached a mature state, almost ready for its publication after a final validation round.
- Optimisation of OTIC RESTful API service and making it compatible with Teanga workflow manager.
- Transformation of BabelNet sense linking components to a RESTful API service in a way that is compatible with Teanga workflow manager.

An exploration of data validation for Teanga pipelines showed that for some cases it is not relevant (e.g. for Apertium-OTIC pipeline), but applicable for others (Apertium-BabelNet pipeline). The publication of datasets created as one of the results of this STSM is still underway.

As a follow-up collaboration, we are planning to introduce more flexibility to the Apertium-BabelNet pipeline, allowing users of this pipeline to choose, whether they want sense enrichment for the whole dictionary, or for the list of lexical entries (currently, it is only possible for a single translation at a time).

Finally, as mentioned above, some linking components have been adapted in this collaboration to be integrated in Teanga pipelines. As a further step, a real, practical, implementation of a pipeline based on Teanga that will make use of such linking components will be carried out, in close collaboration with UNIZAR and NUIG.

The latter might be the basis for a joint publication that will describe the main achievements of this STSM.