Interlingual Information Extraction in Temporal Web Collections

STSM Report

Dr. Elena Demidova
STSM duration: 20.03.2017- 24.03.2017
Host: ISST Laboratory, ITMO University, St. Petersburg, Russia

Purpose of the STSM

The goal of this STSM was to initiate a joint grant application between the L3S Research Center and the ISST Laboratory in the area of multilingual Information Extraction for the joint funding program of the DFG (German Research Foundation) and Russian Foundation for Basic Research – RFBR (russ. RFFI).

Description of the work carried out during the STSM

The work during the STSM was carried out according to the work plan. In particular, several discussions regarding the joint grant application and joint publications were conducted.

1. Discussion of the current state of preliminary work at the ISST and L3S.

The work started with the presentations of the ISST and L3S current projects. On the L3S side this included recent work of Dr. Demidova on interlingual information alignment in Wikipedia (to appear in the ACM TWEB) as well as case studies to analyze user interactions with multilingual information. On the ISST side that included a variety of projects in the areas of Semantic Web and NLP, in particular in the financial domain.

2. Discussion of DFG and RFFI requirements.

Further discussions included exchange on the DFG and RFFI requirements. Whereas on the German side, DFG applications are considered in the common open call for basic research, RFFI on the Russian side publishes regular calls. Further open issues regard integration of the proposal. As a follow up a need for the discussion with the research support office was identified to determine an appropriate level of integration of the partner contributions.

3. Discussion of the topical focus of the proposal from the L3S and ISST perspectives.

Based on the initial presentations, possible topical focus for a joint grant application were discussed. In particular, whereas ISST focuses on the NLP processing and semantics for the Russian language, L3S is interested to continue its line of research in interlingual information alignment, including German, English and in particular Russian languages in this collaboration. Application areas of interest for both institutions in this context include financial domain and interlingual Question Answering. In particular, the development of models for interlingual event-centric processing for the English and German languages and their adaptations to include the Russian language will build the core of the L3S application part.

4. Planning the next steps for joint publications on the topic of interlingual Information Extraction from temporal Web collections to strengthen preliminary work for the proposal.

In order to strengthen the preliminary work for the application, joint publications, in particular with the focus on interlingual event-centric news processing were discussed and planned. In particular an article regarding event-centric information extraction and alignment on multilingual news was planned to be submitted to a leading Semantic Web or NLP conference later this year. Event-centric multilingual information models building the core of the publication will contribute to the preliminary work of the joint grant application.

5. Creating a proposal draft and a time plan for writing.

Whereas on the German side, DFG applications are considered in the common open call for basic research (i.e. submission is possible at any time), RFFI on the Russian side publishes regular calls. The time plan for writing will be further detailed and aligned with one of the next RFFI calls as they become available.

6. Creating an overview of the datasets and tools available at L3S and ISST.

During the STSM relevant datasets and tools available at L3S and ISST were discussed. On the L3S side, the project "ALEXANDRIA", focused on the entity-centric and event-centric information processing in Web archives provides access to large scale archived Web datasets. Relevant tools developed at L3S include iCrawl (focused event-centric Web and Web archive collection creation) and MultiWiki (cross-lingual text passage alignment for Wikipedia). ISST provides tools for NLP in Russian language and ontologies in several application domains.

7. Creating an overview of related approaches.

Related work covers several areas and includes approaches to machine translation, interlingual classification and clustering, analyzing interlingual differences in comparable corpora, interlingual alignment at different levels of granularity, as well as interlingual text reuse and plagiarism detection. These areas will be discussed in the joint application in more detail.

Description of the main results obtained

- An overview of the common research interests between the L3S and ISST.
- A plan for the joint grant preparation for DFG and RFFI.
- A plan for the continuation of work on the joint publications to strengthen preliminary work for the application.
- An overview of the most related datasets and tools at both institutions.

Future collaboration with the host institution

- Further possibilities for collaboration discussed during the STSM include participation of ISST in H2020 programs, in the context of future proposals by L3S.

Foreseen publications/articles resulting from the STSM

- A joint publication on the topic of event-centric multilingual information extraction from multilingual news.

Confirmation by the host institution of the successful execution of the STSM

See attached confirmation letter by ISST.

Other comments (if any) None.