Short-Term Scientific Mission (COST-STSM-IC1302-34073) – REPORT

KEYSTONE COST Action IC1302

STSM Applicant: Liubov Kovriguina, Ph.D.

Host Institution: Dr. Elena Demidova, Web and Internet Science Group, Department of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

Duration: 2016-0-6-27 – 2016-07-11

**STSM Title: Clustering Similar Events Across Languages Using Simple Event Ontology Population**

## 1. Purpose of the STSM:

The purpose of the STSM was the development of methods and algorithms enabling extraction and clustering of equivalent and related events from the news articles across languages. The results to be obtained contribute to several unsolved problems: 1) paraphrase detection, 2) similar events clustering, 3) process mining and 4) multilingual natural language processing.

## 2. Description of the work carried during the STSM:

During the visit to the WAIS Group I had the opportunity to work with Dr. Elena Demidova to join our expertise for discussing novel applications of events clustering. We started by discussing how news aggregation algorithms can be improved (similar events clustering) and how alternative points of view on the same fact in newswire texts can be automatically extracted and grouped (multilingual events clustering). The discussion resulted in a decision to develop a prototype for automatic scenarios mining (process modeling), see system pipeline in figure 1. We decided to include news on the following topics into the dataset for event extraction and process modeling: Olympic Games 2016 (and the pre-Olympic reports), USA elections, Turkey coup of 2016 and some others.

The approach combines statistical and semantic approaches and cluster similar news articles using the Simple Event Model Ontology as a fact model, populating this ontology with different classes of named entities and qualify events that have the same actors, place and time as similar or related. This allows to cluster predicates of the events, which surface syntax and lexica

are different (paraphrases). When all similar events are extracted and clustered for a given topic, it is possible to specify main states of the process using the size of the cluster. Given the process states, the process can be described using Markov models.

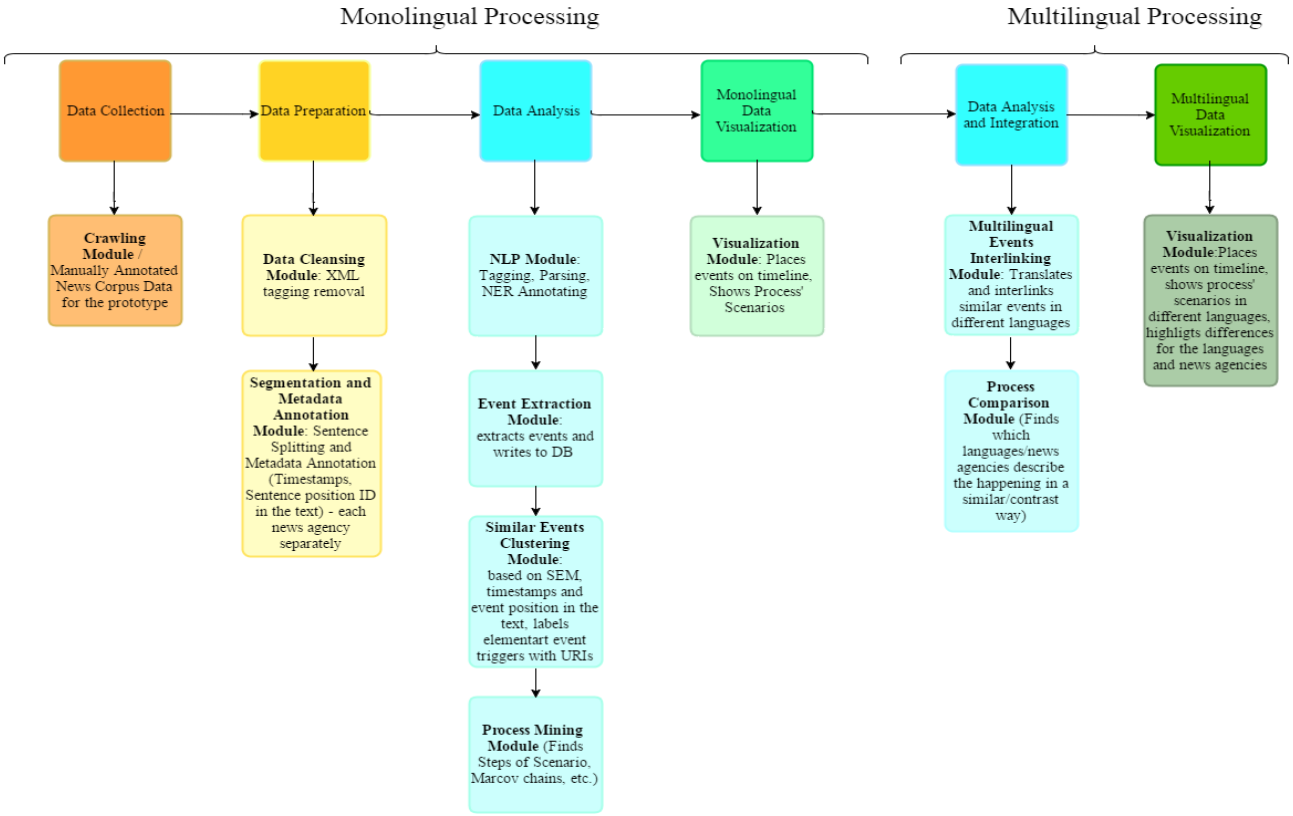


Figure 1. Pipeline of Multilingual Process Mining System.

Among detailed aspects elaborated during the STSM are:

- Overview of existing "event" notions and models used for the task of event extraction;
- Overview of existing event ontologies and ontologies describing temporal and causal relations between events;
- Overview of standards and approaches to event annotation;
- Overview of machine learning algorithms used for paraphrasing and sentence similarity computing;
- Overview of process mining algorithms;

I also had a chance to present and discuss research projects of ISST Laboratory, ITMO University with the staff and Ph.D. students of WAIS Group and got valuable feedback.

## 3. Description of the main results obtained

(i) a draft of the system architecture,

(ii) database scheme,

(iii) a draft of the paper.

## 4. Future collaboration with the host institution (if applicable)

Together with Dr. Elena Demidova we plan to continue research activities in the field of text mining systems. We plan to focus on development of algorithms for automatic process modeling and scenario mining from unstructured data.

5. **Foreseen publications/articles resulting from the STSM (if applicable).**

(i) a paper concerning automatic paraphrase extraction from the clusters of similar events to be submitted to the European Chapter of the Association for Computational Linguistics (EACL 2017), evidence from news articles,

(ii) a paper on multilingual process mining, evidence from news articles in Russian, Spanish and English.

6. **Other comments (if any)**