Temporal Entity-Centric Information Propagation in Multilingual Texts

KEYSTONE STSM - Keystone COST Action IC1302

Applicant: Simon Gottschalk, L3S Research Center,
Leibniz Universität Hannover, Germany

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

Period: October 3 – October 14

Purpose of the STSM

Real-world happenings and their implications are reported in several sources like social media, knowledge bases or Wikipedia. In the latter case, such happenings can result in permanent changes in the description of the related entities (such as persons, locations, organisations, etc). In most cases, this information is originally adopted from news reports that play an important role in this process of information propagation consisting of the initial happening, the following reports and the final adoption in the aforementioned resources. Given the Internet as a global resource bringing together news from multiple cultures and language communities, this process becomes more complex, including information asymmetries and controversies.
To this end, a language-independent model representing events, their media coverage and adoption helps to understand how events and their consequences are described in secondary sources like Wikipedia and where the actual information comes from.

Work carried out during the STSM

In a first step during the visit we discussed where the process of information propagation becomes evident. Here, we differentiated between two different resources: (i) news reports and (ii) Wikipedia articles (entity and event descriptions). News reports can take both the role of the original information provider and a secondary source. In any case, they are focused on current information and only refer to past happenings when putting the news into context. On the other hand, Wikipedia articles ought to be more encyclopedic. This means the coverage of ongoing happenings is supposed to contribute to the longterm description of the specific event or entity. Even though its structure and multilinguality lends itself to our

analysis, we decided not just to focus on Wikipedia as an information resource, because it does not cover the whole dynamics and diversity of news reporting and adoption.

To enable a generic analysis, we plan to synthesize the information coming from Wikipedia and news: Both sources can help to build state models representing the flow of events and its media coverage. To this end, we discussed potential data models, features and methods. Wikipedia can help to extract the sequence of important, language-independent, key terms typical to given event categories. To put them into order, we will map these terms onto a relative time dimension starting with the origin of the described happening. Given the information in the states modeled like this, we will be able to connect them to news. For example, consider a powerful earthquake: First, reports about it are propagated through the whole world, but then reports about minor aftershocks may be limited to local reports.

Our analysis aims at two different purposes: First, we plan to conduct a descriptive analysis to detect communalities between information propagation processes e.g. with respect to specific event classes. More generally, we can answer questions like how long it takes for specific information to be spread across language communities.
Second, we think of concrete applications where we will utilise the mapping of news documents to the process models. This can e.g. help to bring order into non-chronologic news reports or to group together news documents from different sources and languages into clusters representing specific sub states of an event.

Main results

To summarize, our main results are threefold: We first identified important sources useful for our temporal entity-centric information propagation in multilingual texts: multilingual news collections to address the dynamics of information propagation, and Wikipedia articles to build event models. Second, we modeled an event process models involving key terms, a relative time dimension, different event classes and news reports. Third, we discussed several application scenarios like the clustering of news documents into different sub states of events.
In future collaboration we plan to continue on this work and discussed a submission of this work to the ACM SIGIR 2017.