XLike: Cross-lingual Knowledge Extraction

Marko Grobelnik[1] and Achim Rettinger[2]

[1] Josef Stefan Institute, Slovenia
marko.grobelnik@ijs.si
[2] Karlsruhe Institute of Technology, Germany
rettinger@kit.edu

1 Administrative Facts

Call:	FP7-ICT-2011-7
Funding scheme:	Small and medium scale focused research project (STREP)
Grant agreement no:	288342
Duration:	Jan 2012 to Dec 2014 (36 months)
Partners:	Josef Stefan Institute, Karlsruhe Institute of Technology, Universitat Politecnica de Catalunya, University of Zagreb, Tsinghua University, Intelligent Software Components, Bloomberg, Slovenian Press Agency
Coordinator:	Marko Grobelnik
Website:	http://www.xlike.org

2 Introduction

The vision of the XLike project is to develop technologies to monitor and aggregate knowledge spreading across global mainstream and social media and to enable cross-lingual services for publishers, media monitoring and business intelligence. To achieve this we are combining scientific insights from several scientific areas to contribute in the area of text understanding.

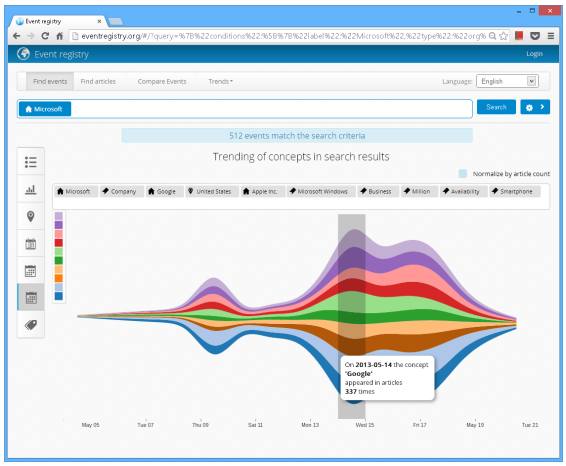
The project is aiming to solve the following two open research problems:

1. Extraction and integration of knowledge from multilingual texts with cross-lingual knowledge bases, and
2. Adapt linguistic techniques and crowdsourcing to deal with irregularities in informal language used primarily in social media.

3 Contributions to ESWC2014

We intend to present the general research ideas and use cases plus show demonstrators of our major achievements:

1. XLike deployed methods for semantic role labeling for English, Spanish, German, Catalan and Chinese. These are statistical models that recognize predicate-argument relations after syntactic analysis. These models constitute the backbone linguistic structure used in order to extract relational information that is relevant for the project use cases.
2. XLike developed a new version of annotation tool which can annotate documents with the knowledge resources such as DBpedia and other Linked Open Data sources, with added cross-lingual groundings and annotation tool that can annotate documents with these resources.
3. XLike developed cross-lingual document linking tools to cover 100 languages by improving a novel scalable approach based on hub languages.
4. XLike developed an approach for cross-lingual event extraction. In order to identify events we developed a clustering algorithm that is able to group articles based different article features. Each identified cluster of articles initially represents a separate event. Once a cluster of articles is identified we extract key information about the event – title of the event, time when it happened, location and main entities involved in it. Since each cluster contains only articles in the same language we also need to identify and merge it with clusters in other languages that are describing the same event. We achieve this using the cross-lingual document linking approach. After the processing of an event is complete, we store it in the event registry.



4 Networking at ESWC2014

We are interested in any projects related to knowledge extraction from text and multilinguality. In addition we are interested in projects related to our use cases in the publishing industry and projects that are interested in events from mainstream news.