XLike: Cross-lingual Knowledge Extraction

European Commission
FP7 Language Technologies (ICT-2011.4.2)
Small and medium scale focused research project (STREP)
288342
http://www.xlike.org

List of partners
Jožef Stefan Institute (JSI), Slovenia (coordinator)
Karlsruhe Institute of Technology (KIT), Germany
Technical University of Catalonia (UPC), Spain
University of Zagreb (UZG), Croatia
Tsinghua University (THU), China
Intelligent Software Components (ISOCO), Spain
Bloomberg (BLO), USA
Slovenian Press Agency (STA), Slovenia
New York Times (NYT), USA (associated partner)
Indian Institute of Technology (IIT), India (associated partner)

Project duration: Januray 2012 — December 2014

Summary

The goal of the XLike project is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence. The aim is to combine scientific insights from several scientific areas to contribute in the area of cross-lingual text understanding. By combining modern computational linguistics, machine translation, machine learning, text mining and semantic technologies we plan to deal with the following two key open research problems: (1) to extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases, and; (2) to adapt linguistic techniques and crowdsourcing to deal with irregularities in informal language used primarily in social media. The developed technology will be language-agnostic, while within the project we specifically address English, German, Spanish, Chinese as major world languages and Catalan, Slovenian and Croatian as minority languages. Knowledge resources from Linked Open Data cloud (e.g. Wikipedia, DBpedia, Wordnets etc.) will be used with special focus on general common sense knowledge base CycKB, that will be used as Interlingua. A number of different methods to translate from natural language to the selected formal language that serves as our Interlingua are being explored, among others also SMT. For languages where no required linguistic resources are available, we use SMT systems trained from parallel or comparable corpora (e.g. drawn from the Wikipedia) to come up with the Interlingua representation.