Cross-lingual Knowledge Extraction (XLike)

European Commission
The Seventh Framework Programme
Language Technologies (ICT-2011.4.2)
Project ID number: 288342
http://www.xlike.org

List of partners
Institut Jožef Stefan, Ljubljana, Slovenia
Karlsruher Institut für Technologie, Karlsruhe, Germany
Universitat politecnica de Catalunya, Barcelona, Spain
Sveučilište u Zagrebu, Zagreb, Croatia
Tsinghua University, Beijing, China
Intelligent software components S.A., Madrid, Spain
Slovenska tiskovna agencija d.o.o., Ljubljana, Slovenia
Bloomberg, New York, USA New York Times, New York, USA (associated partner) Indian Institute of Technology, Mumbai, India (associated partner)

Project duration: January 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 effort will combine scientific capabilities and insights from several areas of science – modern computational linguistics and NLP, machine learning, text mining and semantic technologies – in order to enable cross-lingual text "understanding" by machines. Specifically, we plan to pursue the following two key open research problems: (1) to extract and integrate formal knowledge relations from multilingual texts with cross-lingual knowledge bases, and (2) to adapt linguistic techniques and crowdsourcing to deal with irregularities in the informal language used primarily in social media. The developed technology will be language-independent to the largest possible extent, while within the project we will specifically address English, German, Spanish, Chinese and Hindi as major world languages and Catalan, Slovenian, and Croatian as less resourced languages. Knowledge resources from Linked Open Data cloud will be used, with special focus paid to using general common sense knowledge base CycKB as "semantic Interlingua". The use of Machine Translation will be oriented towards translation from a natural language as the source language into a formal language of semantic representation as source language. For languages where not enough required linguistic resources are available, we will use a probabilistic Interlingua representation trained from a parallel corpora and/or from comparable corpus derived from the Wikipedia, or use MT as a fallback option for translating the text from less resourced language to English and then process this translation. Specifically, developed solutions will be applied and evaluated in two use cases: a "Bloomberg" use case, covering the domain of financial news, and a "Slovenian Press Agency" use case, covering the domain of general news.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 451.