

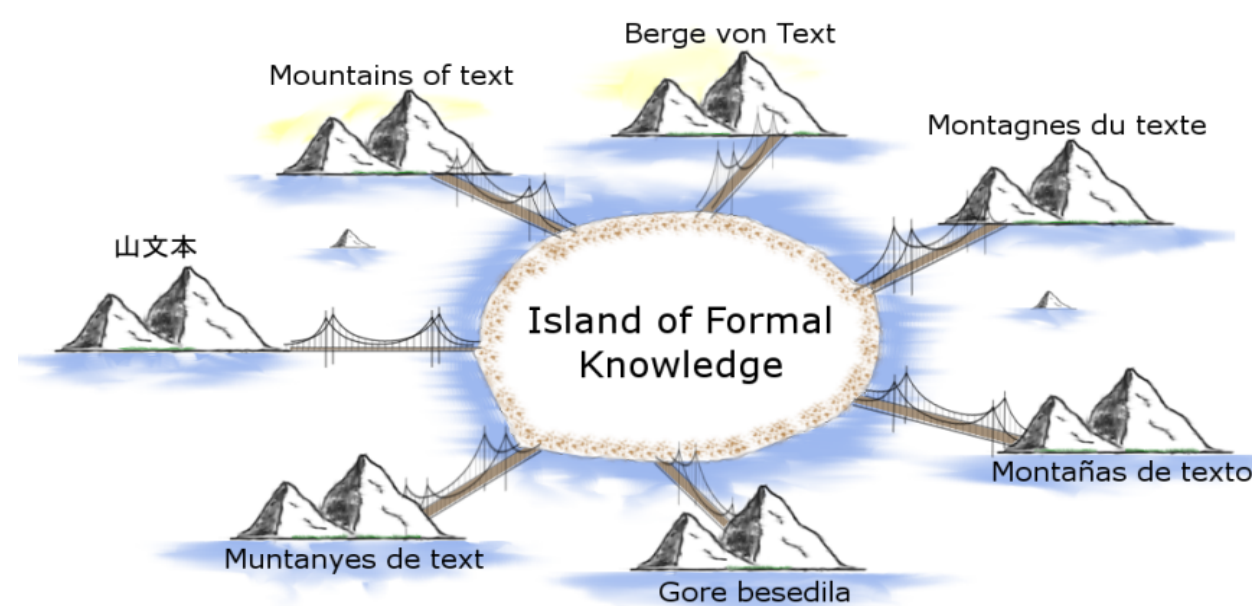
XLike: Cross-lingual Knowledge Extraction

Marko Tadić

University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Ivana Lučića 3, 10000 Zagreb, Croatia

Goal

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.



Research contributions

To extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases. To adapt linguistic techniques and crowdsourcing to deal with irregularities in the informal language used primarily in social media.

Languages covered:

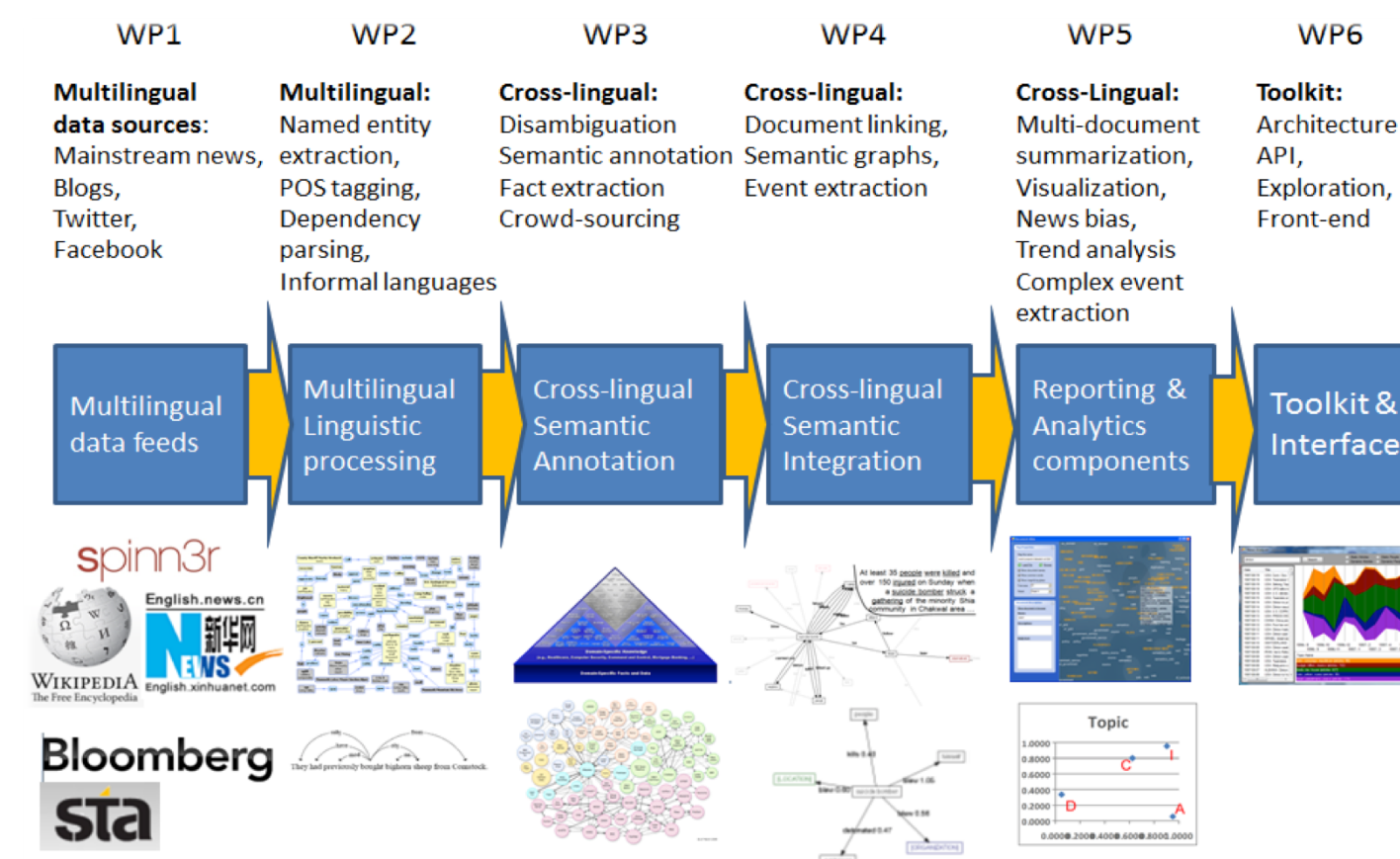
- Major languages: English, German, Spanish, Chinese and Hindi
- Minor languages: Catalan, Slovenian and Croatian

Knowledge resources used as interlingua:

- Linked Open Data (e.g. DBpedia)
- Common sense knowledge base CycKB

For languages where no required linguistic resources are available, we will use a probabilistic Interlingua representation trained from a parallel corpora or comparable corpus derived from the Wikipedia.

Project pipeline



Linguistic processing

Fully automatized pipelines for tokenization, POS-tagging, lemmatization, NERC, dependency parsing, semantic role labelling for all seven main XLike languages.

MT in XLike

Supporting technology in two cases:

- translation from natural language (SL) to semantic representation in formal language (TL);
- translation from under-resourced language(s) (SL) into English (TL) for processing with en-pipeline.

Project info

Funded under: FP7

Area: Language Technologies (ICT-2011.4.2)

Project reference: 288342

Total cost: 4.57M euro

EU contribution: 3.55M euro

Duration: from Jan 2012 to Dec 2014

Contract type: STREP

Coordinator: Marko Grobelnik

www.xlike.org

MT Summit 2013, Nice, France

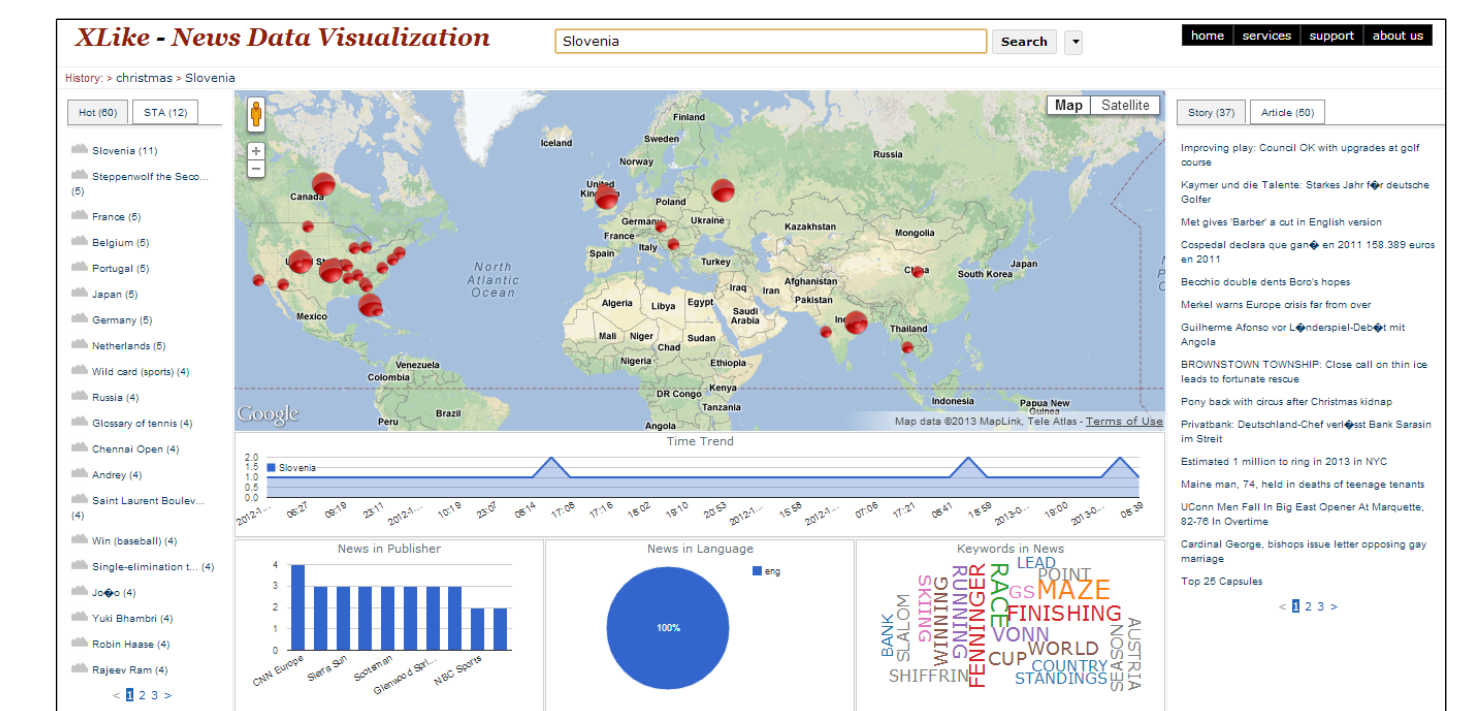
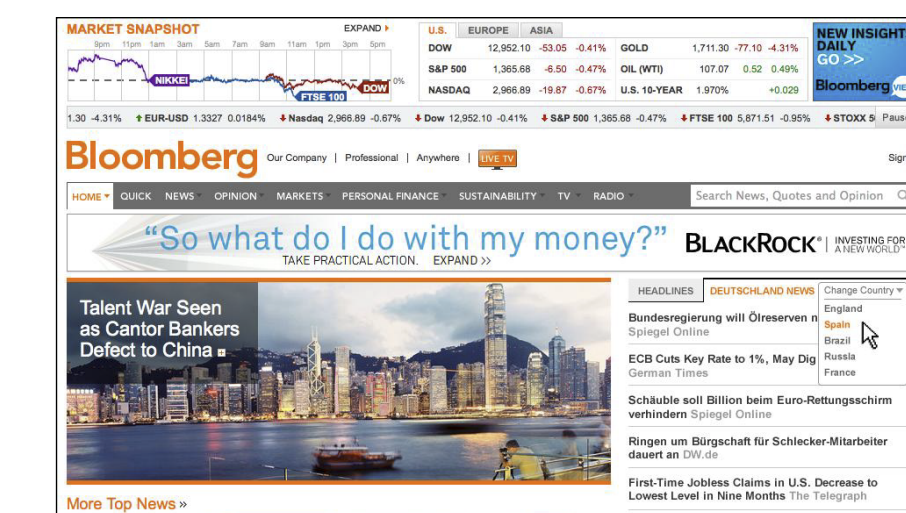
2013-09-05

Use cases

Developed technology will be used for cross-lingual summarization, contextualization, visualization, personalization and plagiarism detection of news stories.

Project use-case partners:

- Bloomberg: financial news
- Slovenian Press Agency: general news.



Project partners

Institut Jožef Stefan, Ljubljana, Slovenia

Karlsruher Institut für Technologie, Karlsruhe, Germany

Universitat politecnica de Catalunya, Barcelona, Spain

Sveučilište u Zagrebu/University of Zagreb, 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)