Cross-lingual Knowledge Extraction

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.**

To solve the above two target research problems, the consortium brings expertise and insights into the project from several research areas including language technologies, cross-lingual technologies, machine learning, text mining, information retrieval, semantic & knowledge technologies.

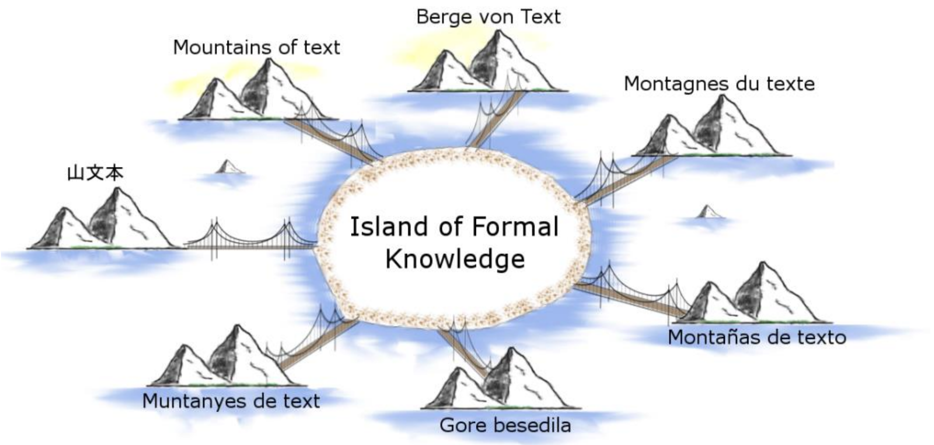
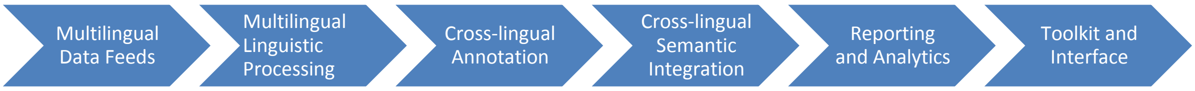


Figure 1. Schematic presentation of semantic Interlingua as an island bridged from the islands of textual information in various languages

The key tangible result of the project will be the “X-LIKE Software Toolkit”, which will serve as a basis for the use case applications. Functionally, the X-LIKE Toolkit consists from a pipeline of six separated stages:



XLike Major Achievements

Project produced several significant advancements towards the objectives:

1. Project developed elaborate data infrastructure based around **NewsFeed service: a clean, continuous, real-time aggregated stream of semantically enriched mainstream news articles from across the world**. Service was substantially extended within the XLike project by adding and improving the support and coverage of non-English content. The integrated several XLike services: (a) enrichment of the content using shallow linguistic processing and semantic annotation, and (b) creation of cross-lingual links to other related articles (cross-lingual story detection).

   The project constructed a **large comparable corpora based on Wikipedia, covering up-to 50 major languages** based on Wikipedia comparable corpus (i.e. languages with sufficient number of articles and amount of content). Cross-lingual links between the articles were used to connect articles across languages.

2. Project designed the software architecture to build **multi-lingual linguistic pipelines**, and implemented the first complete system covering all XLike languages (**English, Spanish, German, Chinese, Slovene, and Catalan**). The architecture is flexible allowing NLP tools to be written in different programming languages and integrated in the pipeline, using web services. The system provides a solid NLP foundation which is used throughout the project.

3. Project developed **scalable approach for cross-lingual document linking Canonical Correlation Analysis (CCA).** Novel techniques for scaling-up were developed using randomized approaches for dimensionality reduction, and approximation using “hub languages”. This resulted in the reduction of the optimization complexity from an NP-hard to a standard eigenvalue problem (polynomial complexity). This made it possible to **train language pair similarity models for millions of documents from Wikipedia** used as a comparable corpora.

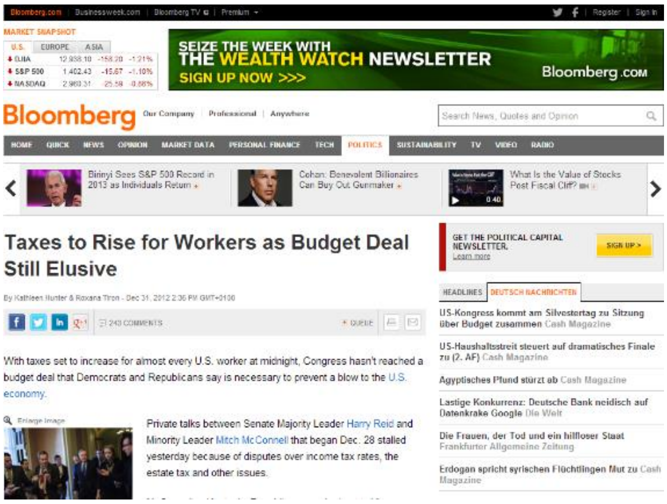
   The project developed novel techniques based on “hub languages” for dealing with language pairs with **small overlap in comparable corpora.** For “hub languages” we use some of the major languages (like English or French) with a large overlap towards the most other languages. This allowed us to substantial **improve the state-of-the-art performance on minority languages**.

4. The project performed **statistical analysis of informal language**. The first part of the analysis evaluates the performance of our syntactic processing tools on English, for which there exist annotated resources. We concluded, that a major source of errors is related to words unknown to the statistical linguistic models. Additional analysis of unknown words for different collections of informal language in English, Spanish and Catalan, concluded that, in fact, they are very frequent.

5. The project developed and analysed three methods for **cross-lingual annotation** by exploiting **cross-lingual groundings of documents in Wikipedia**: cross-lingual mapping of named entities detected by, cross-lingual extension of Wikifier, and cross-lingual document mapping based on “Explicit Semantic Analysis”.

6. The project **developed an early prototype**, integrating all the technology developed within the first year. The **prototype was successfully deployed and validated within Bloomberg and STA production environment**. The validation results provided positive and valuable feedback.

XLike Early Prototype and Use cases



Bloomberg.com | Businessweek.com | Bloomberg TV | Premium | Register | Sign In

MARKET SNAPSHOT

U.S. EUROPE ASIA

DJIA	12,938.10	-158.20	-1.21%
S&P 500	1,402.43	-15.67	-1.10%
NASDAQ	2,960.31	-25.58	-0.86%

SEIZE THE WEEK WITH THE WEALTH WATCH NEWSLETTER SIGN UP NOW >>> Bloomberg.com

Bloomberg Our Company | Professional | Anywhere

Search News, Quotes and Opinion

HOME QUICK NEWS OPINION MARKET DATA PERSONAL FINANCE TECH POLITICS SUSTAINABILITY TV VIDEO RADIO

Birinyi Sees S&P 500 Record in 2013 as Individuals Return

Cohan: Benevolent Billionaires Can Buy Out Gunmaker

What Is the Value of Stocks Post Fiscal Cliff?

Taxes to Rise for Workers as Budget Deal Still Elusive

By Kathleen Hunter & Roxana Tiron - Dec 31, 2012 2:36 PM GMT+0100

243 COMMENTS QUEUE

With taxes set to increase for almost every U.S. worker at midnight, Congress hasn't reached a budget deal that Democrats and Republicans say is necessary to prevent a blow to the U.S. economy.

Enlarge image

Private talks between Senate Majority Leader Harry Reid and Minority Leader Mitch McConnell that began Dec. 28 stalled yesterday because of disputes over income tax rates, the estate tax and other issues.

GET THE POLITICAL CAPITAL NEWSLETTER. Learn more SIGN UP >

HEADLINES DEUTSCH NACHRICHTEN

US-Kongress kommt am Silvestertag zu Sitzung über Budget zusammen Cash Magazine

US-Haushaltsstreit steuert auf dramatisches Finale zu (2. AF) Cash Magazine

Agyptisches Pfund stürzt ab Cash Magazine

Lastige Konkurrenz: Deutsche Bank neidisch auf Datenkrake Google Die Welt

Die Frauen, der Tod und ein hilfloser Staat Frankfurter Allgemeine Zeitung

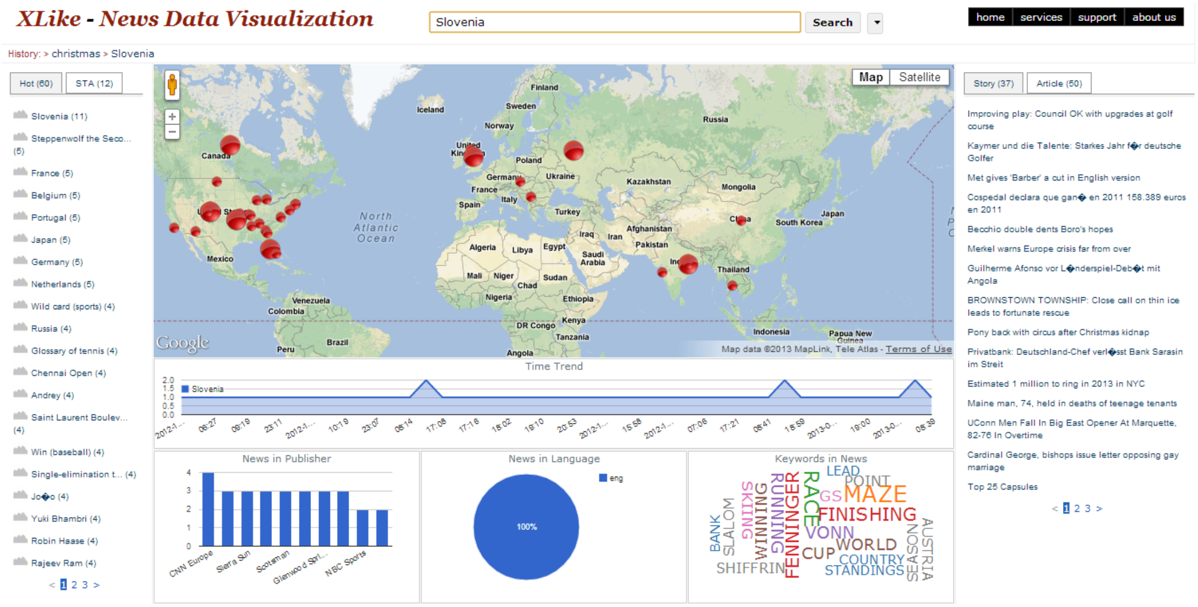
Erdogan spricht syrischen Flüchtlingen Mut zu Cash Magazine

Financial news prototype was developed and tested in the production setting of the **Bloomberg.com** online portal. The prototype focused on the insights discovered by the Bloomberg customer analytics group, that the users prefer to read financial news in their local language. As such, the prototype was used to recommend a content from local financial news sources to visitors of Bloomberg.com.

The prototype relied on several XLike technologies: (a) the source of local news was collected from NewsFeed service, and (b) a model for identifying relevant articles was developed using "Bloombergness" measure. The model was first built for English and transferred to other XLike languages using Canonical Correlation Analysis. The prototype was deployed to the Bloomberg.com portal, and is being exposed to a sample of all users. The initial results show increase of user engagement when XLike module is present on the website.

General news prototype was based on the visualization prototype, and was used by **STA** for the task of for tracking mentions of Slovenian entities in foreign news sources regardless of the language and for tracking top stories in a selected area and time period. The first task is an important assignment for national press agencies, especially in small countries such as Slovenia, while the second task is important for all news desks, especially the ones which operate on a daily basis. Both tasks are time consuming, but important and therefore solutions which would speed-up and at the same time improve the process are highly desired.

Usability of the tool was evaluated through comparison of the results of work with XLike tool compared to the existing system at STA. The comparison was made for editorial and journalistic activities, most relevant in regard to the functionalities of the XLike prototype. The first results are promising, especially in the case of searching mentions and top stories in the global feed, resulting in less time spent and improved coverage.



XLike - News Data Visualization

Slovenia Search

home services support about us

History: > christmas > Slovenia

Hot (60) STA (12)

Slovenia (11)
Steppenwolf the Seco... (5)
France (5)
Belgium (5)
Portugal (5)
Japan (5)
Germany (5)
Netherlands (5)
Wild card (sports) (4)
Russia (4)
Glossary of tennis (4)
Chennai Open (4)
Andrey (4)
Saint Laurent Boulev... (4)
Win (baseball) (4)
Single-elimination t... (4)
Jo�o (4)
Yuki Bhambri (4)
Robin Haase (4)
Rajeev Ram (4)

< 1 2 3 >

Story (37) Article (50)

Improving play: Council OK with upgrades at golf course
Kaymer und die Talente: Starkes Jahr f�r deutsche Golfer
Met gives 'Barber' a cut in English version
Cospedal declara que gan� en 2011 158.389 euros en 2011
Becchio double dents Boro's hopes
Merkel warns Europe crisis far from over
Guilherme Afonso vor L�nderspiel-Deb�t mit Angola
BROWNSTOWN TOWNSHIP: Close call on thin ice leads to fortunate rescue
Pony back with circus after Christmas kidnap
Privatbank: Deutschland-Chef verl�sst Bank Sarasin im Streit
Estimated 1 million to ring in 2013 in NYC
Maine man, 74, held in deaths of teenage tenants
UConn Men Fall In Big East Opener At Marquette, 82-76 In Overtime
Cardinal George, bishops issue letter opposing gay marriage
Top 25 Capsules

< 1 2 3 >