xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D1.3

Report on Existing Data

Editor:	Blaž Novak, JSI
Author(s):	Blaž Novak, JSI; Achim Rettinger, KIT; Dubravko Ćulibrk, UNITN; Henrik Holzhausen, VICO; Jan Rupnik, JSI; Philipp Sorg, ECONDA
Deliverable Nature:	Report (R)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M3 – 31 January 2014
Actual Delivery Date:	M3 – 31 January 2014
Suggested Readers:	All project partners
Version:	1.0
Keywords:	existing data, existing language processing infrastructure

Disclaimer

This document contains material, which is the copyright of certain xLiMe consortium parties, and may not be reproduced or copied without permission.

All xLiMe consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the xLiMe consortium as a whole, nor a certain party of the xLiMe consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	xLiMe – crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP1 Processing Multilingual Multimedia Data
Document Title:	D1.3 – Report on Existing Data
Editor:	Blaž Novak, JSI
Work Package Leader:	Andreas Thalhammer, KIT

Copyright notice

Executive Summary

This document provides a list and a description of datasets and data processing technologies available to xLiMe partners from the start of the project.

Data sources for use cases will be provided by ZATTOO, VICO, ECONDA and JSI. All of them are already prepared and can be used with no further delays.

We have also provided a list of potentially useful public datasets and knowledge resources to be used during research, for training and evaluation of developed software.

A description of existing language technologies that will be used in extraction and data pre-processing tasks is included as a reference.

Table of Contents

Abbreviations

API	Application Programming Interface
CSV	Comma Separated Values
FTP	File Transfer Protocol
HLS	HTTP Live Streaming
HTTP	Hyper-Text Transfer Protocol
IPTV	Internet Protocol TV
JSON	Javascript Object Notation
MPQA	Multi-Perspective Question Answering
MSD	Morphosyntactic Descriptions
POS	Part-of-Speech
SaaS	Software as a Service
SIFT	Scale Invariant Feature Transform
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
XML	Extensible Markup Language

1 Introduction

This document is the result of the task T1.3 and serves as input to work packages WP2, WP3 and WP4.

It contains a reference list of all of the datasets and data sources available within the project which will be provided by partners for internal use, along with a description of what those datasets contain, and how to access them. All of the partner provided data, with the exception of JSI NewsFeed, is proprietary and should not be redistributed without explicit permission of the owner. This data will be mainly used as input to the extraction pipeline produced in WP2.

As a further contribution, we provide a list of publically available datasets that might be relevant for research purposes, such as training and verification of machine learning algorithms, etc. We also provide a list and a description of commonly used knowledge resource databases, which might be used as a grounding vocabulary in the annotation phases of WP3. This data, however, was not generated by xLiMe project partners.

The final section contains a description of existing software technologies that will be used in xLiMe use cases, most importantly the speech recognition web service, provided by technology partner Vecsys, and some of the technology created within the XLime FP7 project, which will be reused and possibly extended here.

2 Existing Data Sources and Datasets

2.1 Primary Datasets

This section describes all of the datasets that will be used as inputs to use cases.

2.1.1 ZATTOO TV Streams

Zattoo Europa AG is a European IPTV provider. It was founded in 2005 and is a market leader in internet based TV streaming in Germany and Switzerland, with additional presence in France, Spain, UK, Luxemburg and Denmark.

The core data source provided to the xLiMe project is a set of up to 250 live TV channels, streamed over the internet. Corresponding audio streams cover 14 languages. The majority of audio is in German language but there is also a good representation of English, French, Italian, Turkish, Portuguese, Spanish, Croatian and Polish languages in the available set of channels.

All of the streams contain audio and video sub-streams. Subtitles for channels which provide them will become available at a later date.

Project partners have been provided with user accounts to access the available video channels using a web based interface. However, for accessing a video stream from an application, an API must be used. Authentication for the API requires an additional key (referred to as the *app_tid* in the documentation*)*, which is available from the project coordinator.

Full documentation for the API is available at https://developer.zattoo.com/. Login credentials for the documentation website are separate from the web stream access and the API key and were also provided at the kick-off meeting.

The API itself uses HTTP as the underlying transport protocol for session management and data streaming. In order to start the video transfer, a sequence of POST requests needs to be sent to the server, in which the API key is provided and the channel selected. Video stream is delivered using the HLS protocol as a response to a 'watch' request.

As a part of a WP2 'speech to text extraction' task, JSI will access selected TV channels on behalf of other partners and convert their audio tracks to a textual representation. This data will be available through an authenticated web service located at http://xlime.ijs.si/ in either XML or JSON format. It will include timestamps, which will enable other partners to align the text with video streams received directly. The same web service will also provide a configurable periodic stream of timestamped video snapshots and audio segments as an alternative supply of image data for the WP2 'text from video extraction' task.

2.1.2 VICO Social Media Streams

VICO Research & Development GmbH is a German company focused on social media measurement and analysis, development of social media monitoring systems and social media consulting.

As a part of their social media monitoring activities, they are retrieving various types of content from the internet in near real-time, as it is being created. This data will be made available to project partners. Currently, they are fetching about 10 million documents from the internet per day, and that number is continuously growing.

The data is divided between approximately 15 languages, with the best coverage of German, English, Italian, French, Spanish, Polish and Chinese.

The following table is a description of available data streams:

Table 1 VICO Social Media Data Streams

Dataset name	Details
Forums	Extracted web forum posts, information about the thread for each post, link to folder or sub-forum for each thread.
Online News	Mainly the posts themselves, in a few cases also comments.
Blogs	Mainly the posts themselves, in a few cases also comments.
Facebook	Public fan pages, based on keyword search.
Twitter	Based on twitter search API, using predefined keywords.
YouTube	Video links, descriptions and comments. Based on keyword search.
Webpages	A set of explicitly defined monitored webpages.
Rating sites	Textual posts; only a couple of sites.
Q&A Sites	Textual posts; only a couple of sites.

This data is gathered in three different ways: either by crawling and parsing both done by VICO, using public APIs or receiving data from external data providers.

As a post-processing step, sentiment analysis is performed on the text.

A somewhat uniform schema is used for data representation: at least content text, title, author(s), date of publication and source type are available for the various content types.

The data can be passed around in either JSON or XML format. The exact protocol for access will need to be determined based on task requirements.

2.1.3 Econda Web Shop Product Information

econda GmbH is a web analytics and recommendation company. It provides services to over 1000 e-business customers. Within the xLiMe project, ECONDA will be providing data about the Deichmann web shop product range, that will be used to develop new reporting functionality and web shop features.

The dataset is a snapshot of the internal product list from Deichmann and of a web crawl of the Deichmann web shop. It is available as a set of the following files in CSV format:

Table 2 Deichmann Web Shop Product Description Dataset

Name	Filename	Description
Products	products.csv	Contains a list of (id, name, brand, link, image URL, material, categories, target groups) lines describing a specific product.
Categories	categories.csv	Contains a list of (id, name, parent_id) lines describing product categories.
Product-category mapping	econda_products2category.csv	Contains matching pairs of product and category ids.
Category hierarchy	econda_categorytree.csv	Contains pairs of parent and child category id pairs, describing product category hierarchy.

Along with the listed CSV files, a set of directories is available archived in a ‘products.zip’ file, with one directory per product. These contain images of the product and a text file with a detailed semi-structured textual description. Datasets are available for German, Spanish and UK stores, with descriptions in their respective languages.

The dataset currently describes approximately 2360 products, and will be manually updated when needed. It is located on the private KIT cloud infrastructure.

2.1.4 JSI NewsFeed

JSI NewsFeed is a news and blog collection service developed at Jožef Stefan Institute during the course of multiple EU FP7 projects.

Main focus of the NewsFeed system is the collection of mainstream news articles with emphasis on good coverage of minority languages and dialects while providing a configurable low latency of discovery for new content; however it also collects a sizeable portion of most popular blog posts.

The system provides approximately 250.000 articles per day, with a good coverage of 50 languages and occasional content in approximately 100 more.

After retrieval, the articles are cleaned and analysed for language and content, clustered into events, and semantically enriched by POS processing, named entity extraction and merging, anaphora resolution, classification into a topic taxonomy, etc. This is only done when the article is written in one of the supported languages. English and Slovenian language are supported internally, while Spanish and German support is being developed as a part of the XLike project, as described in the last section of this document.

Access to the raw data stream in XML format is free for research purposes and can be obtained at http://newsfeed.ijs.si/. Full-text search of the archives is also available using an API.

2.2 Public Datasets

Over the past few decades, there were significant advances made in the understanding of the information generated from various media formats – either text, audio or video. But in recent years, the focus has shifted towards simultaneously processing two or more of these formats at the same time in order to achieve better performance than what is possible by using just a single modality. A significant number of multilingual and multimedia datasets has been created in the process. Table 3 lists some of the currently available public datasets.

Many of these datasets were produced with the purpose of advancing the state of the art by organizing competitions between researchers, by providing them with standardized inputs and a known ground truth.

Table 3 Multi-lingual and Multi-modal Dataset Availability

Number	Dataset	Availability	Text	Video	Audio	Image	Multi-Lingual (**)	Cross-Lingual
1	Affective Text (SemEval)	Free	**Yes**	No	No	No	No	No
2	Hollywood Human Actions	Free	No	**Yes**	No	No	No	No
3	IAPRTC-12	Free	No	No	No	**Yes**	**Yes**	No
4	Image Word Relatedness	Free	**Yes**	No	No	**Yes**	No	No
5	Subjectivity Text v2.0	Free	**Yes**	No	No	No	**Yes**	**Yes**
6	MultiLingual Parallel Video Corpus Descriptions	Free	**Yes**	**Yes**	No	No	**Yes**	**Yes**
7	EUROM1- MultiLingual Speech Corpus	Paid	No	No	**Yes**	No	**Yes**	**Yes**
8	TRECVID Multimedia Event Detection	Registered	No	**Yes**	**Yes**	No	No	No
9	ECI MultiLingual Text	Paid	**Yes**	No	No	No	**Yes**	**Yes**
10	ImageCLEF Wikipedia Image Retrieval Datasets	Registered	**Yes**	No	**Yes**	No	No	No
11	CoPhIR	Registered	**Yes**	No	No	**Yes**	No	No

12	BelgaLogos	Free	**Yes**	No	No	**Yes**	No	No
13	Street View Text	Free	**Yes**	No	No	**Yes**	No	No
14	TRECVID Internet Archive data (IACC.2)	Registered	**Yes**	**Yes**	**Yes**	No	**Yes**	No

****** Multi-Lingual -- Multiple Languages Present

Table 4 shows the extra information for each of the above mentioned datasets.

Table 4 Multi-lingual and Multi-modal Dataset Extended Information

Number	**Dataset**	**Data Access**	**Data Coverage**	**Language**	**Link**
1	Affective Text (SemEval)	Files	250 training sentences, 1000 testing sentences	English	http://www.cse.unt.edu/~rada/affectivetext/
2	HollywoodHuman Actions	Files	475 videos, 3 annotated documents	English	http://www.di.ens.fr/~laptev/actions/
3	IAPRTC-12	Files	20,000 images and image annotations	English, German and Random	http://www.imageclef.org/photodata
4	Image Word Relatedness	Files	167 images and 1000 synsets in extended version	English	http://lit.csci.unt.edu/index.php/index.php?P=research/downloads#MEASURING_SEM_REL
5	Subjectivity Text v2.0	Files	SemCor and MPQA	English, Romanian, Arabic,	http://lit.csci.unt.edu/index.php/index.p

			sentences	German, French and Spanish	hp?P=research/downloads#MULTI_SUB_A
6	MultiLingual Parallel Video Corpus Descriptions	Files and Microsoft .msi files	122,000 descriptions of 2089 videos	16 languages	http://www.cs.utexas.edu/users/ml/clamp/videoDescription/#data
7	EUROM1- MultiLingual Speech Corpus	Files	Many Talker Corpus: 100 numbers, 3 passages, 5 sentences, speech signal Few Talker Corpus: 5x 100 numbers, 15 passages, 25 sentences, Very Few Talker Corpus C(C)VC(V): material embedded in 5 contexts, various phrases, 5x context words	Danish, Dutch, English, French, German, Norwegian, Swedish	http://www.phon.ucl.ac.uk/shop/eurom1.php
8	TRECVID Multimedia Event Detection	Files	4000 hours of video	English	http://www.nist.gov/itl/iad/mig/med12.cfm
9	ECI MultiLingual Text	Files	Not Available	Albanian, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, French, Gaelic, German, Italian, Japanese, Latin, Lithuanian, Mandarin Chinese, Modern Greek, Northern Uzbek, Norwegian, Norwegian Bokmaal,	http://catalog.ldc.upenn.edu/LDC94T5

				Norwegian Nynorsk, Portuguese, Portuguese, Russian, Serbian, Slovenian, Spanish, Standard Malay, Swedish, Turkish	
10	ImageCLEF Wikipedia Image Retrieval Datasets	Files	237,434 images in 2011 tasks and 50 topics	English	http://www.imageclef.org/wikidata
11	CoPhIR	Files	106 million processed images	English	http://cophir.isti.cnr.it/get.html
12	BelgaLogos	Files	10000 images	English	http://www-sop.inria.fr/members/Alexis.Joly/BelgaLogos/BelgaLogos.html#download
13	Street View Text	Files	350 images, 725 labels	Various	http://vision.ucsd.edu/~kai/svt/
14	TRECVID Internet Archive data (IACC.2)	Files	600 hours of video in 7300 files	Various	http://trecvid.nist.gov/trecvid.data.html#tv13

2.3 Knowledge Resources

2.3.1 DBpedia

DBpedia is the largest cross-lingual knowledge database publically available. It is created by automatically extracting structured information available in the info-boxes in Wikipedia dump files. The current version, 3.9, is based on March 2013 dump of Wikipedia database and contains information about over 3 million useful entities, such as people, places, organizations, etc. Facts about these entities are represented as subject-predicate-object triples and are serialized in RDF format. There are approximately 2.5 billion triples extracted from 119 different language versions of Wikipedia, connecting the entities together.

Entities in the DBpedia are described using unique URIs, which cover all of the concepts present in Wikipedia. This makes the set of DBpedia entities a very useful common vocabulary, with which we can annotate information extracted in work package 2. Additionally, the semantically rich entity interconnectedness can be used for advanced inference that is not possible using shallow methods.

DBpedia can be downloaded in compressed form from http://wiki.dbpedia.org/Downloads39. Access to the already indexed DBpedia is also available from KIT, which can provide a SPARQL endpoint to project partners.

2.3.2 Wikidata

Wikidata is a public knowledge base related to Wikipedia, containing information about approximately 14 million items. Unlike DBpedia, which is extracted from existing textual representation of information, Wikidata is organized as a structured document store, containing human edited key-value pairs, describing specific items, with emphasis on provenance and quality.

The Wikidata project has three main goals. The first one was to create a unified semantic representation of concepts (pages) appearing on different language editions on Wikipedia, providing consistent cross-lingual links and consequently allowing text written in different languages to be transformed into a unified semantic representation consisting of Wikipedia URIs. The second one is to provide a unified semantic structure of infoboxes across languages, unifying multilingual input at a higher, more semantic level. The third goal, less relevant to xLiMe, is focused on Wikipedia itself and relates to integrating information in the infoboxes and the free text in the body of Wikipedia articles.

Access to Wikidata is available through KIT.

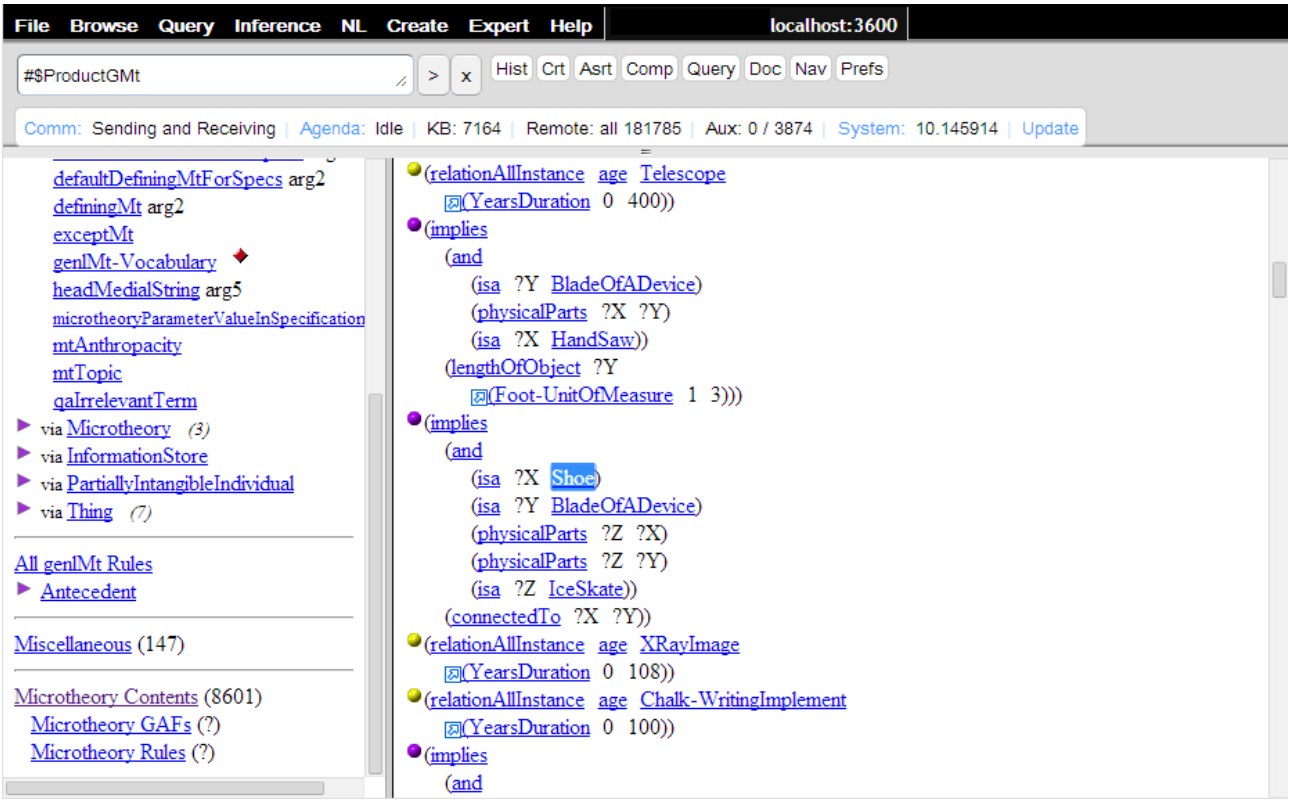
2.3.3 ResearchCyc Knowledge Base

ResearchCyc is an inference engine and a knowledge base that contains systematically manually encoded human knowledge. It has been in development for the past 30 years and has so far received over 1000 person-years of development time.

The database is an ontology of over 500,000 concepts, of which 17,000 are predicates ranging from abstract ones such as “is-a” to specific ones like “restaurant-has-menu-item”. Concepts are connected together by approximately 10 million assertions, describing various aspects of the real world.

If access to ResearchCyc is needed, a virtual machine containing the inference engine service will be provided by the JSI. A client-side Java library exposes an API that lets the user query the database and use the inference engine, while taking care of the communication with the server.

A web interface to the inference engine is also available, as shown in Figure 1.



#$ProductGMt

Comm: Sending and Receiving | Agenda: Idle | KB: 7164 | Remote: all 181785 | Aux: 0 / 3874 | System: 10.145914 | Update

defaultDefiningMtForSpecs arg2
definingMt arg2
exceptMt
genlMt-Vocabulary
headMedialString arg5
microtheoryParameterValueInSpecification
mtAnthropacity
mtTopic
qaIrrelevantTerm
via Microtheory (3)
via InformationStore
via PartiallyIntangibleIndividual
via Thing (7)

All genlMt Rules
Antecedent

Miscellaneous (147)

Microtheory Contents (8601)
Microtheory GAFs (?)
Microtheory Rules (?)

```
(relationAllInstance age Telescope
   (YearsDuration 0 400))
(implies
   (and
      (isa ?Y BladeOfADevice)
      (physicalParts ?X ?Y)
      (isa ?X HandSaw))
   (lengthOfObject ?Y
      (Foot-UnitOfMeasure 1 3)))
(implies
   (and
      (isa ?X Shoe)
      (isa ?Y BladeOfADevice)
      (physicalParts ?Z ?X)
      (physicalParts ?Z ?Y)
      (isa ?Z IceSkate))
   (connectedTo ?X ?Y))
(relationAllInstance age XRayImage
   (YearsDuration 0 108))
(relationAllInstance age Chalk-WritingImplement
   (YearsDuration 0 100))
(implies
   (and
```

Figure 1 Web Interface to the CycKB Inference Engine

3 Existing Language Processing Technology

3.1 Vecsys MediaSpeech

Vecsys is a French company that provides a speech to text software solution called MediaSpeech. MediaSpeech is used by media monitoring companies and on-line press groups in France, UK, Spain, USA and Gulf countries. It will also be used as the main text from speech extraction engine on the xLiMe project.

MediaSpeech is a SaaS (software-as-a-service) cloud service, however it will be possible to get virtual machine image files to run locally, should processing requirements needed by the project exceed the available cloud resources. In that case, one or multiple computers at JSI will be dedicated to running it.

Currently, the service supports transcription of audio in English, Spanish, French, Italian, Russian, Arabic and Standard Chinese, with support for German language planned in coming months. This covers all of the major xLiMe languages; support for Catalan, however, will have to be developed separately.

Interface to the MediaSpeech consists of a combination of SOAP HTTP requests and FTP file transfers.

Processing is done by converting chunks of audio at a time. The longer the chunks are, the higher the quality of the conversion, so it is recommended to use at least 2-5 minute long segments.

The conversion process begins by calling a *login* function with the username and password provided by Vecsys.

After a successful login, an audio file is uploaded using the FTP protocol to the MediaSpeech server. Supported audio file formats are ‘wav’, ‘mp3’, ‘ogg’, ‘flv’, ‘bwf’, ‘mpg’, ‘wma’, ‘m4v’, ‘mp4’, ‘m4a’, ‘amr’, ‘mp2’, ‘aif’, ‘caf’, ‘alaw’ and ‘al’.

When the upload is complete, conversion can be started by issuing a *transcribe* SOAP request with the name of the file, language and quality settings. Completion of conversion is signalled through *status* function polling, and the transcribed text is finally retrieved by calling the *result* function with the job ID and the desired output format, which can be either XML, plain text, or a specialized subtitle file format.

Further documentation for the SOAP interface can be found at https://mediaspeech.com/webservice/index.php.

3.2 Multi-lingual Linguistic Technologies developed by the XLike Project

We will reuse and improve the cross-lingual information extraction technologies and text processing pipeline developed within the XLike project, which will be used to process the textual part of multi-media input on the use cases.

3.2.1 Linguistic Enrichment Pipeline

For English, Spanish, Catalan, German, Croatian, Slovenian and Chinese languages, the XLike project has developed an extensive pipeline for text processing and semantic enrichment. This covers all of the main xLiMe languages.

For each language, an independent pipeline has been constructed from a set of web services provided by participants. The pipeline performs language identification as well as shallow and deep semantic analysis.

Shallow analysis consists of sentence tokenization, part-of-speech and morphosyntactic tagging, lemmatization and named entity recognition and classification. Named entities and words from this stage are linked to DBpedia and can already be used for annotation and semantic inference.

The deep linguistic processing stage performs syntactic parsing, where the syntactic structure of sentences is determined, semantic role labelling, where arguments of lexical predicates in a sentence are identified and labelled with semantic roles, and frame extraction, which produces a set of semantic frames – units of meaning with associated actors and actions.

All of these services are implemented as RESTful web services, which allows for easy reuse within xLiMe project. We will try to reuse as many of these components for text extracted from audio and social media streams as well.

3.2.2 Statistical Cross-Lingual Document Linking

3.2.2.1 Motivation

We now describe the resources developed in XLike that enable cross-lingual text analysis. The main motivation for the technology that is described here is that it enables applying standard machine learning techniques on cross-lingual text mining problems such as cross-lingual classification, clustering and document retrieval. These methods are based on finding language independent representations of documents. Some of the problems that are being addressed by the XLike project are: demand prediction (given the reading habits of German readers extracted from social media data, predict which articles published on the Bloomberg website are most relevant to them), cross-lingual content tracking (e.g., a press agency would like to know who translates their content).

3.2.2.2 Technologies: Canonical Correlation Analysis (CCA) and Explicit Semantic Analysis (ESA)

CCA and ESA are both techniques that map cross-lingual documents into a common vector space, where they can be compared (i.e. a similarity computation) and analysed. The representations are based on expressing documents in terms of their similarity to a given set of topic vectors (distributions over words). Given an aligned data set of pairs of documents in two languages (for example, the pairs may represent documents that are translations of each other), CCA extracts those topics that are highly correlated across languages, and ESA uses the aligned set as resultant topics without any additional computation. CCA models are more compact, but require more computation to extract.

3.2.2.3 Software

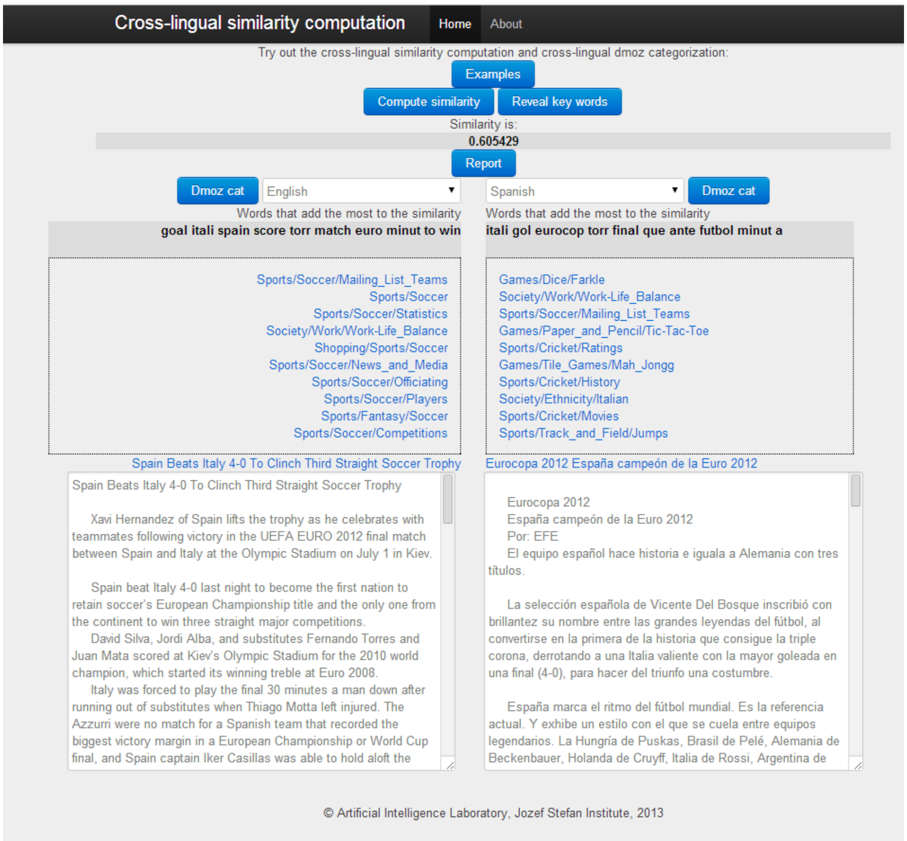
JSI will provide a C++ library (named '*glib*') that contains all of the functionality for building cross-lingual models and for solving machine learning tasks on the data. The library also contains functionality for cross-modal analysis. Incorporating a new modality to the analysis involves implementing a similarity function (called a *kernel*) for the data. For example, to include images, one can use a bag-of-visual-words representation based on SIFT features and cosine similarity as the kernel function.

3.2.2.4 Services

Based on the dataset extracted from Wikipedia, XLike project built a set of web services that include computing similarities between documents and document classification in the taxonomy of the Open Directory Project (ODP). A demo based on these service calls is available at: http://pankretas.ijs.si:1221/wikipedia.html, and a screenshot is provided with Figure 2.

3.2.2.5 Extensions

CCA was designed to analyse two data sources and looks for linear patterns in the data. It can be extended to more than two sources, and an implementation is already available that enables joint analysis of documents written in 100 languages. The method can be extended to finding nonlinear patterns by using kernel methods, which makes it useful for cross-modal analysis. For example, having an aligned data set of images and their textual descriptions, the method can identify a common representation that enables image retrieval based on textual query or vice versa.



Try out the cross-lingual similarity computation and cross-lingual dmoz categorization:

Similarity is:
0.605429

English

Spanish

Words that add the most to the similarity
goal itali spain score torr match euro minut to win

Words that add the most to the similarity
itali gol eurocop torr final que ante futbol minut a

Sports/Soccer/Mailing_List_Teams
Sports/Soccer
Sports/Soccer/Statistics
Society/Work/Work-Life_Balance
Shopping/Sports/Soccer
Sports/Soccer/News_and_Media
Sports/Soccer/Officiating
Sports/Soccer/Players
Sports/Fantasy/Soccer
Sports/Soccer/Competitions

Games/Dice/Farkle
Society/Work/Work-Life_Balance
Sports/Soccer/Mailing_List_Teams
Games/Paper_and_Pencil/Tic-Tac-Toe
Sports/Cricket/Ratings
Games/Tile_Games/Mah_Jongg
Sports/Cricket/History
Society/Ethnicity/Italian
Sports/Cricket/Movies
Sports/Track_and_Field/Jumps

Spain Beats Italy 4-0 To Clinch Third Straight Soccer Trophy

Spain Beats Italy 4-0 To Clinch Third Straight Soccer Trophy

Xavi Hernandez of Spain lifts the trophy as he celebrates with teammates following victory in the UEFA EURO 2012 final match between Spain and Italy at the Olympic Stadium on July 1 in Kiev.

Spain beat Italy 4-0 last night to become the first nation to retain soccer's European Championship title and the only one from the continent to win three straight major competitions.

David Silva, Jordi Alba, and substitutes Fernando Torres and Juan Mata scored at Kiev's Olympic Stadium for the 2010 world champion, which started its winning treble at Euro 2008.

Italy was forced to play the final 30 minutes a man down after running out of substitutes when Thiago Motta left injured. The Azzurri were no match for a Spanish team that recorded the biggest victory margin in a European Championship or World Cup final, and Spain captain Iker Casillas was able to hold aloft the

Eurocopa 2012 España campeón de la Euro 2012

Eurocopa 2012
España campeón de la Euro 2012
Por: EFE
El equipo español hace historia e iguala a Alemania con tres títulos.

La selección española de Vicente Del Bosque inscribió con brillantez su nombre entre las grandes leyendas del fútbol, al convertirse en la primera de la historia que consigue la triple corona, derrotando a una Italia valiente con la mayor goleada en una final (4-0), para hacer del triunfo una costumbre.

España marca el ritmo del fútbol mundial. Es la referencia actual. Y exhibe un estilo con el que se cuela entre equipos legendarios. La Hungría de Puskas, Brasil de Pelé, Alemania de Beckenbauer, Holanda de Cruyff, Italia de Rossi, Argentina de

© Artificial Intelligence Laboratory, Jozef Stefan Institute, 2013

Figure 2 Web Demo for the CCA Statistical Cross-lingual Document Similarity Calculation Service