Global Media Monitoring

http://eventregistry.org/

Marko Grobelnik
Jozef Stefan Institute
Ljubljana, Slovenia

Contributions from Gregor Leban, Blaz Fortuna, Janez Brank, Jan Rupnik, Andrej Muhic
Outline

• Introduction
• Collecting Media Data
• Document Enrichment
• Cross-linguality
• News Reporting Bias
• Event Representation
• Event Tracking
• Future Projects
Introduction
What questions we’ll try to answer?

• Where to get global media data?
• What is extractable from media documents?
• How to connect information across languages?
• What is an event?
• How to approach diversity in news reporting?
• How to visualize global event dynamics?
Systems/Demos used within the presentation

- NewsFeed (http://newsfeed.ijs.si/)
  - News and social media crawler

- Enrycher (http://enrycher.ijs.si/)
  - Language and Semantic annotation

- XLing (http://xling.ijs.si/)
  - Cross-lingual document linking and categorization

- DiversiNews (http://aidemo.ijs.si/diversinews/)
  - News Diversity Explorer

- Event Registry (http://eventregistry.org/)
  - Event detection and topic tracking
The overall goal

• The goal is to establish a real-time system
  • ...to collect data from global media in real-time
  • ...to identify events and track evolving topics
  • ...to assign stable identifiers to events
  • ...to identify events across languages
  • ...to detect diversity of reporting along several dimensions
  • ...to provide rich exploratory visualizations
  • ...to provide interoperable data export
Global Media Monitoring pipeline

Input data

Mainstream news

Blogs

Pre-processing steps

Article semantic annotation
Extraction of date references
Cross-lingual article matching
Detection of article duplicates

Event construction

Article clustering
Cross-lingual cluster matching
Event formation
Event info. extraction
Identifying related events

Event storage & maintenance

Event registry

API Interface

GUI/Visualizations

http://EventRegistry.org
Collecting Media Data

http://newsfeed.ijs.si/
Where to get references to news publishers?

• Good start is Wikipedia list of newspapers:
From a newspaper home-page to an article

http://www.nytimes.com/

HTML

RSS Feed (list of articles)

Article to be retrieved
Collecting global media data

• Data collection service News-Feed
  • [http://newsfeed.ijs.si/](http://newsfeed.ijs.si/)
  • ...crawling global main-stream and social media

• Monitoring
  • ~60k main-stream publishers (RSS feeds+special feeds)
  • ~250k most influential blogs (RSS feeds)
  • free Twitter feed

• Data volume: ~350k articles & blogs per day (+5M tweets)
• Languages: eng (50%), ger (10%), spa (8%), fra (5%)
Downloading the news stream (1/2)

• The stream is accessible at http://newsfeed.ijs.si/stream/

• To download the whole stream continuously, you can use the python script (http://newsfeed.ijs.si/http2fs.py)

• The script does the following:

```python
timestamp = [when you want to start downloading, e.g. now() minus 1 hour]
while True:
    fetch http://newsfeed.ijs.si/stream/?after=timestamp
    if [404 error]:
        # there is no new data
        pause 1 minute
    else:
        save data
        timestamp = [extract it from the Content-Disposition HTTP header]
```
News Stream Contents and Format

- The root element, `<article-set>`, contains zero or more articles in the following XML format:

...more details:

Document Enrichment

http://enrycher.ijs.si/
What can extracted from a document?

• **Lexical level**
  • *Tokenization* – extracting tokens from a document (words, separators, ...)
  • *Sentence splitting* – set of sentences to be further processed

• **Linguistic level**
  • *Part-of-Speech* – assigning word types (nouns, verbs, adjectives, ...)
  • *Deep Parsing* – constructing parse trees from sentences
  • *Triple extraction* – subject-predicate-object triple extraction
  • *Name entity extraction* – identifying names of people, places, organizations

• **Semantic level**
  • *Co-reference resolution* – replacing pronouns with corresponding names; merging different surface forms of names into single entity
  • *Semantic labeling* – assigning semantic identifiers to names (e.g. LOD/DBpedia/Freebase) including disambiguation
  • *Topic classification* – assigning topic categories to a document (e.g. DMOz)
  • *Summarization* – assigning importance to parts of a document
  • *Fact extraction* – extracting relevant facts from a document
Slovenia's dramatic win over Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near domination from the likes of Brazil, Italy and Germany, international soccer is entering the era of the Cinderella. It may not happen this time around, but given the increasing flow of talent, training and information across borders, it's almost certain that a small upstart nation blessed with good athletes and better luck will make a legitimate run at the world covetted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Wednesday.

“Enrycher” is available as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization, Sentiment

Diego Maradona Semantics:
owl:sameAs: http://dbpedia.org/resource/Diego_Maradona
owl:sameAs: http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrCN5Y29ycA
rdf:type: http://dbpedia.org/class/yago/ArgentinianInternationalFootballers
rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers

Robbie Keane Semantics:
owl:sameAs: http://dbpedia.org/resource/Robbie_Keane
rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly
rdf:type: http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers
**Enrycher** is a web service consisting of a set of interlinked modules...

...covering lexical, linguistic and semantic annotations

...exporting data in XML or RDF

To execute the service, one should send an HTTP POST request, with the raw text in the body:

```bash
curl -d "Enrycher was developed at JSI, a research institute in Ljubljana. Ljubljana is the capital of Slovenia." http://enrycher.ijs.si/run
```
Cross-linguality

http://xling.ijs.si/
Cross-linguality

How to operate in many languages?

• Cross-linguality is a set of functions on how to transfer information across the languages
  • ...having this, we can track information independent of the language borders
  • Machine Translation is expensive and slow, so the goal is to avoid machine translation to gain speed and scale

• The key building block is the function for comparing and categorization of documents in different languages
  • XLing.ijs.si is an open web service to bridge information across 100 languages
<table>
<thead>
<tr>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>Chinese</td>
<td>Hungarian</td>
<td>Malagasy</td>
<td>Simple English</td>
<td>Slovak</td>
<td></td>
</tr>
<tr>
<td>Albanian</td>
<td>Chuvash</td>
<td>Icelandic</td>
<td>Malay</td>
<td>Slovenian</td>
<td>Slovenian</td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>Croatian</td>
<td>Ido</td>
<td>Malayalam</td>
<td>Spanish</td>
<td>Spanish</td>
<td></td>
</tr>
<tr>
<td>Aragonese</td>
<td>Czech</td>
<td>Indonesian</td>
<td>Marathi</td>
<td>Sundanese</td>
<td>Spanish</td>
<td></td>
</tr>
<tr>
<td>Armenian</td>
<td>Danish</td>
<td>Irish</td>
<td>Nepali</td>
<td>Swahili</td>
<td>Sundanese</td>
<td></td>
</tr>
<tr>
<td>Asturian</td>
<td>Dutch</td>
<td>Italian</td>
<td>Newar / Nepal Bhasa</td>
<td>Swedish</td>
<td>Swedish</td>
<td></td>
</tr>
<tr>
<td>Azerbaijani</td>
<td>English</td>
<td>Japanese</td>
<td>Norwegian (Bokmå)</td>
<td>Tagalog</td>
<td>Thai</td>
<td></td>
</tr>
<tr>
<td>Bashkir</td>
<td>Esperanto</td>
<td>Javanese</td>
<td>Norwegian (Nynorsk)</td>
<td>Tamil</td>
<td>Turkish</td>
<td></td>
</tr>
<tr>
<td>Belarusian</td>
<td>Estonian</td>
<td>Kazakh</td>
<td>Occitan</td>
<td>Tagalog</td>
<td>Ukrainian</td>
<td></td>
</tr>
<tr>
<td>Belarusian (Taraškievica)</td>
<td>Finnish</td>
<td>Kirghiz</td>
<td>Persian</td>
<td>Turkish</td>
<td>Urdu</td>
<td></td>
</tr>
<tr>
<td>Bengali</td>
<td>French</td>
<td>Korean</td>
<td>Piedmontese</td>
<td>Uzbek</td>
<td>Uzbek</td>
<td></td>
</tr>
<tr>
<td>Bishnupriya Manipuri</td>
<td>Galician</td>
<td>Kurdish</td>
<td>Polish</td>
<td>Vietnamese</td>
<td>Vietnamese</td>
<td></td>
</tr>
<tr>
<td>Bosnian</td>
<td>Georgian</td>
<td>Latin</td>
<td>Portuguese</td>
<td>West Frisian</td>
<td>West Frisian</td>
<td></td>
</tr>
<tr>
<td>Breton</td>
<td>Greek</td>
<td>Latvian</td>
<td>Quechua</td>
<td>Western Panjabi</td>
<td>Yoruba</td>
<td></td>
</tr>
<tr>
<td>Burmese</td>
<td>Gujarati</td>
<td>Lithuanian</td>
<td>Romanian</td>
<td>Yoruba</td>
<td>Yoruba</td>
<td></td>
</tr>
<tr>
<td>Cantonese</td>
<td>Haitian</td>
<td>Lombard</td>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>Hebrew</td>
<td>Low Saxon</td>
<td>Serbian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cebuano</td>
<td>Hindi</td>
<td>Luxembourgish</td>
<td>Serbo-Croatian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Macedonian</td>
<td>Sicilian</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
News Reporting Bias

http://aidemo.ijs.si/diversinews/
News Reporting Bias example

UK soldiers cleared in Iraqi death – Seven British soldiers were acquitted on Thursday of charges of beating an innocent Iraqi teenager to death with rifle butts. A judge at a specially convened military court in eastern England ordered the adjudicating panel to return ‘not guilty’ verdicts against the seven because he did not believe there was sufficient evidence against them, the Ministry of Defence said. . . .

British murderers in Iraq acquitted – The judge at a court-martial on Thursday dismissed murder charges against seven soldiers, from the 3rd Battalion, the Parachute Regiment, who’re accused of murdering Iraqi teenager; claiming there’s insufficient evidence to secure a conviction, The Associated Press reported Thursday. . . .
Detecting News Reporting Bias

• The task:
  • Given a news story, are we able to say from which news source it came?

• We compared CNN and Aljazeera reports about the same events from the war in Iraq
  • ...300 aligned articles describing the same story from both sources

• The same topics are expressed in both sources with the following keywords:
  • CNN with:
    • Insurgents, Troops, Baghdad, Iran, Militant, Police, Suicide, Terrorist, United, National, Hussein, Alleged, Israeli, Syria, Terrorism…
  • Aljazeera with:
    • Attacks, Claims, Rebels, Withdrawing, Report, Fighters, President, Resistance, Occupation, Injured, Army, Demanded, Hit, Muslim, …
DiversiNews iPad App (1/2)

- DiversiNews iPad App is using newsfeed.ijs.si and enrycher.ijs.si services

- ...in its initial screen is shows list of current hot topics and current trending events
DiversiNews iPad App (2/2)

• DiversiNews “diversity search” screen allows dynamic reranking of articles describing an event along three dimensions:
  • **Geography** – where is a content being published from
  • **Subtopics** – what are subtopics of an event
  • **Sentiment** – what are good and what are bad news

• For each query it provides
  • Automatically generated summary
  • List of corresponding articles
Event Representation

http://eventregistry.org/
What is an event?  
(abstract description)

• more practical question: what definition of is computationally feasible?

• In general, an event is something which “sticks out” of the average in some kind of (high dimensional) data space
  • could be interpreted as an “anomaly”
  • densification of data points (e.g. many similar documents)
  • significant change of distribution (e.g. a trend on Twitter)

• In practice, the event could be:
  • A cluster of documents / change of a distribution in data
    • Detected in an unsupervised way
  • A fit to a pre-built model
    • Detected in a supervised way
How to represent an event?

• Baseline data for a news event is usually a cluster of documents
  • ...with some preprocessing we extract linguistic and semantic annotations
  • ...semantic annotations are linked to ontologies providing possibility for multiresolution annotations

• Three levels of event representation:
  • Feature vector event representation:
    • ...light weight representation that can be easily represented as a set of feature vectors augmented with external ontologies – suitable for scalable ML analysis
  • Structured event representation:
    • Infobox representation (slots filling) using open schema or event taxonomy
  • Deep event representation
    • Semantic representation linked to a world-model (e.g. CycKB common sense knowledge)
      – suitable for reasoning and diagnostics
Feature vector event representation

- Feature vectors easily extractable from news documents:
  - **Topical dimension** – what is being talked about? (keywords)
  - **Social dimension** – which entities are mentioned? (named entities)
  - **Temporal aspect** – what is the time of an event? (temporal distribution)
  - **Geographical aspect** – where an event is taking place? (location)
  - **Publisher aspect** – who is reporting? (publisher identifiers)
  - **Sentiment/bias aspect** – emotional signals (numeric estimates)

- Scalable Machine Learning techniques can easily deal with such representation
  - …in “Event Registry” system we use this representation to describe events
Example of “feature vector” event representation: Event Registry “Chicago” related events

Query: “Chicago”

Where? (geography)

When? (temporal distribution)

Who? (named entities)

What? (keyword/topics)
Structured event representation

- Structured event representation describes an event by its “Event Type” and corresponding information slots to be filled

- Event Types should be taken from “Event Taxonomy”

- At this stage of development this level of representation still requires human intervention to achieve high accuracy (Precision/Recall) extraction

- Example on the right – Wikipedia event infobox:
  - 2011 Tōhoku earthquake and tsunami
“Event Taxonomy” – preview to the current development

Disasters_and_accidents_newsEvent
Health_and_environment_newsEvent
Arts_and_culture_newsEvent
Business_and_finance_newsEvent
Sports_newsEvent
Religion_newsEvent

Natural_disaster
Birth
Death

Earthquake
Avalanche
Landslide
Flood
Sinkhole
Hailstorm
Cyclone
Volcanic_erosion
Tropical_storm
Industrial_disaster
Nuclear_disaster
Transportation_accident
Structural_collapse
Chemical_disaster
Man_made_disaster
Airplane_accident
Maritime_accident
Medical_accident
Space_accident
Road_accident
Prototype for event Infobox extraction: XLike annotation service

• The goal is to build a system for economically viable extraction of event infoboxes
  • using crowd-sourcing
  • aiming at high Precision & Recall for a small cost
Event sequences & Hierarchical events

• Once having events identifies and represented we can connect events into “event sequences” (also called story-lines)

• “Event sequences” include events which are supposedly related and constitute larger story

• Collection of interrelated events can be also organized in hierarchies (e.g. World Cup event consists from a series of smaller events)
An example event: Microsoft Windows 9

Windows 9 release date, price, features: why Microsoft can't wait to launch Windows 9 in 2015

WHEN: April 23, 2014

We outline what we know about Windows 9 release date, Windows 9 price and Windows 9 features. It's fair to say that Windows 8 has not been a universal success. In many ways a necessary step not terribly well executed, Windows 8's attempt to stitch together desktop and mobile has been way too radical to encourage new users. And as the longer life of existing PCs and laptops combines with the growth of smartphones and tablets to continually slow PC sales, Window 8 has come to be seen as a dud....
### Similar events example: similar events to Microsoft Windows 9 event

<table>
<thead>
<tr>
<th>Similarity</th>
<th>Event Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>83%</td>
<td>Windows 8.1 Update: Five features we love</td>
<td>2014/04/04</td>
</tr>
<tr>
<td>82%</td>
<td>Windows 8.2 Rumors: New Start Menu, Apps on Desktop, More</td>
<td>2013/12/18</td>
</tr>
<tr>
<td>81%</td>
<td>Windows 9 release date, price, features: new Start menu and no Charms bar</td>
<td>2014/08/13</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft's new, new Windows cadence: Once a year is not enough</td>
<td>2014/04/26</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft making Windows free on devices with screens under nine inches</td>
<td>The Verge</td>
</tr>
<tr>
<td>77%</td>
<td>Here are the latest details on Windows 9 and Microsoft's other major Windows updates</td>
<td>2014/06/25</td>
</tr>
<tr>
<td>77%</td>
<td>Mouse and keyboard-friendly Windows 8.1 update arrives April 8</td>
<td>2014/04/02</td>
</tr>
<tr>
<td>77%</td>
<td>Five Places to Learn How to Use Windows 8.1 Better</td>
<td>2014/01/18</td>
</tr>
<tr>
<td>76%</td>
<td>Microsoft's 'Project Siena' lets anyone create Windows 8.1 apps</td>
<td>2013/12/19</td>
</tr>
<tr>
<td>76%</td>
<td>Microsoft strips some Windows 7 users of IE11 patch privileges</td>
<td>2014/07/11</td>
</tr>
<tr>
<td>76%</td>
<td>No Microsoft Start Menu for Windows 8 until 2015</td>
<td>2014/06/02</td>
</tr>
<tr>
<td>76%</td>
<td>Intel plans a CES coup: Android and Windows in the same computer</td>
<td>2014/01/03</td>
</tr>
<tr>
<td>76%</td>
<td>Why Microsoft Is Already Dropping Support for Windows 8.1</td>
<td>2014/04/25</td>
</tr>
<tr>
<td>76%</td>
<td>New Windows 'Threshold' rumor claims virtual desktop support is being tested</td>
<td>2014/08/06</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft Hopes Windows 9 Will Win Over Desktop Users</td>
<td>2014/07/03</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft showcases latest updates to Windows, opportunities for developers</td>
<td>2014/04/03</td>
</tr>
<tr>
<td>75%</td>
<td>Windows 8.1 update sweeps 'Metro' under the rug, boots to desktop by default</td>
<td>2014/02/02</td>
</tr>
<tr>
<td>74%</td>
<td>In Six Months, Microsoft Will Pull The Plug On Some Parts Of Windows 7 Support</td>
<td>MSFT</td>
</tr>
<tr>
<td></td>
<td>Microsoft talks Windows 8.1 Update 1 features, April 8 release, teases future Start menu</td>
<td>2014/04/05</td>
</tr>
</tbody>
</table>
Event sequence identification
Hierarchy of events

- **Story**
  - Margaret Thatcher's Life
  - Margaret Thatcher's Death

- **Events**
  - Margaret Thatcher's Funeral

- **Article**
  - "Last respects paid to Lady Thatcher"
Example Microsoft hierarchy of events
Zoom-in Example
Microsoft hierarchy of events
Event Tracking

http://eventregistry.org/
Live Event tracking with http://EventRegistry.org/

Obama, Harper discuss NATO summit, Iraq in phone call:
White House
Location: Washington D.C., United States
Articles about event: 17

Lt Gen Tijendar Singh to be arrested in connection with Tatra truck deal
Location: New Delhi, India
Articles about event: 11
<table>
<thead>
<tr>
<th>Entities</th>
<th>Concept Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>100</td>
</tr>
<tr>
<td>National Security Agency</td>
<td>84</td>
</tr>
<tr>
<td>United States</td>
<td>70</td>
</tr>
<tr>
<td>Washington D.C.</td>
<td>42</td>
</tr>
<tr>
<td>Josiah Bartlet</td>
<td>41</td>
</tr>
<tr>
<td>President of the United States</td>
<td>39</td>
</tr>
<tr>
<td>United States Foreign Intelligence</td>
<td>28</td>
</tr>
<tr>
<td>Angela Merkel</td>
<td>26</td>
</tr>
<tr>
<td>Presidency of George W. Bush</td>
<td>24</td>
</tr>
<tr>
<td>Chancellor of Germany (Federal)</td>
<td>20</td>
</tr>
<tr>
<td>Executive Office of the President</td>
<td>19</td>
</tr>
<tr>
<td>Yakov Smirnoff</td>
<td>18</td>
</tr>
<tr>
<td>Supreme Court of the United States</td>
<td>17</td>
</tr>
<tr>
<td>Republican Party (United States)</td>
<td>17</td>
</tr>
<tr>
<td>Romeo Miller</td>
<td>16</td>
</tr>
<tr>
<td>Jay Carney</td>
<td>11</td>
</tr>
<tr>
<td>United States Department of Justice</td>
<td>8</td>
</tr>
<tr>
<td>Larry Klayman</td>
<td>8</td>
</tr>
<tr>
<td>European Parliament</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Concept Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveillance</td>
<td>67</td>
</tr>
<tr>
<td>White House</td>
<td>66</td>
</tr>
<tr>
<td>Privacy</td>
<td>63</td>
</tr>
<tr>
<td>Espionage</td>
<td>54</td>
</tr>
<tr>
<td>Review</td>
<td>54</td>
</tr>
<tr>
<td>Telephone</td>
<td>54</td>
</tr>
<tr>
<td>Meeting</td>
<td>49</td>
</tr>
<tr>
<td>Leadership</td>
<td>48</td>
</tr>
<tr>
<td>Official</td>
<td>48</td>
</tr>
<tr>
<td>Legislator</td>
<td>47</td>
</tr>
<tr>
<td>United States Congress</td>
<td>46</td>
</tr>
<tr>
<td>Debate</td>
<td>40</td>
</tr>
<tr>
<td>Board of directors</td>
<td>38</td>
</tr>
<tr>
<td>Foreign policy</td>
<td>35</td>
</tr>
<tr>
<td>Report</td>
<td>35</td>
</tr>
<tr>
<td>Data</td>
<td>32</td>
</tr>
<tr>
<td>Intelligence</td>
<td>30</td>
</tr>
<tr>
<td>National security</td>
<td>29</td>
</tr>
<tr>
<td>Document</td>
<td>29</td>
</tr>
</tbody>
</table>
Collection of events described through Entity relatedness
Trending of concepts in search results

Collection of events described through trending concepts
Collection of events described through three level categorization
Events identified across languages
Collection of events described through Reporting dynamics
Collection of events described through a story-line of related events
Event Registry exports event data through API and RDF/Storyline ontology

- API to search and export event information
  - Export of all the system data in JSON

- Event data is exported in a structured form
  - BBC Storyline ontology

- SPARQL endpoint:
  - http://eventregistry.org/rdf/search
    - http://eventregistry.org/rdf/event/{eventID}
    - http://eventregistry.org/rdf/article/{articleID}
    - http://eventregistry.org/rdf/storyline/{storylineID}
    - Example: http://eventregistry.org/rdf/event/1234
Future / Follow-up projects
Some of the follow-up projects

- Understanding global social dynamics
  - How global society functions?

- Integrating text-based media with TV channels
  - ...requires speech recognition, video processing, visual object recognition, face recognition, ...

- Event prediction / Event-Consequence prediction
  - ...requires understanding of causality in the social dynamics and much more

- Micro-reading / Machine-reading
  - ...full understanding of individual documents – the goal for 10+ years