Benefits of multilingual media monitoring

Ralf Steinberger
Ralf.Steinberger@jrc.ec.europa.eu

& the JRC’s OPTIMA team - Open Source Text Information Mining and Analysis

http://langtech.jrc.ec.europa.eu/
http://press.jrc.it/overview.html

Agenda

• Who we are: European Commission – Joint Research Centre
• Who our users are
• Overview of what we do, what services we provide
  • Europe Media Monitor (EMM) applications - Functionality
  • Publicly accessible at http://press.jrc.it/overview.html
• Importance of multilinguality
• User requirements
• Our insights regarding needs
Joint Research Centre - Who we are

- European Commission (scientific-technical arm of public administration)
- Non-commercial
- Highly multi-disciplinary
- Serving the EU → multilingual
- Relatively small team working on Language Technology

Media Monitoring – Potential Users / Motivation

- Any large organisation wants to monitor
  - newest developments in their field of interest
  - their public image, people’s views on certain issues, etc.
  - …

- Institutions and governments, e.g. EC, UN, National Public Health institutions
  - Security-related events (bombings, hostage takings, piracy, …)
  - Humanitarian events (displaced people, shortage of water/medicines, natural disasters…)
  - Threats to public health (contagious diseases, nuclear or chemical leaks, …)
  - Scientific and technological news
  - …

- The public (journalists, specialists, anybody)
  - General information / following the news
  - Meta-news sites are independent of individual (possibly biased) news sources
Media Monitoring – EMM Users

- **European Commission** (most DGs) and other EU Institutions
- **EU Agencies**:  
  - e.g. Public Health (ECDC), Food Safety (EFSA), Chemicals Bureau (ECHA), etc.
- **EU Member State organisations**: e.g.
  - Public Health,
  - law enforcement authorities,
  - parliaments,
  - crisis management/humanitarian
- **International and extra-European organisations**: e.g.
  - various UN organisations
  - Centres for Disease Prevention and Control in the US, Canada, China, …
- **The public**:  
  - Ca. 30,000 anonymous users of publicly accessible EMM systems.
  - Combined between 1 and 2 Million hits per day
  - [http://press.jrc.it/overview.html](http://press.jrc.it/overview.html)

Functionality of EMM systems. What we do

---

What we do: 1. News Gathering

- EMM news gathering engine
  - Monitors ~2,200 news sources
  - Gathers 80,000 – 100,000 news articles per day
  - In about 50 languages
  - Visits some sites every 5 minutes
  - Extracts text from the web page
  - Converts text into Unicode-encoded RSS
  - Feeds the news into the four publicly accessible media monitoring systems

What we do: 2. News Analysis

- Group related news multi-monolingually

- **Categorise news** into hundreds of categories, including countries

- Extract information (entities and events)

- Track related news over time and across languages

- Collect multilingual information on entities

- Generate social networks based on info extracted from multilingual news

- Provide **tools for human moderation and dissemination**
EMM – NewsBrief (currently 50 languages)

- Categorises news into ~600 categories, using:
  - Boolean search word combinations
  - vicinity operators
  - optional weights
  - regular expressions
- Clusters and tracks news live (multi-monolingually)
- Detects breaking news
- Sends out email notifications for each category
- Also accessible as RSS, KML
EMM-Labs: Real-time Cluster Network

Display of relations between clusters and categories

Cluster 188: Former Nazi death camp guard taken to Germany
Cluster 11: Israel hands over Lebanon cluster bomb maps - UN
Cluster 110: Pope meets Abbas, calls for Palestinian homeland
Cluster 30: U.S. wins a spot on U.N. Human Rights Council
Cluster 15: Obama to meet ME leaders to talk peace

MedISys – Filtering and classification in up to 50 languages

MedISys home page

Latest News - Mustard gas

Number of articles for this alert: 5
Articles per 10,000 searched by the system: 0

Date: 19/09/2007

The country values are calculated by setting the number of articles mentioning a theme and a country in relation to the number of articles about the theme and the number of articles about the country.
Aggregation of multilingual information

- Documents from all languages get classified according to the same countries and categories.
- An increase of the number of media reports on any country-category combination is detected, independently of the reporting language.
- Graphs and alerts may show events not yet reported in your own language.

Detection and visualisation of events (violence/disasters/humanitarian/…)

Objective: global crisis monitoring. Languages: En, Fr, Es, It, Pt, Ru + (Ar)

---

Detection and visualisation of events (violence/disasters/humanitarian/…)

NewsExplorer – Functionality

http://emm.newsexplorer.eu/

Since April 2004
Pakistan court dismisses final Musharraf challenge (32) [es]

Pakistan’s new look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf’s re-election on Wednesday 21.00 PM CET.

Chinese cluster on Tibet and Taiwan [es] as it is sv

Contrary to expectations, China is not doing much to soften its image ahead of the Beijing Olympics by allowing its domestic critics to speak their minds or championing human rights in Sudan. Instead, Chinese leaders are extending authoritarian rules at home and abroad and using aggressive diplomacy against those who

Fidel Castro quits as president, state-run paper reports [72] [es] [en]

Fidel Castro announced his resignation as president of Cuba and commander-in-chief of Cuba’s military on Tuesday, according to a letter published by state-run newspaper Granma. on 9:23:00 AM CET

Kubo: Fidel Castro gibt das Zepter ab [de]

Konrad Adenauer (1949-1963)

Der legendäre Lehrer des Einheitsdeutschlandes

PRÄZISION: Kubanische Fidel Castro po 40 latach rządzi zrezygnował we wnioskach z funkcji lidera rewolucyjnego Rady Państwowej

Kuba: Castro gaur sa [es]

Fidel Castro renuncia a la Presidencia del Consejo de Estado

Cubaanse president Fidel Castro afgezet [nl]

Fidel Castro, Fidel Castro regenta en donderdag de president van Cuba af. in editie elektronica en cotidotimag-

Kuba: Fidel Castro se Jose odpowiedzialność kubańskiego przedziera [pl]

Powyższy kubański Fidel Castro po 40 latach rządów zrezygnował we wnioskach z funkcji lidera rewolucyjnego Rady Państwowej.

Kuba: Castro se otariča od poteša predsednika Kubske Sabora [sl]

Fidel Castro se je odpredstavil zadnji predsednika [sl]

Este es un tema muy importante que ha sido objeto de debate en varios foros internacionales.
NewsExplorer – Time line: biggest clusters per day

International: Hussein’s execution

German national breaking news

NewsExplorer – Aggregation of clusters into longer ‘stories’

Thursday, November 22, 2007

Pakistan court dismisses final Musharraf challenge

Pakistan’s new-look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf’s re-election, extending his five-year rule today. AIJ’s chief justice, Iftikhar Muhammad Chaudhry, announced the decision.

Emergency rule ‘destroying the judiciary’

Bing.com: 8:46 PM CEST

Musharraf’s emergency rule (dictatorship)

Emergency to deep Pakistan: 13:32 PM CEST

Pakistan’s Commonwealth suspension in sorrow, not anger: Britain (AIR)

NewsExplorer

This cluster belong to the following story: Pakistan defends emergency rule
NewsExplorer – Information about people collected from multiple languages and over time

NewsExplorer – Relation exploration

Example: Pervez Musharraf & Iftikhar Chaudhry
Text analysis tools integrated in NewsExplorer

- Document **clustering**
- **Named entity recognition** (persons, organisations)
- Name **variant matching**, including across scripts (transliteration, string similarity calculation)
- **Geo-tagging** (recognition and disambiguation of locations)
- Multi-label document **categorisation**
- **Quotation** recognition (and reference resolution for name parts)
- **Social network generation** (based on co-occurrence, quotations, relation types)
- **Topic detection and tracking** (TDT) (Multi-monolingual cluster similarity calculation)
- **Cross-lingual cluster similarity** calculation (for cross-lingual TDT).

**Importance of multilinguality**
Multilinguality: complementary news coverage in various languages

Locations mentioned in MedISys medical articles across languages – complementary coverage

- Italian - German
- English - French
- Spanish - Portuguese

Multilinguality: Gathering more information about people

Alexander Litvinenko
Information about this person was last updated on Dienstag, 20. März 2007.

<table>
<thead>
<tr>
<th>Names</th>
<th>Key Titles and Phrases</th>
<th>External resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander Litvinenko (Eu,nl)</td>
<td>russo (it,pt - 349)</td>
<td>Image obtained automatically from Wikipedia</td>
</tr>
<tr>
<td>Alexander Litvinenko (de)</td>
<td>agent russe (fr - 134)</td>
<td></td>
</tr>
<tr>
<td>Aleksandr Litvinenko (fr)</td>
<td>nisa (cs - 7NR)</td>
<td></td>
</tr>
<tr>
<td>Aleksandar Litvinenko (nl,sv)</td>
<td>agenter (de,sv - 134)</td>
<td></td>
</tr>
<tr>
<td>Aleksandra Litvinenko (ru)</td>
<td>kritikers (de - 79)</td>
<td></td>
</tr>
<tr>
<td>Aleksande Litvinenko (ru)</td>
<td>agent (en,sv - 130)</td>
<td></td>
</tr>
<tr>
<td>Aleksandar Litvinenko (uk)</td>
<td>russa (it,pt - 76)</td>
<td></td>
</tr>
<tr>
<td>Alexander V. Litvinenko (en)</td>
<td>secret russe (fr - 29)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko (it)</td>
<td>russe (de,fr - 73)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko (es)</td>
<td>formed russian agent (en - 20)</td>
<td></td>
</tr>
<tr>
<td>Alexandre Litvinenko (fr)</td>
<td>morte ci (it - 45)</td>
<td></td>
</tr>
<tr>
<td>Alexandre Litvinenko (en)</td>
<td>russe agenten (sv - 13)</td>
<td></td>
</tr>
<tr>
<td>Oleksandr Lytvynenko (en)</td>
<td>kritiker (de - 13)</td>
<td></td>
</tr>
<tr>
<td>Oleksandr Litvinenko (fr)</td>
<td>43 ans (fr - 17)</td>
<td></td>
</tr>
<tr>
<td>Aleksander Litvinenko (hr)</td>
<td>russe (it - 14)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko (it)</td>
<td>russian (en - 15)</td>
<td></td>
</tr>
<tr>
<td>Александр - Литвиненко (ja)</td>
<td>urugendu (it - 11)</td>
<td></td>
</tr>
<tr>
<td>Alexander Walterowitsch Litvinenko (do)</td>
<td>officer (en - 13)</td>
<td></td>
</tr>
</tbody>
</table>

Multilingual Information Access Technology Transfer Day, Berlin, Germany, 8 December 2009

**Castro quits as president, state-run paper reports [72]**

Fidel Castro announced his resignation as president of Cuba and commander-in-chief of the island’s military on Tuesday, according to a letter published by state-run newspaper Granma.

castro quits as president, state-run paper reports [72]

27

**Multilingual Information Access Technology Transfer Day, Berlin, Germany, 8 December 2009**

**Multilinguality: less news bias and more transparency**

News aggregators

- Google News: 45 languages
- Yahoo News: 35 languages
- Europe Media Monitor: 50 languages

News analysis systems

- NewsExplorer: 19 languages
- NewsTin: 11 languages
- Daylife: 1 language
- SkyBreaker: 1 language
- NewsLINE: 1 language

**Language coverage of various media analysis tools (August 2009)**

in early 2008: 34, 17, 43 languages

in early 2008: the same

28

News aggregators

- Google News: 45 languages
- Yahoo News: 35 languages
- Europe Media Monitor: 50 languages

News analysis systems

- NewsExplorer: 19 languages
- NewsTin: 11 languages
- Daylife: 1 language
- SkyBreaker: 1 language
- NewsLINE: 1 language

**Language coverage of various media analysis tools (August 2009)**

in early 2008: 34, 17, 43 languages

in early 2008: the same
Social networks produced on the basis of many languages are more complete and less biased.

EMM user requirements

EMM insights
EMM user requirements: many languages; summarisation

- Many languages, more languages, depending on user group
  - Majority of EU institutions: the 23 official EU languages
  - Public Health: Asian, African
  - Nuclear Safeguards: countries with nuclear facilities or aspirations
  - Humanitarian Aid: crisis areas, e.g. in Africa, Asia or Latin America

- Multilingual, multi-document summarisation, update summarisation. But: how?

US President Obama to Set Out Afghan Troop Strategy

Nato commits additional Afghanistan troops

EMM user requirements: Machine Translation

- Machine Translation, including for less widely spoken languages
- Although even other simple CLIA methods are useful
- MT not on demand, but pre-translated (expensive)
EMM user requirements: multilingual information fusion

- **Information fusion** before presentation to the users would save time and effort.

EMM insight: source language analysis better than MT

- Translating everything into one target language is good, but has drawbacks.
  - Important specific terms and names are often not translated, or translated badly, e.g.
    - Person and organisation names
    - Locations
    - Names of diseases and toxins

- Our insight: source language information extraction is better for information filtering than searching machine-translated text.
  - MT as a second step.
EMM insight: keep it simple!

- Keep applications simple!
  - Not only from the user’s point of view, but also from the developer’s point of view, e.g. multilingual EMM category definitions
    - Manually produced Boolean combinations or cumulative weights, using wild cards
    - Faster and easier than learning categorisers for hundreds of classes in 50 languages
  - Ideally language-independent rules, to which language-specific resources can be plugged in
  - If possible, avoid language pair-specific resources
    - $N^2(N-1)/2$ language pair combinations
    - NewsExplorer’s 19 languages: 171 language pairs!

---

Summary

- Europe Media Monitor applications (EMM; http://press.jrc.it/overview.html):
  - Users
  - Functionality

- Importance of multilinguality
- EMM user requirements
- Our insights