Highly multilingual news monitoring applications

ECML PKDD 2009
European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

Ralf Steinberger
& the JRC’s OPTIMA team
(Open Source Text Information Mining and Analysis)

http://langtech.jrc.it/
http://press.jrc.it/overview.html
BRUSSELS (BE)
- The Directorate General (DG)
- The Institutional and Scientific Relations Directorate (ISR)
- The Programme and Resource Management Directorate (PRM)

GEFL (BE)
- The Institute for Reference Materials and Measurements (IRMM)

KARLSRUHE (DE)
- The Institute for Transuranium Elements (ITU)

ISPRA (IT)
- Download the Ispra site Brochure (English - Italian)
- The Institute for the Protection and Security of the Citizen (IPSC)
- The Institute for Environment and Sustainability (IES)
- The Institute for Health and Consumer Protection (IHCP)
- The Ispra site Directorate (IS)

PETTEN (NL)
- The Institute for Energy (IE)

SEVILLE (E)
- The Institute for Prospective Technological Studies (IPTS)
Agenda

• Media Monitoring and multilinguality
• Europe Media Monitor (EMM) applications - Functionality
  • Publicly accessible at http://press.jrc.it/overview.html

• Some language technology components in detail
  • Multilingual person name recognition
  • Name variant matching across many languages
  • Social networks based on multilingual information extraction
  • Cross-lingual cluster linking
• Summary and Ongoing work
• 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, …)
  • Threats to public health (contagious diseases, nuclear or chemical leaks, …)
  • …

• The public (journalists, specialists, anybody)
  • General information / following the news
  • Meta-news sites are independent of individual (possibly biased) news sources
European context – Motivation for multilinguality

• The European Union has 23 official languages,
  • Plus other, non-official ones
  • Plus an interest in non-EU languages (Turkish, Croatian, Arabic, Russian, Farsi, Chinese, …)

• Political and practical need to cover many languages

• Ideally no pivot language, but direct information access between the language pairs

• The EC’s Translation Service DGT is one of the largest in the world (2350 staff plus freelancers, see http://ec.europa.eu/dgs/translation/index_en.htm)

• Complementary coverage of the news in different languages
Multilinguality: complementary news coverage in various languages

Locations mentioned in MedISys medical articles across languages – complementary coverage

Italian - German

English - French

Spanish - Portuguese
Language coverage of various media analysis tools (August 2009)

News aggregators

<table>
<thead>
<tr>
<th>Service</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google News</td>
<td>45</td>
</tr>
<tr>
<td>Yahoo News</td>
<td>35</td>
</tr>
<tr>
<td>Europe Media Monitor</td>
<td>50</td>
</tr>
</tbody>
</table>

News analysis systems

<table>
<thead>
<tr>
<th>Service</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NewsExplorer</td>
<td>19</td>
</tr>
<tr>
<td>NewsTin</td>
<td>11</td>
</tr>
<tr>
<td>Devlife</td>
<td>1</td>
</tr>
<tr>
<td>SiloBreaker</td>
<td>1</td>
</tr>
<tr>
<td>NewsVine</td>
<td>1</td>
</tr>
</tbody>
</table>

in early 2008: 34, 17, 43 languages

in early 2008: the same
Agenda

• Media Monitoring and multilinguality
• **Europe Media Monitor (EMM) applications** - Functionality
  • Publicly accessible at [http://press.jrc.it/overview.html](http://press.jrc.it/overview.html)
  • Some language technology components in detail
    • Multilingual person name recognition
    • Name variant matching across many languages
    • Social networks based on multilingual information extraction
    • Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc
• Summary and Ongoing work
Europe Media Monitor (EMM) - News data

- 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

- Combined between 1 and 2 Million hits per day
- 30,000 – 50,000 distinct users per day

EMM – NewsBrief (currently 50 languages)

- Public site: http://press.jrc.it/NewsBrief/ (since 2002)
- 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

Live display of latest geo-located news clusters
EMM-Labs: Real-time Cluster Network

Display of relations between clusters and categories
MedISys – Filtering and classification in up to 50 languages
• 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.
Objective: global crisis monitoring.  

Languages: En, Fr, Es, It, Ru + (Ar)

В Турции террористы захватили 15 заложников

Last update 2009-05-13T17:35+0200

Две неизвестных в масках совершили в среду вооруженный напад на филиал банка в турецком курортном городе Кушадасы на побережье Эгейского...

Event Type: Kidnapping/Hostage Taking. Severity: 0 Killed, 0 Injured and 15 Victims were no one killed and no one injured.

На турецком курорте грабители взяли в заложники посетителей банка

Освобождены захваченные в турецком банке заложники

Last update 2009-05-13T17:46+0200

Турецкая полиция обезвредила налетчика, захватившего 10 человек в банке в курортном городе Кушадасы на побережье Эгейского моря, передают местные СМИ. Полиция пошла на штурм банка....

Сдался полиции преступник, захвативший 10 заложников в банке Кушадасы

Last update 2009-05-13T17:21+0200

Турецкая полиция обезвредила налетчика, захватившего в заложники более 10 человек в банке в курортном городе Кушадасы на побережье Эгейского моря,......

Полиция Турции обезвредила налетчика, захватившего заложников в банке

Last update 2009-05-13T16:54+0200

Дrama с заложниками продолжалась более шести часов - с 11.00 местного времени (12.00 мск). По сообщение телеканалов NTV и CNN-Turk, полиции удалось убедить налетчика сдаться. Подробности проведения операции пока неизвестны....

In Turkey, an unknown masked seized 13 hostages in the bank

Publication time: 14:29

The police stormed the bank in the western Turkish city of Kusadasi, where the hostages were 13 people. Robbers arrested, ITAR-TASS reported with reference to state television.
http://press.jrc.it/NewsExplorer

Since April 2004
Clustered news for Thursday, November 22, 2007
Read more...

Pakistan court dismisses final Musharraf challenge [32]
das fr id br pl pt ru sl sr tr
Pakistan's new-lock Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf's re-election.

Iran heeding transparency pledge, more needed - IAEA [32] de es fr it nl

Chinese blister on Tibet and Taiwan [29] de es fr it nl

Cyclone leaves 242 dead in Bangladesh

Claimants warned of fraud fears

This Month's New Stories
November 14, 2007 - November 22, 2007

NewsExplorer – Multilingual daily news overview
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.

cnn 9:23:00 AM CET

Fidel Castro resigns as president, state-run paper reports [72]

Kuba: Fidel Castro gibt das Zepter ab

Der legendäre Kubaner, ein der Partei und seiner Zeitstifter treu ergebener Mann, hat prolongierte Erkrankungen und die Belastungen seiner Rolle nicht mehr bewältigen können. Die Hoffnung, dass er sich bald wieder zurückziehen wird, ist erfüllt. 

Fidel Castro resigned as president of Cuba

Fidel Castro stepped down as Cuba’s president and head of the Communist party on Tuesday after more than half a century in power, state news agency Cubana de Información (Cubinegro) quoted an aide as saying.

Fidel Castro renunció a la Presidencia del Consejo de Estado

Fidel Castro renunció a la presidencia de Cuba

Kubaner Castro regresa a su papel de escritor

Fidel Castro a anuntat, marti, ca renunta la presedintia Cudebi, in editia electronica a cotidianului "NewsExplorer".

Fidel Castro se retractó de la presidencia Cuidihi

Fidel Castro a anunțat, marți, că renunță la președinția Cudebi, în editia electronică a cotidianului "NewsExplorer".

NewsExplorer – Time line: biggest clusters per day

International: Hussein’s execution

German national breaking news

Stoiber will zurücktreten

Timelines and visualizations of news clusters by day.
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.

euronews-en: 23:01 PM CET

Pakistan Court Rules for Musharraf

ADNNews: 2:43 PM CET

‘Bush unembarrassed by Pakistan emergency’

dailytimes:PK: 12:28 AM CET

Pakistan: Emergency and chaos

Related People

Pervez Musharraf (26)

Related Topics

Pakistan (26)

Countries

Pakistan (26)
United States (12)
Saudi Arabia (10)

Places

Islamabad (PK)
Rawalpindi (PK)
Karachi (PK)
Washington (US)
Al Muyar (SA)
Jiddan (SA)

Other Names

Supreme Court (27)
Human Rights Watch (3)
Pakistan Muslim League (2)
High Court (2)
High Commission (2)
GEO Television (2)
Human Rights Commission (1)
Al Qaeda (1)
NewsExplorer – Information about people collected from multiple languages and over time

Pervez Musharraf
Information about this person was last updated on Friday, November 23, 2007.

Names
- Pervez Musharraf (Eu, sv)
- General Pervez Musharraf (da, sv)
- Gen Musharraf (en)
- Pervez Mušarraf (sl)
- Gen Pervez Musharraf (en)
- Pervez Musharraf (da, sv)
- Pervez Musharraf (fr, sv)
- Pervez Musharraf (fr)
- Pervez Musharraf (ar)
- Pervez Musarraf (sl)
- Pervez Musharraf (de)
- Perves Musharraf (de)

Key Titles and Phrases
- pakistani president (en - 827)
- president (de, sv - 3230)
- president gen (en - 506)
- président pakistanaïs (fr - 398)
- pakistaanse president (nl - 235)
- presidente paquistanès (pt - 277)
- pakistani president gen (en - 181)
- presidente paquistaní (es - 147)
- président (de - 617)
- general (en, sv - 450)
- præsident (da - 210)
- presidente (es, pt - 589)
- president (en - 621)
- president (de - 101)

External resources

Latest Clusters – English
[de] [pt] [es] [nl] [fr] [ar] [sv] [it] [da] [pl] [si] [ro] [ru] [no] [bg]

Pakistani court dismisses final Musharraf challenge Struggle to help cyclone survivors in Bangladesh
euronews 22-NOV-07

Musharraf’s Quotes from – English
[en] [ru] [sv] [pt] [nl] [de] [fr] [no] [bg] [it]

[-has said]: Where it fails to live up to those values, it needs to act with credibility and consistency. I think Pakistan is a test in that respect.
bday 22-NOV-07

[has repeated]: The (presidential) oath can be taken ... by the weekend or immediately thereafter.

Quotes about – English
[en] [bg] [de] [pt] [ro] [fr] [no] [nl]

Rice [-said]: And look, a lot of that was done by Musharraf himself. And so for him at this point to help put his country back on the road to democratic reform is important. We’re looking for him to take off his uniform.
expressindia 22-NOV-07

Brad Adams [-said]: Rather than making

Brown [-said]: He (Musharraf) has assured
Example: Pervez Musharraf & Iftikhar Chaudhry
• 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).
Agenda

- Media Monitoring and multilinguality
- Europe Media Monitor (EMM) applications  - Functionality
  - Publicly accessible at [http://press.jrc.it/overview.html](http://press.jrc.it/overview.html)

- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc
  - Summary and Ongoing work
News clustering
Clustering: Monolingual document representation

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)
- Enhanced with log-likelihood weighted country score

“Michael Jackson Jury Reaches Verdicts”

**Monday, June 13, 2005**

**Jackson cleared of all 10 charges**

Jackson cleared of all 10 charges during *9/8/2005 23:36*

**Timetable Of Events Which Led To Court**

18/05/2005 19:30

**Week 2 Of Jackson Deliberations**

ABCnews 13/06/2005 12:07

**Jackson jurors set for second week**

ABCNews 13/06/2005 12:12

**Music’s misunderstood superstar**

13/06/2005 23:03

**The Extraordinary Life Of A Music Icon**

13/06/2005 22:12

**Jury finds Michael Jackson innocent of all charges**

13/06/2005 23:16

**Michael Jackson found not guilty on all charges**

13/06/2005 23:34

**JACKSON NOT GUILTY**

13/06/2005 23:29

**Michael Jackson cleared of abuse**

12/06/2005 23:25

**Analysis: Testing times ahead**

12/06/2005 23:48

**Jackson arrives at court**

12/06/2005 23:57

---

**Keyness** | **Keyword** | **Keyness** | **Keyword**
---|---|---|---
109.2478 | jackson | 7.5620 | testimony
41.5450 | neverland | 6.5014 | maria
37.9347 | santa | 4.0957 | michael
32.6105 | molestation | 1.7368 | reached
24.5193 | boy | 1.6857 | ap
24.4351 | pop | 1.5610 | *gb*
20.6824 | documentary | 1.5610 | *il*
18.7973 | accuser | 1.5610 | *br*
13.5945 | courthouse | 1.0520 | appeared
11.1224 | jury | 0.5384 | child
10.4184 | *us* | 0.5045 | trial
10.0838 | ranch | 0.4502 | monday
9.6021 | california | 0.2647 | children
9.3905 | verdict | 0.0946 | family

---

*Original cluster*
Multi-monolingualse news clustering

- Document similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the binary hierarchical clustering tree (dendrogram)
  - Retain only “big” nodes in the tree with a high cohesion (empirically refined minimum intra-node similarity: 45%)
- Use the title of the cluster’s medoid as the cluster title
Multilingual named entity recognition and variant mapping
<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>death of former Prime Minister Rafik Hariri, blamed by many opposition</td>
</tr>
<tr>
<td>es</td>
<td>asesinato del exprimer ministro Rafic al-Hariri, que la oposición atribuyó</td>
</tr>
<tr>
<td>fr</td>
<td>l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom</td>
</tr>
<tr>
<td>nl</td>
<td>na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een</td>
</tr>
<tr>
<td>de</td>
<td>libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B</td>
</tr>
<tr>
<td>sl</td>
<td>danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si</td>
</tr>
<tr>
<td>et</td>
<td>möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl</td>
</tr>
<tr>
<td>ar</td>
<td>اغتيال رئيس الوزراء السابق رفيق الحريري بأياد يهودية وما حدث سابقا</td>
</tr>
<tr>
<td>ru</td>
<td>Бывший премьер-министр Ливана Рафик Харири, который</td>
</tr>
</tbody>
</table>
Multilingual named entity recognition (NER) – brief overview

- **Lookup of ~1.15 million known names** and variants from database
  - Currently about 870,000 names (status June 2009) + 275,000 variants
  - Dealing with morphological variation by pre-generating morphological variants (Slovene example):
    
    \[\text{Tony}(a|o|u|om|em|m|ju|jem|ja)\]?s+\text{Blair}(a|o|u|om|em|m|ju|jem|ja)\]

- **Guessing new names** using empirically-derived *lexical patterns*
  - President, Minister, Head of State, Sir, American
  - “death of”, “[0-9]+-year-old”, …
  - Known first names + uppercase words
  - Identification of a current average of 608 unknown names per day
Name Variants

- Adding names (and images) from Wikipedia
- Merging NewsExplorer name variants
  - Transliteration
  - Normalisation
  - Similarity measure


Hamid Karzai

From Wikipedia, the free encyclopedia

Hamid Karzai (Persian and Pashto: هامد کرژی; born December 24, 1957) is the current President of Afghanistan (since December 7, 2004). He became the dominant political figure after the removal of the Taliban government. From 2001, Hamid Karzai was the Chairman of the Administration and Interim President. He won the 2004 election.
Currently, EMM NewsExplorer transliterates from Arabic, Farsi, Greek, Russian and Bulgarian.

Transliterate each character, or sequence of characters, by a Latin correspondent:

- \( \psi \rightarrow ps \)
- \( \lambda \rightarrow l \)
- \( \mu \pi \rightarrow b \)

Examples of transliterations:

- Κόφι Ανάν, Greek \( \rightarrow \) Kofi Anan
- Кофи Анан, Russian \( \rightarrow \) Kofi Anan
- Кофі Анан, Bulgarian \( \rightarrow \) Kofi Anan
- ﺑﻮﻠﺳّ, Arabic \( \rightarrow \) Kufi Anan
- कोफी अनन, Hindi \( \rightarrow \) Kofi Anan
Name normalisation – Why? (1)

- Transliteration rules depend on the target language, e.g.
  Владими́р Усти́нов (Russian)
  - Владимир Устинов (English)
  - Влассимир Устинов (German)
  - Владимир Ouстинов (French)

- Various ways to represent the same sound: sh, sch, ch, š, e.g.
  - Ба́шар ал Асса́д
  - Bashar al Assad
  - Bašar al Assad
  - Baschar al Assad
  - Bașar al Assad
Name normalisation – Why? (2)

• Diacritics are often omitted, e.g.
  • Wałęsa → Walesa
  • Saïd → Said
  • Schröder → Schroder
  • Skarsgård → Skarsgard
  • Jørgen → Jorgen

→ Edit distance is large for naturally occurring word variants:
  • “Rafik Hariri” vs. "Rafiq Hariri“ → 2
  • “Rfk Hrr” vs. "Rafiq Hariri“ → 6
Name normalisation (3) - ~ 30 Rules

- Latin normalisation:
  - accented character → non-accented equivalent
    - Malik al-Sai’doullaïev
    - Malik al-Saidoullaïev
  - double consonant → single consonant
    - Malik al-Saidoulaiev
  - ou → u
    - Malik saïdoulaiev
  - “al-” →
    - Malik Saidulaiev
  - vl (beginning of name) → vl
    - Malik Saidulaiev
  - ow (end of name) → ov
    - Malik Saidulaiev
  - ck → k
    - Malik Saidulaiev
  - ph → f
    - Malik Saidulaiev
  - ž → j
    - Malik Saidulaiev
  - š → sh
    - Malik Saidulaiev
  - x → ks
    - Malik Saidulaiev
  - Remove vowels

<table>
<thead>
<tr>
<th>Name</th>
<th>Normalised form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohammed Siad Barre, Mohamed Siad Barré,</td>
<td>mhmd sd br (mohamed siad bare)</td>
</tr>
<tr>
<td>Mohamed Siad Barre, Mohamed Siad Barre,</td>
<td></td>
</tr>
<tr>
<td>Mahmoud Ahmadinejed, Mahmūd Ahmadīnežād</td>
<td>mhmd hmdnjd (mahmud ahmadinejad)</td>
</tr>
<tr>
<td>Сергей Куприянов, Sergei Kupriyanov,</td>
<td>srg kprnv (sergei kuprianov)</td>
</tr>
<tr>
<td>Сергей Куприянов, Sergei Kupriyanov,</td>
<td></td>
</tr>
<tr>
<td>Ban Ki-moon, Ban Ki Moon, Пан Ги Мун</td>
<td>bn k mn (ban ki mun)</td>
</tr>
</tbody>
</table>
Similarity measure for name merging

To compare ~600 new names every day with ~1,145,000 known name variants:

- Only if the transliterated, normalised form with vowels removed is identical
  → Calculate edit distance variant similarity using two different representations:

\[
\begin{align*}
\text{transliteration} & \quad \text{normalisation} \\
\text{jeck sharki} & \quad \text{jack sharki} \\
20\% & \quad 80\% \\
\text{consonants} & \quad \text{consonants} \\
\text{Exact Match (quick)} & \quad \text{Edit distance (slow)}
\end{align*}
\]
Merging name variants – some results

- Threshold: 0.94 (100% Precision in test set)
- NewsExplorer: currently, 608 new names every day
  - 83 are automatically merged (14%)
  - Some are additionally saved for expert judgment

<table>
<thead>
<tr>
<th>Name 1</th>
<th>Name 2</th>
<th>Similarity</th>
<th>Merged?</th>
<th>Same person?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barzan al-Tikriti</td>
<td>Barzan al Tikriti</td>
<td>0.99</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ismail Hanieh</td>
<td>Ismail Hanyieh</td>
<td>0.98</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Farouq al-Qaddoumi</td>
<td>Farouk al-Kadoumi</td>
<td>0.97</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Abdullah bin Abdul Aziz</td>
<td>Abdullah bin Abdel Aziz</td>
<td>0.96</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Barzan al-Tikriti</td>
<td>Barazan al-Takriti</td>
<td>0.94</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Manfred Wörner</td>
<td>Manfred Werner</td>
<td>0.93</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Michel Ancel</td>
<td>Michael Ancel</td>
<td>0.92</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Jorge Costa</td>
<td>Jorge Acosta</td>
<td>0.92</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Falun Gong</td>
<td>Falun Gong</td>
<td>0.90</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Roberto Panella</td>
<td>Roberto Pianelli</td>
<td>0.87</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Peter Struck</td>
<td>Peter Starck</td>
<td>0.82</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Jamie Foxx</td>
<td>Jaime Foxx</td>
<td>0.77</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Pervez Musharraf

Information about this person was last updated on Friday, November 23, 2007.

Names
- Pervez Musharraf (Eu, sv)
- General Pervez Musharraf (da, sv)
- Gen Musharraf (en)
- Pervez Mušaraf (sl)
- Gen Pervez Musharraf (en)
- Pervez Musharraf (da, sv)
- Pervez Musharraf (fr)
- Pervez Musharraf (fr)
- Pervez Musharraf (ar)
- Pervez Musharraf (sl)
- Pervez Musharraf (de)
- Perves Musharraf (de)

Key Titles and Phrases
- pakistan president (en - 827)
- president (de, sv - 3230)
- president gen (en - 506)
- président pakistanais (fr - 398)
- pakistaanse president (nl - 235)
- presidente pakistán (es - 147)
- president (de - 617)
- general (en, sv - 450)
- president (da - 210)
- president (es, pt - 589)
- president, gen (en - 62)

External resources

Latest Clusters - English
[pakistán, president, Musharraf, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, president, Pakistan, presen

Quotes from - English
- [has said]: Where it fails to live up to those values, it needs to act with credibility and consistency. I think Pakistan is a test in that respect.
- [has repeated]: The (presidential) oath can be taken ... by the weekend or immediately thereafter.

Quotes about - English
- [said]: And look, a lot of that was done by Pervez Musharraf himself. And so for him at this point to help put his country back on the road to democratic reform is important. We're looking for him to take off his uniform, expressindia 22-NOV-07
- [said]: It's disgraceful that Musharraf is punishing Chief Justice Chie... judge's family under house arrest, HumanRightsWatch 22-NOV-07
- [said]: Rather than making... HumanRightsWatch 22-NOV-07
- [said]: He (Musharraf) has assured... HumanRightsWatch 22-NOV-07

Related People
- Benazir Bhutto (910)
- Nawaz Sharif (552)
- George W. Bush (337)
- Iftikhar Muhammad Chaudhry (502)
- Osama bin Laden (430)
- Shaukat Aziz (395)
- Tariq Azeem (251)
- Ha...

Related Stories
- Pakistan court dismisses final Musharraf challenge... euronews 22-NOV-07
- Struggle to help cyclone survivors... cnn 21-NOV-07

Quotes
- [said]: Where it fails to live up to those values, it needs to act with credibility and consistency. I think Pakistan is a test in that respect. expressindia 22-NOV-07
- [said]: It's disgraceful that Musharraf is punishing Chief Justice Chie... judge's family under house arrest, HumanRightsWatch 22-NOV-07
- [said]: Rather than making... HumanRightsWatch 22-NOV-07
- [said]: He (Musharraf) has assured... HumanRightsWatch 22-NOV-07
Social network extraction from multilingual news
Social networks and relationship extraction

1. Co-occurrence networks (long-term / daily)

2. Quotation networks
   http://langtech.jrc.it/picNews.html

3. Relationship network
   http://emm-labs.jrc.it/LiveNews/Nets/Social.html
\[ w_{e_1,e_2} = \ln(C_{e_1,e_2}) \times \frac{2 \cdot C_{e_1,e_2}}{(C_{e_1} + C_{e_2})} \times \frac{1}{1 + \ln(A_{e_1} \cdot A_{e_2})} \]

- Number of clusters they appear in together
- Weighting depending on the total number of clusters they appear in
- Weighting depending on the total number of persons they are associated with
Daily co-occurrence networks (reset every midnight)

http://langtech.jrc.it/entities/socNet/last.html
Daily mixed-language co-occurrence networks (~100,000 links/day)

Daily co-occurrence networks: Edward Kennedy – Barack Obama

499 articles in the news since midnight talking about those two persons.


Edward Kennedy ... Barack Obama

RSS for two Entities Wed, 26 Aug 2009 14:13:59 +0200

Barack Obama Edward Kennedy

2009-08-26T14:11+0200 Murió el senador Edward Kennedy [clarin - es]
... El senador estadounidense [Edward M. Kennedy] ..., miembro de una de las familias más rele (…) dor de nuestro tiempo", dijo el presidente [Barack Obama] en un comunicado. El presidente se mostró & ...

2009-08-26T14:11+0200 Barack Obama Edward Kennedy [bbc-arabic - ar]
... أوباما، وفاة كينيدي نهاية حفل عام في السياسة الأمريكية ...

2009-08-26T14:10+0200 National Review Online: Their Lion, Our Bane [NPRnews - en]
... the death either. This is his remembrance of [Ted Kennedy]. ... There is a lot one could say (…) are reform and also the campaign of Senator [Barack Obama]. There are the personal failings and traged ...

2009-08-26T14:06+0200 Barack Obama Ted Kennedy meros 77; conheça trajetória do irmão de John Kennedy [folha - pt]
... os 77 anos, o senador democrata americano [Edward M. Kennedy], ... vítima de um câncer de cérebro, informo (…) ào democrata que confirmou a candidatura de [Barack Obama] à Presidência. Mesmo sofrendo de dores caus ...

2009-08-26T14:04+0200 Senaattori Kennedy muistetaan liberaalina leijonana [keskisuomalainen - fi]
... u ympärä maailmaa. Yhdysvaltain presidentti [Barack Obama] kuvisi oloaan loiduttomaksi. – Viden vuos (…) änella diagnosoitiin aivosyöpä viime vuonna. [Ted Kennedy] ... varittiin senaatirin vuonna 1962 veljensä J ...

2009-08-26T14:01+0200 Potere assoluto [La giornata] [lifoglio - it]
... ni fa a Chappaquiddick la tragedia che segnò [Ted Kennedy] ... L’arroganza della strife più amata tolse al (…) ensi durante il pranzo di festeggiamiento di [Barack Obama], dopo il giuramento presidenziale del 20 ge ...

2009-08-26T14:00+0200 Barack Obama Ted Kennedy hylles av Obama [hegnar - no]
... bort tirsdag. Slik hylles han av psident Barack Obama. Senator [Edward Kennedy] ..., en av de mest sentrale og karismatiske se ...

2009-08-26T13:59+0200 Senator Edward Kennedy dies at age 77 [reuters - en]
... BOSTON (Reuters) - U.S. Senator [Edward Kennedy], a towering figure in the Democratic Party (…) Democrats as they seek to answer President [Barack Obama]'s call for an overhaul of the healthcare sy ...

2009-08-26T13:57+0200 09:00 ABD'nin en eski senatörü Edward Kennedy öldü [netgazete - tr]
... tor ve Demokrat Partinin en son önemli şahsisi [Edward Kennedy], ... , 77 yaşında hayatını kaybetti. CNN, Amerik (…) nde de Demokrat Partinin aday adaylarından [Barack Obama] yi desteklediğiğini açıkladı. Kennedy'nin des ...

2009-08-26T13:57+0200 Edward „Ted“ Kennedy - Der Tod einer Legende [bildt-online - de]
... 0 Senator [Edward Kennedy], ... politisches Schwergewicht der USA und jün (…) teilt. Bei der Amtseinführung von Präsident [Barack Obama] (48) im vergangenen Januar brach Edward Ken ...
Chief Superintendent Stewart Gull said: “Wright, from Ipswich, has been charged with the murder of all five women.”

"I am against Turkey's full membership, and I have also told this to Turkish Prime Minister Erdogan," Sarkozy said, during a meeting with Angela Merkel.

Sarkozy, "Türkiye'nin AB'ye tam üye olmasını karşıyım. Bunu Başbakan Erdoğan'a da söyledim" dedi.

Gull → Wright (1)
Sarkozy → Erdogan (2)
Quotation network: http://langtech.jrc.it/entities/socNet/quotes.html
“He (Zuma) gave his views on regional issues from Zimbabwe, Somalia and Sudan. His advice on these issues was extremely helpful,” Clinton said a media briefing.
We have sold Ronaldo, Liverpool have sold Alonso, Arsenal have sold Adebayor. It has been really quiet in terms of the big four.
Quotation network: another sub-network

> Jaime Alguersuari, den mit 19

It's good for the sport but it

> Michael Schumacher, septuagé

Realmente no sé qué ocurrirá en las próximas siete carreras. Será muy, muy duro alcanzar a Button. Ya cuenta con demasiados puntos sobre mí. Incluso aunque consiga el máximo de puntos en las últimas siete carreras, no seré capaz de superarle", comentó Alguersuari, por lo que Schumacher

> In agreement with Stefano Domenicali

Am fost nevoit să îl informez luni scâră pe pregătirea scurterii Ferrari, Luca di Montezemolo, și pe directorul Stefano Domenicali că nu îl voi putea înlocui pe Felipe Massa în Marele Premiu al Europei. Am încercat totul pentru a putea reveni în circuit, dar, din păcate, nu am reușit. Nu am putut să scap de durerile pe care le am la gât, deși, din punct de vedere medical am făcut totul", a spus Schumacher.

> Luca di Montezemolo

Esa historia de Schumacher...

Don't think he (Schumacher)

Es wäre eine Ehre etwas für dich, Willi Weber

> Willi Weber

What would we expect from Schu

> Dan Williams

While we wait for Michael Schumacher
Network of specific relationships (English only)

http://emm-labs.jrc.it/LiveNews/Nets/Social.html


Cross-lingual Cluster Linking

Castro quits as president, state-run paper reports [72] de es fr it nl ar bg da et fa no pl pt ro ru sv tr

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.
cnn 9:23:00 AM CET
• How to find out whether two texts in different languages are related?

• Most common approach: use MT or bilingual dictionaries to translate into English, then use monolingual methods to calculate similarity.

  • **Using MT** (e.g. Leek et al. 1999 for Chinese-Mandarin to English); 
  ~ 50% performance loss when using MT;

  • **Using bilingual dictionaries** (e.g. Wactlar 1999 for Serbo-Croatian to English; Urizar & Loinaz for Basque, Spanish and English 2007)

  • In TDT 1999, the better results were achieved using MT
Automatically produce bilingual lexical space for bilingual document representation and document similarity calculation, e.g.

- Bilingual *Lexical Semantic Analysis (LSA)* (Landauer & Littman 1991)
- *Kernel Canonical Correlation Analysis (KCCA)* (Vinokourov et al., 2002)

+ Achieved results are relatively good
- Bilingual approach is restricted to a few languages (OK for English as target lang.):

  \[
  \text{Language pairs} = N \times (N - 1) / 2
  \]

  \(N = \text{number of languages}\)

- NewsExplorer: 19 languages → 171 language pairs (342 language pair directions)!
• Alternative: use entities and thesauri as anchors:
  • Names of persons and organisations
  • Names of locations
  • Nodes from a multilingual thesaurus (e.g. Eurovoc)
  • Dates
  • Terms from multilingual specialist dictionaries (MeSH for medicine, etc.)
  • …

• Normalise these expressions

→ Use as kind of an interlingua; no language pair-specific resource needed

Language-independent features for multilingual document representation
No MT or bilingual dictionaries
19 languages

\[
\text{CLDS} = \alpha \cdot S_1 + \beta \cdot S_2 + \gamma \cdot S_3 + \delta \cdot S_4
\]

Northwest and Delta become the latest US airlines for file for Chapter 11 bankruptcy protection.

Deux compagnies aériennes américaines se mettent en faillite pour survivre.
Multilingual Geo-tagging

- Lookup of place names from a gazetteer
- **Challenges:**
  - Multilingual gazetteer: combination of multiple sources
  - Inflection similar to lookup of known persons
  - Homography
    - places-places
    - places-persons
    - places-words
- Usage of various heuristics for disambiguation
**Eurovoc subject domain categorisation**

![Eurovoc Thesaurus](image)

<table>
<thead>
<tr>
<th>Description (ID)</th>
<th>Description (en)</th>
<th>Resource metadata share (THES)*2</th>
<th>Coverage</th>
<th>Rank</th>
<th>Values</th>
<th>Values</th>
<th>Values</th>
<th>Rank</th>
<th>Prec. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1641040706000000</td>
<td>FISHING INDUSTRY</td>
<td>0.200</td>
<td>28.416</td>
<td>10</td>
<td>20</td>
<td>32</td>
<td>10</td>
<td>30</td>
<td>1.90</td>
</tr>
<tr>
<td>1641040706000000</td>
<td>FISHING INDUSTRY</td>
<td>0.050</td>
<td>1.216</td>
<td>20</td>
<td>6</td>
<td>10</td>
<td>30</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td>1641040706000000</td>
<td>FISHING INDUSTRY</td>
<td>0.050</td>
<td>1.216</td>
<td>20</td>
<td>6</td>
<td>10</td>
<td>30</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td>1641040706000000</td>
<td>EUROVOC</td>
<td>0.170</td>
<td>4.0400</td>
<td>21</td>
<td>15</td>
<td>37</td>
<td>1.90</td>
<td>1.90</td>
<td></td>
</tr>
</tbody>
</table>
Multi-label indexing with Eurovoc

- **Eurovoc Thesaurus**
  - > 6,000 subject domains
  - Exists in one-to-one translations in all official EU languages, and more
  - Used by many European parliaments for manual classification → use for training classifiers

- **Challenges**:
  - Concepts rather than words, e.g. PROTECTION OF MINORITIES, CONSTRUCTION AND TOWN PLANNING
  - Large number of classes (> 6000)
  - Very unevenly distributed
  - Various text types (heterogeneous training set)
  - Multi-label categorisation (both for training and assignment)

→ **Profile-based category ranking** task (Supervised Learning)
  - Training: Identification of most significant words for each class
  - Assignment: combination of measures to calculate similarity between profiles and new document

- **Result**: Long, weighted list of (numerical) subject domain identifiers → language-independent

---

**Title:** Legislative resolution embodying Parliament's opinion on the proposal for a Council Regulation amending Regulation No 2847/93 establishing a control system applicable to the common fisheries policy (COM(95)0256 - C4-0272/95 - 95/0146(CNS)) (Consultation procedure)

<table>
<thead>
<tr>
<th>Descriptor ID</th>
<th>Descriptor text</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>5641040706000000</td>
<td>FISHING CONTROLS [G]</td>
<td>0.360</td>
</tr>
<tr>
<td>5641020000000000</td>
<td>FISHING GROUNDS [nt]</td>
<td>0.308</td>
</tr>
<tr>
<td>5641040200000000</td>
<td>COMMON FISHERIES POLICY [G]</td>
<td>0.280</td>
</tr>
<tr>
<td>5641040100000000</td>
<td>FISHERY MANAGEMENT [nt]</td>
<td>0.279</td>
</tr>
<tr>
<td>5641040700000000</td>
<td>FISHING REGULATIONS [G]</td>
<td>0.270</td>
</tr>
<tr>
<td>5641040704000000</td>
<td>FISHING PERMIT [G]</td>
<td>0.261</td>
</tr>
<tr>
<td>5641040101000000</td>
<td>CONSERVATION OF FISH STOCKS [S]</td>
<td>0.253</td>
</tr>
<tr>
<td>5641040600000000</td>
<td>FISHING AREA [G]</td>
<td>0.252</td>
</tr>
<tr>
<td>5206040100000000</td>
<td>CONSERVATION OF RESOURCES [S]</td>
<td>0.251</td>
</tr>
<tr>
<td>5641050000000000</td>
<td>FISHERY RESOURCES</td>
<td>0.232</td>
</tr>
<tr>
<td>5641040800000000</td>
<td>CATCH OF FISH</td>
<td>0.213</td>
</tr>
<tr>
<td>5641040000000000</td>
<td>FISHERIES POLICY</td>
<td>0.203</td>
</tr>
<tr>
<td>5641040705000000</td>
<td>FISHING LICENCE</td>
<td>0.181</td>
</tr>
<tr>
<td>5641060100000000</td>
<td>FISHING FLEET</td>
<td>0.179</td>
</tr>
<tr>
<td>5641010000000000</td>
<td>FISHING INDUSTRY</td>
<td>0.176</td>
</tr>
<tr>
<td>5641040201000000</td>
<td>EUROPECHE</td>
<td>0.176</td>
</tr>
</tbody>
</table>
Two freely available multilingual resources (22 languages)

• DGT-TM translation memory
  • For all 231 language pairs (462 language pair directions)
  • Made available by DGT, through JRC’s website.

• JRC-Acquis multilingual parallel corpus (v 3.0)
  • Made available by JRC
  • Sentence-aligned for all 231 language pairs using two different aligners
  • Over 1 Billion words
  • Manually subject domain-classified using the Eurovoc thesaurus.

• **Freely available** for research purposes on our web site: [http://langtech.jrc.it/](http://langtech.jrc.it/) (go to ‘Resources’)

Language-independent features for multilingual document representation
No MT or bilingual dictionaries
19 languages

\[ \text{CLDS} = \alpha \cdot S_1 + \beta \cdot S_2 + \gamma \cdot S_3 + \delta \cdot S_4 \]

Sim1 (40%): Multilingual Eurovoc subject domains

Sim2 (30%): Geo-locations

Sim3 (20%): Names + variants

Sim4 (10%): Cognates and numbers (without country score)
Agenda

• Media Monitoring and multilinguality
• Europe Media Monitor (EMM) applications - Functionality
  • Publicly accessible at http://press.jrc.it/overview.html

• Some language technology components in detail
  • Multilingual person name recognition
  • Name variant matching across many languages
  • Social networks based on multilingual information extraction
  • Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc

• Conclusion and Ongoing work
Conclusion

Ongoing work
Summary

• Europe Media Monitor (EMM) applications - Functionality

• Some language technology components in detail
  • Multilingual person name recognition
  • Name variant matching across many languages
  • Social networks based on multilingual information extraction
  • Cross-lingual cluster linking

➔ 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: More information about relations between people

Social networks produced on the basis of many languages is less biased.
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</td>
<td>russo (it, pt - 349)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko</td>
<td>agent russe (fr - 134)</td>
<td></td>
</tr>
<tr>
<td>Alexandre Litvinenko</td>
<td>ruso (es - 208)</td>
<td></td>
</tr>
<tr>
<td>Aleksandr Litvinenko</td>
<td>agenten (de, sv - 134)</td>
<td></td>
</tr>
<tr>
<td>Aleksander Litvinenko</td>
<td>kritikers (de - 79)</td>
<td></td>
</tr>
<tr>
<td>Aleksandra Litvinenko</td>
<td>agent (en, sv - 130)</td>
<td></td>
</tr>
<tr>
<td>Aleksandr Litvinenko</td>
<td>russe (de, fr - 73)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko</td>
<td>former russian agent (en - 20)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko</td>
<td>morte di (it - 45)</td>
<td></td>
</tr>
<tr>
<td>Alexandre Litvinenko</td>
<td>ryske agenten (sv - 13)</td>
<td></td>
</tr>
<tr>
<td>Alexandre Litvinenko</td>
<td>kritiker (de - 19)</td>
<td></td>
</tr>
<tr>
<td>Aleksandr Litvinenko</td>
<td>43 ans (fr - 17)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenko</td>
<td>russi (it - 14)</td>
<td></td>
</tr>
<tr>
<td>Aleksandar Litvinenko</td>
<td>russian (en - 15)</td>
<td></td>
</tr>
<tr>
<td>Alexander Litvinenk</td>
<td>omicidio di (it - 11)</td>
<td></td>
</tr>
<tr>
<td>Alexander Walterowitsch Litvinenko (de)</td>
<td>officer (en - 13)</td>
<td></td>
</tr>
<tr>
<td>Alexander Walterowitsch Litvinenko (de)</td>
<td>former (en - 16)</td>
<td></td>
</tr>
</tbody>
</table>

Image obtained automatically from Wikipedia

Read Wikipedia entry

---

Ongoing work

• Ongoing activities:
  • Further integration of the four applications
  • Blog monitoring

• Current research:
  • Sentiment analysis for the news
  • Multilingual multi-document summarisation

• Challenge: Using simple methods that do not require many language-specific resources