Linking News Content Across Languages

Ralf Steinberger
European Commission – Joint Research Centre
21027 Ispra (VA), Italy
http://langtech.jrc.it/ – http://press.jrc.it/overview.html
Ralf.Steinberger@jrc.it

1 Introduction

Organisations and individuals that need to monitor what the media say about certain issues face an extreme information overload, especially if they are interested in the news written in more than one language. News aggregators sometimes pre-filter potentially user-relevant articles or automatically group related articles into clusters. However, the enormous amount of available online information calls for further automatic information processing to enable users to sieve through even larger amounts of textual data in less time and to navigate and explore the document collections efficiently.

2 NewsExplorer

NewsExplorer is a freely available news analysis system that offers such functionality in 19 languages. NewsExplorer integrates various text analysis applications including clustering, multi-label document classification, named entity recognition, name variant matching across languages and writing systems, topic detection and tracking, and more. The purpose of this presentation is to present this news exploration and analysis system and to especially address the multilinguality issue and the cross-lingual functionality of the application. References to prior art will be made, where appropriate.

3 News Data and the EMM family of applications

NewsExplorer is part of the Europe Media Monitor (EMM) family of applications (http://press.jrc.it/overview.html). EMM gathers a daily average of 80,000 news articles from about 2,200 web news sources in 43 languages. *NewsBrief* and the Medical Information System *MedISys* classify the news, cluster related articles and alert users of breaking news when unexpected spikes are detected. *EMM-Labs* gives access to data visualisation tools and to the results of a collection of advanced text processing tools such as relation extraction, event scenario template filling, and various types of social networks. The freely available EMM online applications attract between one and two Million hits per day.

References

Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2008). *Using language-independent rules to achieve high multilinguality in text mining*. In: Françoise Fogelman-Soulié, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds): Mining Massive Data Sets for Security. IOS-Press, Amsterdam, Holland.

Pouliquen Bruno & Ralf Steinberger (2009). *Automatic Construction of Multilingual Name Dictionaries*. In: Cyril Goutte, Nicola Cancedda, Marc Dymetman & George Foster (eds.): Learning Machine Translation. MIT Press, NIPS series.

Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). *Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus*. In: Proceedings of the EUROLAN Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities. Bucharest, Romania.

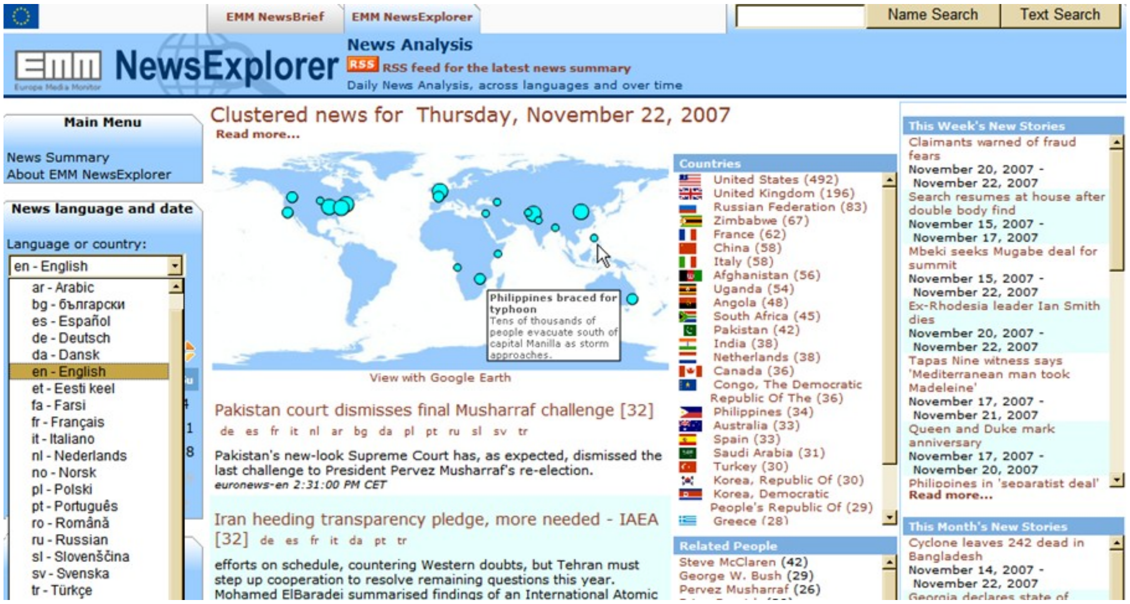
Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). *Geocoding multilingual texts: Recognition, Disambiguation and Visualisation*. Proceedings of the 5th International Conference on Language Resources and Evaluation LREC, pp. 53-58. Genoa, Italy.

Pouliquen Bruno, Ralf Steinberger & Clive Best (2007). *Automatic Detection of Quotations in Multilingual News*. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP. Borovets, Bulgaria.

Pouliquen Bruno, Olivier Deguernel & Ralf Steinberger (2008). *Story tracking: linking similar news over time and across languages*. In: Proceedings of the CoLing'2008 workshop: Multi-source, multilingual information extraction and summarization, Manchester, August 2008.

Acknowledgements

NewsExplorer and the other EMM media monitoring applications are complex applications, which were built in a group effort. I would thus like to thank the whole EMM team for their good work and especially Bruno Pouliquen, who was the main driving force behind the development of NewsExplorer.



Clustered news for Thursday, November 22, 2007

Read more...

Philippines braced for typhoon
Tens of thousands of people evacuate south of capital Manilla as storm approaches.

Pakistan court dismisses final Musharraf challenge [32]
de es fr it nl ar bg da pl pt ru sl sv tr

Pakistan's new-look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf's re-election.
euronews-en 2:31:00 PM CET

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

efforts on schedule, countering Western doubts, but Tehran must step up cooperation to resolve remaining questions this year. Mohamed ElBaradei summarised findings of an International Atomic

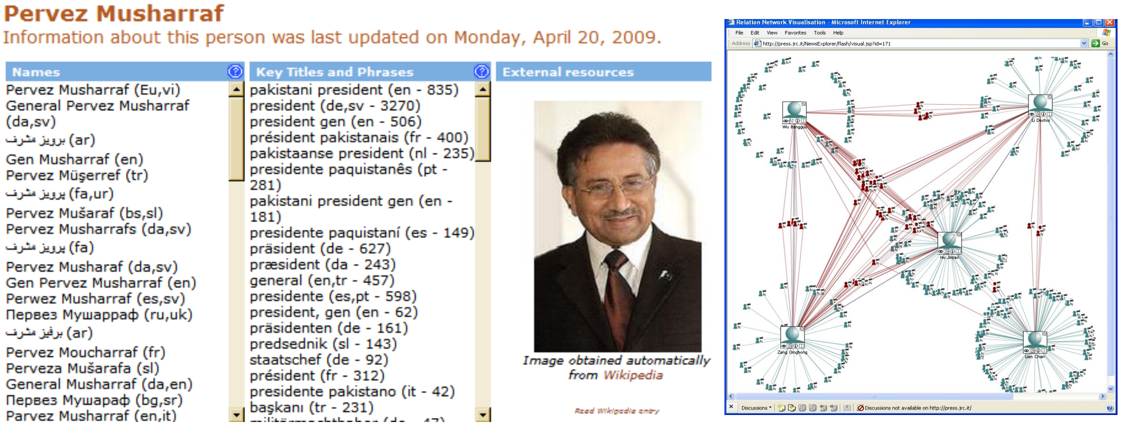
Countries
United States (492)
United Kingdom (196)
Russian Federation (83)
Zimbabwe (67)
France (62)
China (58)
Italy (58)
Afghanistan (56)
Uganda (54)
Angola (48)
South Africa (45)
Pakistan (42)
India (38)
Netherlands (38)
Canada (36)
Congo, The Democratic Republic Of The (36)
Philippines (34)
Australia (33)
Spain (33)
Saudi Arabia (31)
Turkey (30)
Korea, Republic Of (30)
Korea, Democratic People's Republic Of (29)
Greece (28)

Related People
Steve McClaren (42)
George W. Bush (29)
Pervez Musharraf (26)

This Week's New Stories
Claimants warned of fraud fears
November 20, 2007 - November 22, 2007
Search resumes at house after double body find
November 15, 2007 - November 17, 2007
Mbeki seeks Mugabe deal for summit
November 15, 2007 - November 22, 2007
Ex-Rhodesia leader Ian Smith dies
November 20, 2007 - November 22, 2007
Tapas Nine witness says 'Mediterranean man took Madeleine'
November 17, 2007 - November 21, 2007
Queen and Duke mark anniversary
November 17, 2007 - November 20, 2007
Philippines in 'separatist deal'
Read more...

This Month's New Stories
Cyclone leaves 242 dead in Bangladesh
November 14, 2007 - November 22, 2007
Georgia declares state of

Figure 1. Screenshot of NewsExplorer, showing a map with the location of today's news, the largest English language news clusters, links to related news in the other 18 languages, lists of countries, people and organizations mentioned in the news that day, and lists of the biggest 'stories' (daily news clusters linked over time) this week, month and year.



Pervez Musharraf

Information about this person was last updated on Monday, April 20, 2009.

Names
Pervez Musharraf (Eu,vi)
General Pervez Musharraf (da,sv)
برويز مشرف (ar)
Gen Musharraf (en)
Pervez Müşerref (tr)
پرويز مشرف (fa,ur)
Pervez Mušaraf (bs,sl)
Pervez Musharrafs (da,sv)
پرويز مشرف (fa)
Pervez Musharaf (da,sv)
Gen Pervez Musharraf (en)
Perwez Musharraf (es,sv)
Первез Мушарраф (ru,uk)
برفيز مشرف (ar)
Pervez Moucharraf (fr)
Perveza Mušarafa (sl)
General Musharraf (da,en)
Первез Мушараф (bg,sr)
Parvez Musharraf (en,it)

Key Titles and Phrases
pakistani president (en - 835)
president (de,sv - 3270)
president gen (en - 506)
président pakistanais (fr - 400)
pakistaanse president (nl - 235)
presidente paquistanês (pt - 281)
pakistani president gen (en - 181)
presidente paquistaní (es - 149)
präsident (de - 627)
præsident (da - 243)
general (en,tr - 457)
presidente (es,pt - 598)
president, gen (en - 62)
präsidenten (de - 161)
predsednik (sl - 143)
staatschef (de - 92)
président (fr - 312)
presidente pakistano (it - 42)
başkanı (tr - 231)

Image obtained automatically from Wikipedia

Figure 2. Screenshot of (part of) NewsExplorer's page on Pakistani President Pervez Musharraf. That page shows automatically collected name variants (including variants in different scripts such as Arabic, Farsi, and Russian), titles, lists of related persons, lists of related news clusters, quotations by and about Musharraf, multi-day 'stories' in which he is mentioned, and more. Relations between two or more persons can be visualised graphically.