

Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.)

Multilingual Resources and Multilingual Applications

Proceedings of the Conference of the
German Society for Computational Linguistics and
Language Technology (GSCL) 2011



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Sonderforschungsbereich
Mehrsprachigkeit



Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.)

Multilingual Resources and Multilingual Applications

Proceedings of the Conference of the German Society for
Computational Linguistics and Language Technology (GSCL) 2011

© *Hanna Hedeland, Thomas Schmidt, Kai Wörner*
Hamburger Zentrum für Sprachkorpora
Max Brauer-Allee 60
D-22765 Hamburg

Die „Arbeiten zur Mehrsprachigkeit – Folge B“ publizieren Forschungsarbeiten aus dem Sonderforschungsbereich 538 *Mehrsprachigkeit*, der von der Deutschen Forschungsgemeinschaft im Juli 1999 an der Universität Hamburg eingerichtet wurde. Wir danken der DFG für ihre Unterstützung.

Die „Arbeiten zur Mehrsprachigkeit – Folge B“ sind bei der Deutschen Bibliothek in Frankfurt/Main mit der Seriennummer ISSN 0176-559X eingetragen.

Redaktion:

Martin Elsig, Svenja Kranich, Thomas Schmidt, Manuela Schönenberger

Technische Umsetzung:

Thomas Schmidt

Combining various text analysis tools for multilingual media monitoring

Ralf Steinberger

European Commission – Joint Research Centre (JRC)

21027 Ispra (VA), Italy

E-mail: Ralf.Steinberger@jrc.ec.europa.eu, URL: <http://langtech.jrc.ec.europa.eu/>

Abstract

There is ample evidence that information contained in media reports is complementary across countries and languages. This holds both for facts and for opinions. Monitoring multilingual and multinational media therefore gives a more complete picture of the world than monitoring the media of only one language, even if it is a world language like English. Wide coverage and highly multilingual text processing is thus important. The JRC-developed *Europe Media Monitor* (EMM) family of applications gathers about 100,000 media reports per day in 50 languages from the internet, groups related articles, classifies them, detects and follows trends, produces statistics and issues automatic alerts. For a subset of 20 languages, it also extracts and disambiguates entities (persons, organisations and locations) and reported speech, links related news over time and across languages, gathers historical information about entities and produces various types of social networks. More recent R&D efforts focus on event scenario template filling, opinion mining, multi-document summarisation, and machine translation. This extended abstract gives an overview of EMM from a functionality point of view rather than providing technical detail.

Keywords: news analysis; multilingual; automatic alerting; text mining; information extraction.

1. EMM: Background and Objectives

The JRC with its 2700 employees working in five different European locations in a wide variety of scientific-technical fields is a Directorate General of the European Commission (EC). It is thus a governmental body free of national interests and without commercial objectives. Its main mandate is to provide scientific advice and technical know-how to European Union (EU) institutions and its international partners, as well as to EU member state organisations, with the purpose of supporting a wide range of EU policies. Lowering the language barrier in order to increase European integration and competitiveness is a declared EU objective.

The JRC-developed *Europe Media Monitor* (EMM) is a publicly accessible family of four news gathering and analysis applications consisting of *NewsBrief*, the *Medical Information System MedISys*, *NewsExplorer* and *EMM-Labs*. They are accessible via the single URL <http://emm.newsbrief.eu/overview.html>. The first EMM website went online in 2002 and it has since been extended and improved continuously. The initial objective was to complement the manual news clipping

services of the EC, by searching for news reports online, categorising them according to user needs, and providing an interface for human moderation (selection and re-organisation of articles; creation of layout to print in-house newspapers). EMM users thus typically have a specific information need and want to be informed about any media reports concerning their subject of interest. Monitoring the media for events that are dangerous to the public health (PH) is a typical example. EMM thus continuously gathers news from the web, automatically selects PH-related news items (e.g. on chemical, biological, radiological and nuclear (CBRN) threats including disease outbreaks, natural disasters and more), presents the information on targeted web pages, detects unexpected information spikes and alerts users about them. In addition to PH, EMM categories cover a very wide range of further subject areas, including the environment, politics, finance, security, various scientific and policy areas, general information on all countries of the globe, etc. For an overview of EMM, see Steinberger et al. (2009).

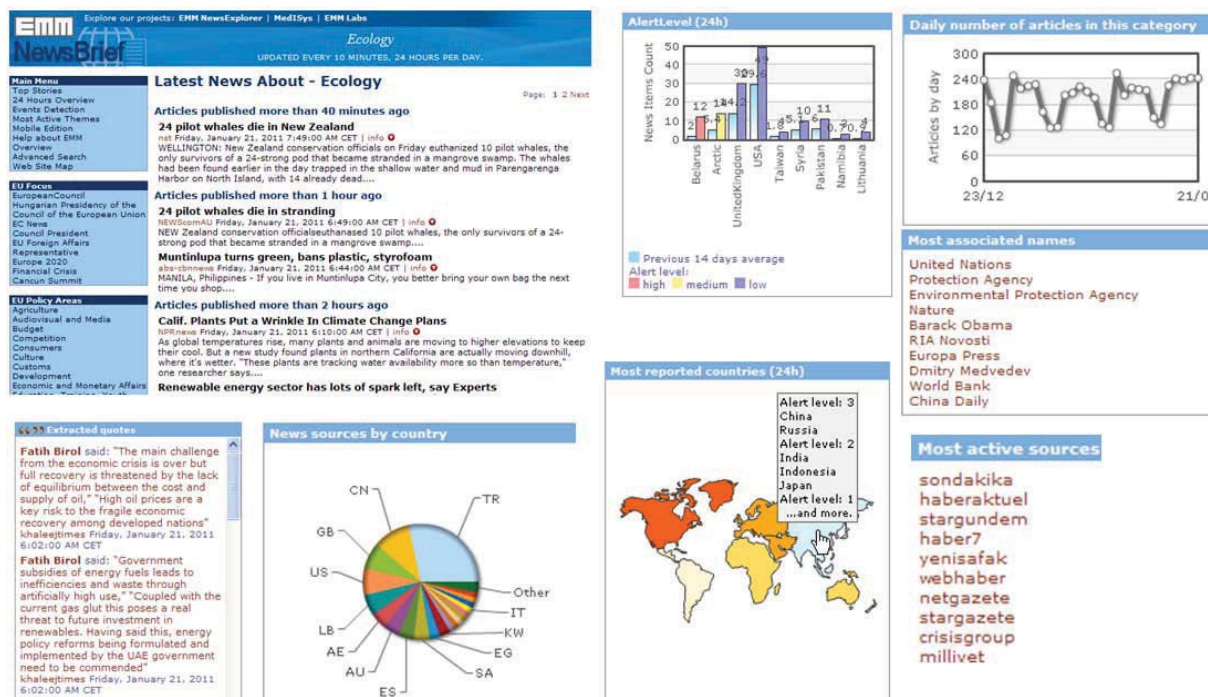


Figure 1. Various aggregated statistics and graphs showing category-based information for one category (ECOLOGY) derived from reports in multiple languages.

2. Information complementarity across languages and countries; news bias

While national EMM clients are mostly interested in the news of their own country and that of surrounding countries (e.g. for disease outbreak monitoring), they also need to follow mass gatherings (e.g. for religious, sport-related or political reasons) because participants may bring back diseases. In addition to the news in the 23 official EU languages, EMM thus also monitors news in Arabic, Chinese, Croatian, Farsi, Swahili, etc., to mention just a few of the altogether 50 languages. While major events such as wars or natural disasters are usually well-covered in world languages such as English, French and Spanish, many small events are typically only mentioned in the national or even in regional press. For instance, disease outbreaks, small-scale violent events and accidents, fraud cases, etc. are usually not reported outside the national borders. The study by Piskorski et al. (2011) comparing media reports in six languages showed that only 51 out of 523 events (of the event types violence, natural disasters and man-made disasters) were reported in more than one language. 350 out of the 523 events were found in non-English news.

Due to this information complementarity across languages and countries, it is crucial that monitoring systems like EMM process texts in many different languages. Using Machine Translation (MT) into one language (usually English) and filtering the news in that language is only a partial solution because specialist terms and names are often badly translated. The benefits of processing texts in the original language was also formulated by Larkey et al. (2004) in their *native language hypothesis*.

We observed the following benefits of applying multilingual text mining tools:

- 1) Different languages cover different geographical areas of the world, for specific subject areas as well as generally. EMM-NewsBrief's news clouds (see <http://emm.newsbrief.eu/geo?type=cluster&format=html&language=all>) show this clearly.
- 2) More information on entities (persons and organisations; see NewsExplorer entity pages) can be extracted from multilingual text. This is due to different contents found, but also to varying linguistic coverage of the text mining software.
- 3) Many more named entity variant spellings (including across scripts) are found when analysing different languages (see NewsExplorer entity pages). These

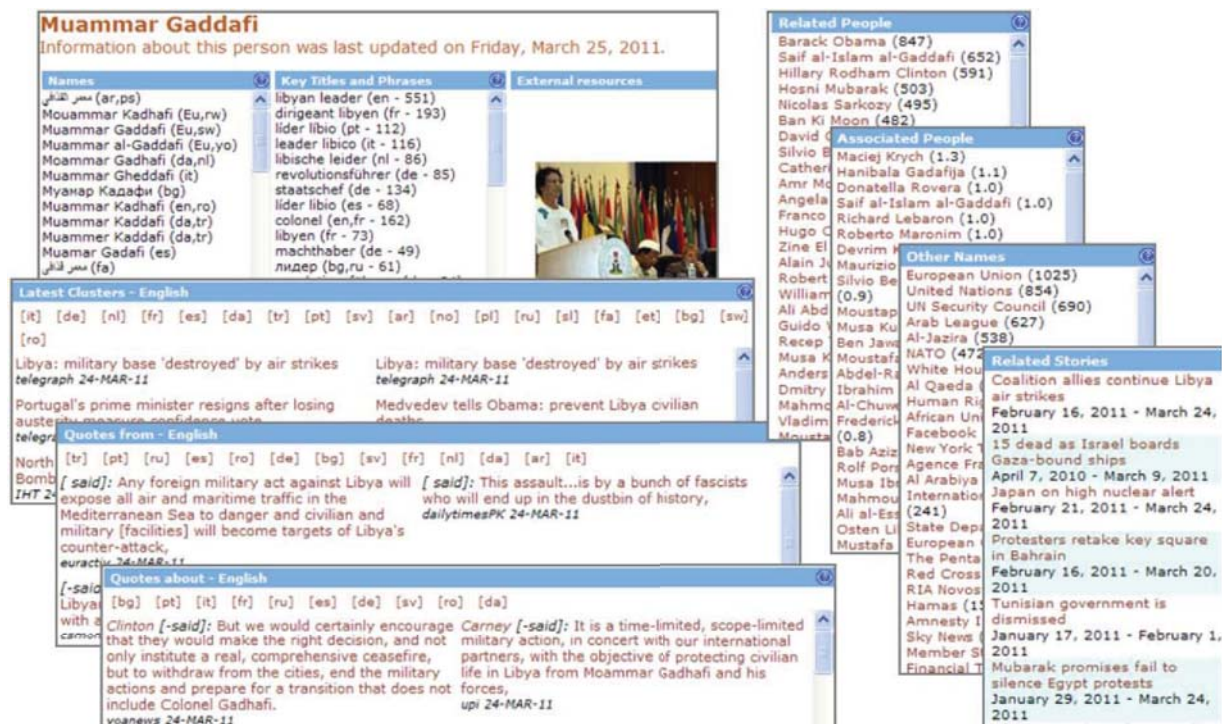


Figure 4. Information automatically gathered over time by EMM-NewsExplorer from media reports in twenty or more languages on one named entity.

The software additionally tracks related news over time, produces time lines and displays extracted meta-information about the news event. For details about the linking of related news items across languages and over time, see Pouliquen et al. (2008).

3.3. Multilingual information gathering on named entities

EMM-NewsExplorer identifies references to person and organisation names in twenty languages. It automatically identifies whether newly found names (within the same script or across different scripts) are simply spelling variants of another name or whether they are new names (for details, see Pouliquen & Steinberger, 2009). The EMM database currently contains up to 400 different automatically collected spellings for the same entity. Any EMM application making use of named entity information uses unique entity identifiers instead of concrete name spellings, allowing to merge information across documents, languages and scripts. The EMM software furthermore keeps track of titles and other expressions found next to the name, keeps statistics on where and when the names were found, and which entities get frequently mentioned together. The latter

information is used to generate social networks that are derived from the international media, thus being independent of national viewpoints. EMM software also detects quotations by and about each entity. The accumulated multilingual results are displayed on the NewsExplorer entity pages (see Figure 4), through which users can explore entities, their relations and related news. Click on any entity name in any of the EMM applications to explore this application.

3.4. Multilingual event scenario template filling

For a smaller subset of currently seven languages, the EMM-NEXUS software extracts structured descriptions of events relevant for global crisis monitoring, such as natural disasters; accidents; violent, medical and humanitarian events, etc. (Tanev et al., 2009; Piskorski et al., 2011). For each news cluster about any such event, the software detects the event type; the event location; the count of dead, wounded, displaced, arrested etc. persons; the perpetrator in the event, as well as the weapons used, if applicable. Contradictory information found in different news articles (such as differing victim counts) are resolved to produce a best guess. The aggregated



Figure 5. EMM-Labs geographical visualisation of events extracted from media reports in seven languages.

event information is then displayed on *NewsBrief* (in text form) and on *EMM-Labs* (in the form of a geographic map¹; see **Figure 5**).

4. JRC's multilingual text mining resources

The previous section gave a rather brief overview of EMM functionality without giving technical detail. Scientific-technical details and evaluation results for all applications have been described in various publications available at <http://langtech.jrc.ec.europa.eu/>.

The four main EMM applications are freely accessible for everybody. Additionally, the JRC has made available a number of resources (via the same website) that will hopefully be useful for developers of multilingual text mining systems. The *JRC-Acquis* parallel corpus in 22 languages (Steinberger et al., 2006), comprising altogether over 1 billion words was publicly released in 2006, followed by the *DGT-Translation Memory* in 2007. A new resource that can be used both as a translation memory and as a parallel corpus for text mining use is currently under preparation. *JRC-Names*, a collection of over 400,000 entity names and their multilingual spelling variants gathered in the course of seven years of daily news analysis (see Section 3.3), has been released in

¹ <http://emm.newsbrief.eu/geo?type=event&format=html&language=all> displays continuously updated live maps.

September 2011 (Steinberger et al., 2011). *JRC-Names* also comprises software to look up these known entities in multilingual text. Finally, the *JRC Eurovoc Indexing software JEX*, which categorises text in 23 different languages according to the thousands of subject domain categories of the Eurovoc thesaurus², will also be released soon.

5. Ongoing and forthcoming work

EMM customers have been making daily use of the media monitoring software for years. While being generally satisfied with the service, they would like to have more functionality and even higher language coverage. JRC's ongoing research and development work focuses on three text mining main areas: (1) Multilingual multi-document summarisation: The purpose is to automatically summarise the thousands of news clusters generated every day; (2) Machine Translation (MT): While commercial MT software currently translates Arabic and Chinese EMM texts into English and hyperlinks to *Google Translate* are offered for all other languages, the JRC is working on developing its own MT software, based on *Moses* (Koehn et al., 2007); (3) Opinion mining / sentiment analysis: EMM users are not only interested in receiving contents, but they would also like to see opinions on certain subjects. They would like to see differences of opinions across different countries and media sources, as well as trends showing changes over time. See the JRC's Language Technology website for publications showing the current progress in these fields.

6. Acknowledgements

Developing the EMM family of applications was a major multi-annual team effort. We would like to thank our present and former colleagues in the OPTIMA group for all their hard work.

7. References

Koehn P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007): *Moses: Open Source Toolkit for Statistical Machine Translation*. Proceedings of the Annual Meeting of the Association for Computational

² See <http://eurovoc.europa.eu/>.

- Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
- Larkey, L., Feng, F., Connell, M., Lavrenko, V. (2004): Language-specific Models in Multilingual Topic Tracking. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.
- Piskorski, J., Belyaeva, J., Atkinson, M. (2011): Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. Proceedings of the 8th International Conference 'Recent Advances in Natural Language Processing'. Hissar, Bulgaria, 14-16 September 2011.
- Pouliquen B., Steinberger, R. (2009): Automatic Construction of Multilingual Name Dictionaries. In: C. Goutte, N. Cancedda, M. Dymetman & G. Foster (eds.), Learning Machine Translation. MIT Press - Advances in Neural Information Processing Systems Series (NIPS), pp. 59-78.
- Pouliquen B., Steinberger, R., Deguernel, O. (2008): Story tracking: linking similar news over time and across languages. In Proceedings of the 2nd workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2008) held at CoLing'2008. Manchester, UK, 23 August 2008.
- Pouliquen, B., Steinberger, R., Belyaeva, J. (2007): Multilingual multi-document continuously updated social networks. Proceedings of the Workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2007) held at RANLP'2007, pp. 25-32. Borovets, Bulgaria, 26 September 2007.
- Steinberger R. (forthcoming): A survey of methods to ease the development of highly multilingual Text Mining applications. Language Resources and Evaluation Journal LRE.
- Steinberger R., Pouliquen, B., van der Goot, E. (2009): An Introduction to the Europe Media Monitor Family of Applications. In: F. Gey, N. Kando & J. Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.
- Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 2142-2147. Genoa, Italy, 24-26 May 2006.
- Steinberger R., Pouliquen, B., Kabadjov, M., van der Goot, E. (2011): JRC-Names: A freely available, highly multilingual named entity resource. Proceedings of the 8th International Conference 'Recent Advances in Natural Language Processing'. Hissar, Bulgaria, 14-16 September 2011.
- Tanev, H. (2007): Unsupervised Learning of Social Networks from a Multiple-Source News Corpus. Proceedings of the Workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2007), held at RANLP'2007, pp. 33-40. Borovets, Bulgaria, 26 September 2007.
- Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., Steinberger, R. (2009): Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. *LinguaMÁTICA*, 2, pp. 55-66.