Near Real Time Information Mining in Multilingual News

M. Atkinson
EC Joint Research Center
Via E. Fermi 27549
Ispra, Italy
+39 0332 789350
martin.atkinson@jrc.it

E. Van der Goot
EC Joint Research Centre
Via E. Fermi 27549
Ispra, Italy
+39 0332 785900
erik.van-der-goot@jrc.it

ABSTRACT
This paper presents a near real-time multilingual news monitoring and analysis system that forms the backbone of our research work. The system integrates technologies to address the problems related to information extraction and analysis of open source intelligence on the World Wide Web. By chaining together different techniques in text mining, automated machine learning and statistical analysis, we can automatically determine who, where and, to a certain extent, what is being reported in news articles.

Categories and Subject Descriptors
D.3.1 [**Content Analysis and Indexing**]: Linguistic Processing

General Terms
Algorithms, Experimentation.

Keywords
Open Source Text, Information Mining and Analysis, Multi-linguality, Automated Media Monitoring.

1. INTRODUCTION
Information overload is a main challenge in our world today. Applying techniques from text mining, automated machine learning and statistical analysis can help to reduce this overload of information. In this paper we present a fully functional system that exploits such techniques to automatically extract information like "who, what, where and when" from news articles in near real time. Moreover, being a small team with limited resources we show the alternative techniques and technologies that we have exploited in order to have a fully operational system that supports our research work.

One application of our system is known as the European Media Monitor [1], a web based multi-lingual news aggregation system that collects over 75000 news articles per day from some 2000 news feeds in 42 languages. The system conducts some basic information mining to automatically determine what is happening to whom and where in the world. Every 10 minutes it automatically groups articles and displays the 10 largest groups per language. It also applies some more sophisticated information analysis techniques, for instance, to automatically detect violent events, derive reported social networks and analyze media impact.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

The public website (http://emm.jrc.it) provides a user interface to all this information. This public website is visited on a regular basis by some 30000 human users, and gets some 1.2M hits per day. The system runs 24/7 only on a few servers without the use of any database technology.

2. BASIC INFORMATION EXTRACTION
At the core of the EMM system is a processing chain of lightweight extensible processes each independently running and chained together using a very basic but reliable in-house developed web service architecture based on HTTP post. Articles flow through the processing chain as thin RSS items that grow as meta-data gets added at each stage of the processing chain.

The first process in this chain is a Scraper. This process periodically checks all web sites of interest and generates a simple RSS feed containing a list of the current items published on the site. In the case of HTML sites a custom html parser is used to convert web pages into XHTML which is then transformed into RSS. In the case of RSS feeds, information from multiple feeds is combined into one. The next system in the chain, Grabber, is then notified with the resulting RSS feed. Grabber detects the new articles published and using a patent pending text extraction process determines the main content of the new articles. Grabber produces a new RSS feed for each site, containing title, link, description and text for all new articles detected, which is then passed on to the next process in the chain, the automated language detection process.

The automatic language detection process uses word frequency tables to automatically detect the language of the RSS content in the title and description. The Information Clustering and Story tracking processes populate these frequency tables using a technique similar to an infinite input response filter.

Next the Entity Recognition process detects people and organizations in the article from a home grown information base of entities and organizations which is populated by an automated (offline) entity recognition system [2].

Homonyms are detected and disambiguated using a multilingual, classified geospatial information base of place names, provinces, regions and countries. The disambiguation module also uses the meta-information of previously recognized entities, in order to perform geo-tagging.

Then, the classification system works on two levels, first it classifies articles on Boolean combinations of multilingual keywords, second it classifies on Boolean combinations of previously discovered classes and other metadata. Finally a tonality module assigns tonality to the RSS Item using a similar keyword based approach as the Classification system. The articles

then flow into the Clustering and Story Tracking Cache. All of these processes that implement basic information extraction rely on the use of highly efficient finite state machine pattern matchers.

3. INFORMATION CLUSTERING

Every 10 minutes the last 4 hours of articles are hierarchically clustered in every language individually. Initially each news article is considered to be a cluster, the process is agglomerative and employs average group linkage to determine the distance between the clusters using a simple cosine measure. The clustering process continues until the maximum cosine distance falls below a certain set threshold which is a function of the theoretical density of the feature vectors, where a higher density leads to a higher threshold value. The article feature vectors are simple word count vectors, constructed using a simple bag of words approach, with some additional ad-hoc rules like: ignore top 100 frequent words, ignore words of 2 characters. A cluster only remains if it has at least 2 articles that are not duplicates from at least 2 different news stories.

This algorithm has also been modified and tested on ideogram based texts in order to cluster Chinese language articles.

The system also tracks the evolution of stories over time. It is represented as the evolution of a news cluster as it grows or shrinks in time as shown in figure 1. If the cluster grows suddenly breaking news alert is generated and sent via email to subscribed users

4. INFORMATION ANALYSIS

Once we have clusters of articles we can start to employ more advanced information analysis techniques. One application area is the automatic detection of Events or to automatically identify "who did what to whom, when, with what methods (tools), where and why [3]. Here, we employ a lightweight linguistic approach in order to remain as much as possible language independent. First the cluster is passed through a linguistic pre-processor called CORLEONE [4] which includes fine-grained tokenization, sentence splitting, domain-specific dictionary look-up (e.g., recognizing numbers, quantizes, person titles), labeling of key terms indicating Unnamed person groups(e.g. civilians, policemen, Shiite), and morphological analysis. Then these results are passed through a cascade of extraction grammars to fill the slots of the events (i.e. who, whom, with what ..etc.). The patters for these grammars are created though machine learning using seed patterns combined with clustered news articles. We are currently working on expanding the grammar patterns to support more languages. This is currently the final process in our processing chain

5. SOCIAL NETWORKS AND IMPACT

Current research work includes the automated construction of Social Networks. These networks, within the context of media monitoring, try to identify who is talking about/to whom. We have experimented with two approaches [5], cluster co-occurrence analysis and automated grammar extractor that both provide complimentary results.

We are also looking into Media Impact or opinion mining, who is talking about what and how. Here, we gather and combine statistics on article classification, tonality and article source, then display this information using traditional graphing techniques. In addition we are also looking into clustering articles by category over time to show the main topics within any category over a longer period of time

6. CONCLUSION

In our experience yesterday's news is old news, therefore, systems for automated media information extraction need to work in real time. We have presented our system for information extraction that builds from simple pattern matching techniques to more sophisticated, but none the less language agnostic, extraction grammars. For our research it is important to have a fully operational real-world system to test the performance of our algorithms.

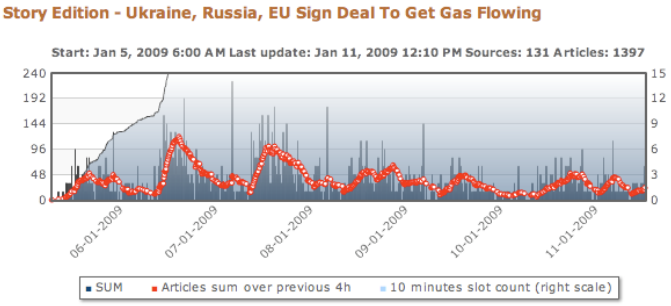

Figure 1. Graphs of single cluster story's evolution over time, the blue area shows the cumulative number of articles, the bars show the instantaneous increase of articles in the story and the red line shows story 4 hour evolutions

7. ACKNOWLEDGEMENTS

Our thanks to OPTIMA team who's research and programming work made this system possible.

8. REFERENCES

[1] Best, C. van der Goot, E. Blackler, K. Garcia, T. and Horby, D. Europe Media Monitor. Technical Report EUR 22173 EN, European Commission, 2005.

[2] Steinberger, R. and Pouliquen, B. Cross-lingual Named Entity Recognition. Journal Linguisticae Investigationes, Special Issue on Named Entity Recognition and Categorisation, LI 30:1 (2007) John Benjamins Publishing Company. ISSN 0378-4169. (2007) 135-162.

[3] Piskorski, J. Tanev, H. Atkinson, M. and van der Goot, E . Cluster-Centric Approach to News Event Extraction. Proceedings of the International Conference on Multimedia & Network Information Systems (Wroclaw, Poland September 2008). IOS Press.

[4] Piskorski, J. CORLEONE - Core Linguistic Entity Online Extraction. Technical Report EUR 23393 EN, European Commission, 2008.

[5] Pouliquen, B, Tanev, H. and Atkinson, M. (2008). Extracting and Learning Social Networks out of Multilingual News. Proceedings of the Social Networks and Application tools workshop (Skalica, Slovakia, Septembe