

NOAM: News Outlets Analysis and Monitoring System

Ilias Flaounas¹, Omar Ali¹, Marco Turchi², Tristan Snowsill¹,
Florent Nicart³, Tijl De Bie¹, and Nello Cristianini¹

¹Intelligent Systems Laboratory, Univ. of Bristol, MVB, Woodland Rd, BS8-1UB, UK

²European Commission - JRC (IPSC) Via E. Fermi, 2749 I-21027 Ispra (VA), Italy

³Université de Rouen, LITIS EA 4108, 76800 Saint-Étienne-du-Rouvray, France

ABSTRACT

We present NOAM, an integrated platform for the monitoring and analysis of news media content. NOAM is the data management system behind various applications and scientific studies aiming at modelling the mediasphere. The system is also intended to address the need in the AI community for platforms where various AI technologies are integrated and deployed in the real world. It combines a relational database (DB) with state of the art AI technologies, including data mining, machine learning and natural language processing. These technologies are organised in a robust, distributed architecture of collaborating modules, that are used to populate and annotate the DB. NOAM manages tens of millions of news items in multiple languages, automatically annotating them in order to enable queries based on their semantic properties. The system also includes a unified user interface for interacting with its various modules.

Categories and Subject Descriptors

H.2.8 [Database applications]: Data mining

General Terms

Algorithms, Design

1. INTRODUCTION

In this demo we present a data management infrastructure developed as part of research in Artificial Intelligence (AI) with the purpose of extracting semantic level properties from unstructured multilingual text, and using them to detect patterns in news content. We combine different AI techniques from fields such as data mining, machine learning and natural language processing into a reliable working system. The combination of multiple AI methods for automated data analysis remains an important question in the AI community; most works focus on a single method e.g., clustering or classification of data. We focus on the field of social sciences and in particular the problem of the automation of news media content analysis. Automation of news analysis on a large scale has only recently become feasible due to digitisation of media content and advances in modern AI techniques.

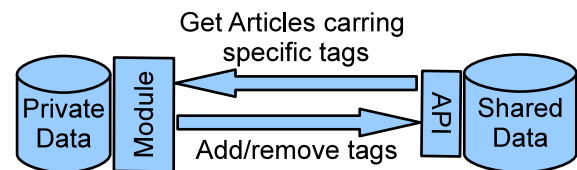


Figure 1: Functionality of a generic module.

Our solution to these problems is the development of a data management system that we name NOAM. It is an integrated platform for the monitoring and automation of analysis of media content. NOAM is centred around a database, which is populated by a web pipeline and by a statistical machine translation module, operating on multilingual news content. The idea of a data management system that autonomously generates parts of its contents, e.g., by translation and annotation, is a central part of the philosophy of our tool.

Our demo focuses on presenting the internal architecture of NOAM that has enabled the continuous analysis of news content for over two years. NOAM is the system behind a series of recent scientific results on media content analysis, e.g., visualisation of the mediasphere network for the first time [5], analysis of factors that affect media content [4], detection of gender biases among different topics in news [1], analysis of relations between countries as they are reflected in their media content [3], and detection of memes [6]. Currently we monitor more than a thousand multilingual news outlets, and analyse 40K news items per day. So far, we have analysed more than 30 million news items. The analysis is performed using state of the art methods organised under a data management architecture of cooperative, independent modules.

Commercial systems like Google News and Yahoo! News have different goals to NOAM. We aim to analyse the mediasphere content rather than provide access to the actual news. We offer novel perspectives and discover patterns that those systems do not offer, like comparison of countries based on the topics that their media promote.

2. ARCHITECTURE

NOAM is designed using a modular, scalable, and distributed architecture. We refer to the core system units as ‘modules’. Figure 1 illustrates the functionality of a generic

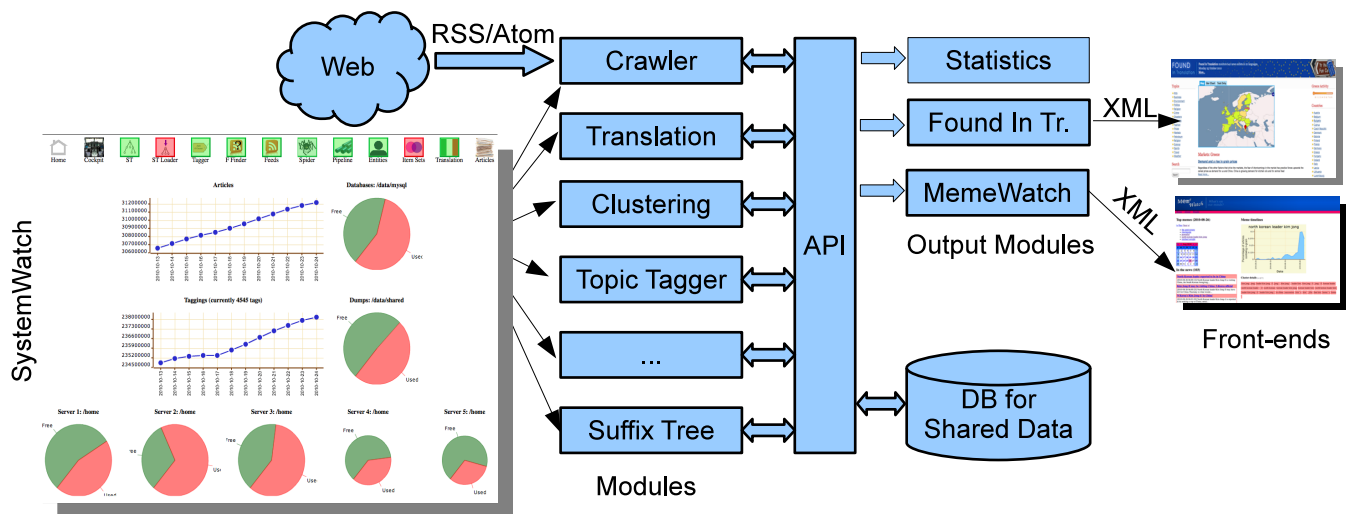


Figure 2: The modular system architecture.

module. Each module is designed to perform a specific task and typically implements a state of the art method from the field of machine learning, data mining or natural language processing. The functionality of a module is to retrieve from the DB a set of articles that carry a specific set of associated tags. It will work on those articles and will write back information in the form of new tags. Access to the shared data, that is annotated articles, is performed using the same API for all modules. Each module can store its own private data as required, e.g., the module ‘Topic tagger’ stores the classifiers parameters after their learning phase.

The overall system architecture is illustrated in Fig. 2. It mainly consists of an expandable set of modules. All modules share and access data (articles and tags), through the common API. The API enables the logical separation of the DB from the implementation of modules and it provides the interaction layer across modules – based on tags. Specialised modules, that implement crawlers, retrieve new data from the internet. Output modules create XML files that are used in e.g., live news monitoring front-ends or statistical reports about the mediasphere.

Modules autonomously annotate the data in order to facilitate semantic level management and retrieval. They are independent from another and work in parallel. If a module stops working the others can keep on performing their tasks. Modules can also be organised to work in a serial way, without changing the architecture. This is achieved by designing the required tags that an article has to carry to become input to a module. Here is a scenario that highlights the interactions of modules: Each day, articles are gathered (Crawler module), translated into English if they are found in a different language (Translation module), English and Machine-English articles are tagged according to their topic (Topic tagger module), the topics distributions of the leading stories of the given day are measured (statistics output module).

3. IMPLEMENTED MODULES

We summarize the implemented modules that currently make up the NOAM system:

- *Crawler* This module is responsible for crawling a pre-defined lists of news feeds, in RSS or Atom format, that provide the content of news outlets of interest. News feeds contain links to the articles’ bodies. An HTML scraper identifies and collects the textual article content from each of these article webpages. The crawler tags these newly discovered articles with a set of tags that will enable other modules to work on them, for example the language of article.
- *Translation* This module translates all non-English articles into English using a statistical machine translation approach. Currently we translate all 21 main EU languages [7].
- *Topic Taggers* The topic of articles is discovered using topic classification based on Support Vector Machines [2, 7].
- *Clustering* This module clusters articles into stories, that is sets of articles that discuss the same event [5, 4].
- *Suffix Tree* This module implements a suffix tree that enables the detection of frequent phrases, known as memes, and significant events [6].
- *Entities Detection* This module detects the named entities present in articles [1]. Name entities are people, organisations and locations.
- *Feed Finder* The addition of new feeds into the system is performed in a semi-automatic way. This module crawls the net and tries to discover new outlets and corresponding feeds. The outlets and feeds have to be confirmed and annotated manually for addition to the system.
- *Output Modules* Outputs of the system are in XML format and they are used by the NOAM front-ends. They are activated typically once per day producing results based on the data of the previous day. For example it is possible for the system to retrieve news

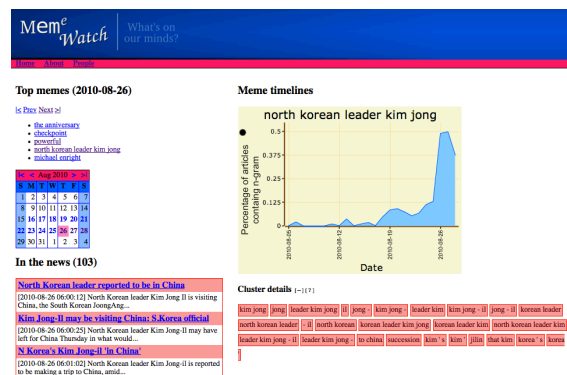
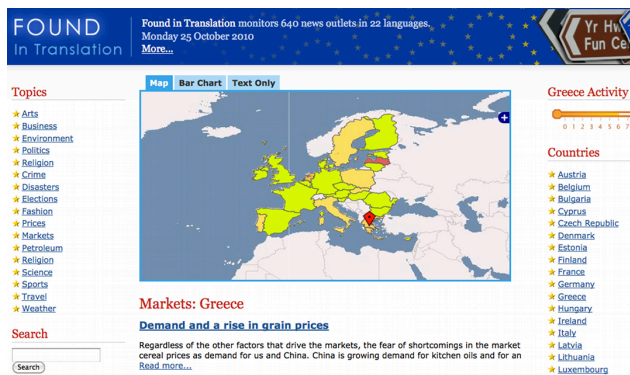


Figure 3: Two front-ends of NOAM that enable interaction with final user. On the left ‘Found in Translation’ and on the right ‘MemeWatch’.

items in German that discuss environmental issues, or all news items in all languages covering the same event on a given day.

A web based graphical user interface, that we call ‘System Watch’, has been developed for controlling the whole NOAM system. It provides an overall view of the system or a module-by-module access interface. This is the place where module parameters can be tuned e.g., the crawler GUI provides an easy way to import new outlets and feeds into the crawler and annotate them. Furthermore it provides information about the status of the physical servers where the modules and DB live. Currently for our projects we use five dedicated servers.

4. DISCUSSION

The audience will be able to interact with at least two web interfaces presenting statistical patterns found in the contents of the mediasphere. The two front-ends of the system we will focus on are ‘Found In Translation’¹ [7] and ‘MemeWatch’² [6] both illustrated in Fig. 3. The first translates EU news from 21 languages, and colours the EU map based on the prevalence of topics in each country. The audience will be able to click on maps, to generate histograms and lists, as well as to read machine-generated text. The second interface detects fast growing memes in the media sphere and is used for event detection. The audience will be able to see the current ‘hot’ memes that dominate the news, timelines of their volumes, and compare their influence.

Behind this, there is NOAM, our modular data management system. We will present to the audience a view of the system from behind the scenes. This is possible using the ‘System Watch’ back-end interface. The participants will be guided through possible user scenarios, involving media analysts studying certain macro-features of news content, which are based on actual studies that have taken place in the past years.

¹ Accessible at <http://foundintranslation.enm.bris.ac.uk>

² Accessible at <http://memewatch.enm.bris.ac.uk>

5. ACKNOWLEDGMENTS

The authors want to thank Philip Naylor for computer infrastructure support and the entire group on ‘Pattern Analysis and Intelligent Systems’ at the University of Bristol for discussions. This work is partially supported by European Commission through the PASCAL2 Network of Excellence (FP7-216866). I. Flaounas is supported by Alexander S. Onassis Public Benefit Foundation, and N. Cristianini is supported by a Royal Society Wolfson Merit Award.

6. REFERENCES

- [1] O. Ali, I. Flaounas, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Automating news content analysis: An application to gender bias and readability. In *JMLR W&CP: Workshop on Applications of Pattern Analysis*, pages 36–43, 2010.
- [2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.
- [3] I. Flaounas, N. Fyson, and N. Cristianini. Predicting relations in news-media content among EU countries. In *Cognitive Information Processing, 2nd International Workshop on*, pages 269–274. IEEE, 2010.
- [4] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. The Structure of EU Mediasphere. *PLoS ONE*, page e14243, 2010.
- [5] I. Flaounas, M. Turchi, T. De Bie, and N. Cristianini. Inference and validation of networks. In *Machine Learning and Knowledge Discovery in Databases, European Conference*, pages 344–358. Springer, 2009.
- [6] T. Snowsill, F. Nicart, M. Stefani, T. De Bie, and N. Cristianini. Finding surprising patterns in textual data streams. In *Cognitive Information Processing, 2nd International Workshop on*, pages 405–410, 2010.
- [7] M. Turchi, I. Flaounas, O. Ali, T. De Bie, T. Snowsill, and N. Cristianini. Found in translation. In *Machine Learning and Knowledge Discovery in Databases, European Conference*, pages 746–749. Springer, 2009.