Frontex Real-time News Event Extraction Framework

Martin Atkinson
Joint Research Centre
Via Fermi 2749
21027 Ispra, Italy
Firstname.Lastname@jrc.ec.europa.eu

Jakub Piskorski
Research & Development Unit, Frontex
Rondo ONZ 1
00-124 Warsaw, Poland
Firstname.Lastname@frontex.europa.eu

ABSTRACT

An ever-growing amount of information relevant for early detection of certain threats can be extracted from on-line news. This led to an emergence of news mining tools to help analysts to digest the overflow of information and to extract valuable knowledge from on-line news sources.

This paper gives an overview of the fully operational Real-time News Event Extraction Framework developed for Frontex, the EU Border Agency, to facilitate the process of extracting structured information on border security-related events from on-line news. In particular, a hybrid event extraction system has been constructed, which is applied to the stream of news articles continuously gathered and pre-processed by the Europe Media Monitor—a large-scale multilingual news aggregation engine. The framework consists also of an earth browser, in which events are visualized and an event moderation tool, which allows to access the database of automatically extracted event descriptions and to clean, validate, group, enhance and export them into other knowledge repositories.

Categories and Subject Descriptors

H.3.3 [**Information Systems and Retrieval**]: Text Mining; H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms, Security, Experimentation

Keywords

Multilingual News Mining, Event Extraction, Visualization

1. INTRODUCTION

On-line news has been considered by various security authorities and organisations around the globe as an important source of information that can be exploited for early detection of certain threats and for situation monitoring during crisis. This is mainly due to the fact that: (a) information on certain security-related events might not be available from any other sources or it might be incomplete (e.g., developments in third countries), (b) there might be a significant delay before such information is made available via official channels, and (c) information from on-line media can be used for cross-checking with information available from other sources. The aforementioned observations together with an ever-growing amount of information transmitted through the Internet led to an emergence of advanced tools that combine techniques from text mining, machine learning, statistical analysis and computational linguistics to help intelligence experts to manage the overflow of information, filter out the relevant from the irrelevant, and to extract valuable and actionable knowledge from on-line sources.

A significant number of approaches to news mining and news exploration systems have been reported recently. The most prevalent way to organize news by such systems is to classify the incoming news into predefined or automatically discovered categories and to group topically simmilar news articles into clusters. However, as has been emphasized in [7], in order to facilitate an in-depth analysis of the news it is essential to extract structured information on the events from the news, i.e., to derive detailed information about them, ideally identifying *who did what to whom, through what methods (instruments), when, where and why* [1]. The current capabilities of news event extraction technology deployed in the security domain are exemplified in [12] (epidemiology) and [6] (armed conflicts), whereas [2] reports on general trends in the field of event extraction.

This paper gives an overview of the Real-time News Event Extraction Framework developed for Frontex[1], to facilitate the process of extracting structured information on border-security related events from on-line media in order to support situation monitoring and intelligence gathering, with a particular focus on incidents related to illegal migration (e.g., illegal entry attempts), cross-border crime (e.g., smuggling), and crisis situations (e.g., violent events, natural disasters, biohazards) at and beyond the EU external borders.

The specific tasks of Frontex impose three requirements on the event extraction tools, in particular, they should: (a) extract information in close to real time, (b) extract as fine-grained event descriptions as possible, (c) process news articles in many different languages, since a significant fraction of relevant events are only reported in non-English, lo-

[1]Frontex - the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union

KDD'11, August 21–24, 2011, San Diego, California, USA.

cal news. To meet these requirements a hybrid multilingual event extraction engine has been constructed, which is applied to the stream of news articles continuously gathered by a large-scale multilingual news aggregation engine, and is capable to extract event information in 7 languages. It is also equipped with tools for storing, moderation and visualization of the automatically extracted event descriptions.

The event extraction tools presented in this paper are currently being integrated with the pilot version of EUROSUR (an Integrated European Border Surveillance System)[2], that will allow to share information of common interest related to border security between Member States of the European Union. The event extraction tools will be specifically deployed to semi-automatically populate the incident information layer of EUROSUR.

2. FRAMEWORK ARCHITECTURE

The event extraction framework architecture is depicted in Figure 1. First, news articles are gathered by a large-scale news aggregation engine, the Europe Media Monitor (*EMM*)[3] developed at the Joint Research Centre of the European Commission (JRC) [4]. *EMM* retrieves more than 100,000 news articles per day from more than 2500 news feeds in 42 languages. These news articles are geo-located, tagged with meta-data and further filtered (classified) using standard keyword-based techniques in order to select those articles, which potentially refer to security-related incidents and events. In addition, the news articles harvested within a 4-hour window are grouped into clusters in every language individually according to content similarity (using hierarchical agglomerative clustering). The filtering and clustering process is performed every 10 minutes. Figure 2 shows the web interface to EMM.

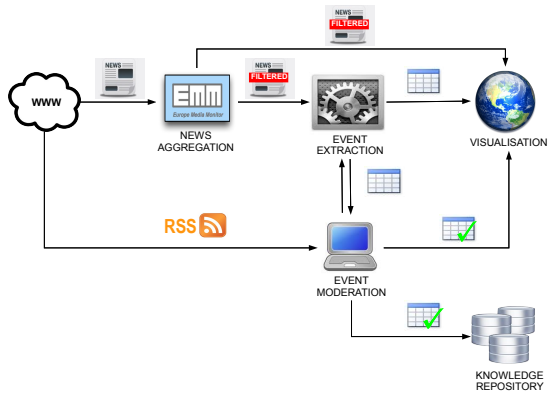

Figure 1: The system architecture.

Next, the stream of filtered news articles and clusters is passed every 10 minutes to the event extraction engine, which consists of two core event extraction systems, namely, *NEXUS*[10, 9], developed at JRC, and *PULS*[5, 12] developed by the University of Helsinki. *NEXUS* follows a shallow cluster-centric approach, which makes it more suitable for extracting information from the entire cluster of topically-related articles, whereas *PULS* follows a non-cluster centric approach and performs a more thorough analysis of the full text of each news article. The main drive behind the deployment of the two aforementioned systems was to: (a) obtain richer coverage wrt. language and event types (both systems have been tuned to detect event types which are not entirely overlapping), and (b) provide a platform to investigate two different approaches to event extraction.

For all event types of interest there is a harmonised event template structure, which includes the following slots: `TYPE`, `SUBTYPE`, `DESCRIPTOR` (free text), `SNIPPET` (text fragment triggering the event), `PUB_DATE`, `DATE`, `LOCATION`, `CONFIDENCE` (system's), `RELEVANCE` (for the user), `SEVERITY`, `SOURCE`, `PERPETRATOR`, `VICTIM`, `ITEM`, `MEANS`, `NUM_AFFECTED`, `NUM_INJURED`, `NUM_KILLED`, `NUM_ARRESTED`, and `WOMEN/MINORS_INVOLVED` (boolean-valued).

The output produced by the core event extraction engine, i.e., a stream of instantiated event templates, is made accessible to the earth browser for visualization of the extracted events on a map. News articles filtered using keyword-based heuristics as potentially relevant to border security domain are also accessible in the earth browser, in a separate layer.

In order to bridge the gap between the automated event extraction phase and an in-depth analysis phase, an event moderation system provides functionality to access the database of automatically extracted event descriptions and to clean, validate, compare, group, filter, enhance and export them. In particular, cleaned and validated event descriptions can be made accessible in a separate 'moderated events' layer in the earth browser and they can be exported into other knowledge repositories. The core event extraction engine can also be applied on any RSS feed on demand.

Figure 2: Europe Media Monitor.

3. EVENT EXTRACTION ENGINE

Currently, the core event extraction engine consists of two event extraction systems, namely *NEXUS* and *PULS*.

NEXUS, follows a shallow cluster-centric approach [10, 9]. Each cluster of topically related articles undergoes a shallow linguistic analysis, i.e., fine-grained tokenization, morphological analysis, gazetteer look-up, sentence boundary detection, etc., and a cascade of simple finite-state extraction grammars is applied to each article in the cluster. While the lower-level grammars are used to extract person names (e.g., *Osama bin Laden*), person groups (e.g., *Algerian immigrants*), numerical expressions (e.g., *two hundred*), quantifiers (e.g., *More than*), and other small-scale structures

[2]http://eur-lex.europa.eu/LexUriServ/-LexUriServ.do?uri=COM:2008:0068:FIN:EN:PDF
[3]http://press.jrc.it

(e.g., *small boats*), the top-level grammar consists of 1–2 slot extraction patterns, e.g., `"PER-GROUP <IMMIGRANT> "sbarcati clandestinamente"`, which applied to the text *15 palestinesi sbarcati clandestinamente in Sicilia* would result in assigning the group of 15 Palestinians (*15 palestinesi*) the semantic role `IMMIGRANT`. The grammars are encoded and processed with *ExPRESS*, a highly efficient IE-oriented pattern matching engine [8].

The system processes only initial sentences and the title of each article since: (a) usually the most important parts of the story are placed in the beginning of the article and the least important facts are left toward the end, (b) processing the entire text might involve handling more complex language phenomena (e.g., anaphora, ellipsis and complex syntax), which is hard and requires knowledge-intensive processing[4], (c) if some crucial information has not been captured from one article in the cluster, it might be extracted from other articles in the same cluster.

Since the information about events is scattered over different articles, the last step consists of cross-article cluster-level information fusion in order to produce full-fledged event descriptions, i.e., information extracted locally from each single article in the same cluster is aggregated and validated. This process encompasses mainly three tasks: entity role disambiguation (as a result of extraction pattern application the same entity might be assigned different roles), victim counting and event type classification, all accomplished through heuristics. If the same entity has two roles assigned in the same cluster, preference is given to the role assigned by the most reliable group of patterns, e.g., 2-slot extraction patterns are considered more reliable than 1-slot extraction patterns. It is important to note that *NEXUS* extracts only the main event for each cluster ('one sense per discourse'). See [10, 9] for details on fusion techniques deployed.

NEXUS is capable of processing news in English, Italian, Spanish, French, Portuguese, Russian and Arabic. Due to a linguistically light-weight approach and deployment of weakly supervised methods (bootstrapping techniques) for creating language-specific resources (i.e., domain-specific lexica and extraction patterns), *NEXUS* can be adapted to process texts in a new language in a short time [13, 11]. Although the *NEXUS* architecture is language-independent, the system is quite flexible regarding the grammar design approach that it can implement. In fact, while surface pattern-based grammars proved to be largely effective for English language, to get a comparable performance for Romance languages was obtained through integration of extraction patterns referencing more abstract morphological information due to a number of phenomena within this language family (see [13, 11] for details).

The cluster-centric approach to event extraction described above appeared to work satisfactorily in case of crisis-related events (e.g., man-made and natural disasters, violent events, etc.) [9]. However, extracting information on illegal migration incidents and related cross-border crimes poses additional challenges, e.g., (a) information about incident type is not explicitly encoded in language, (b) such incidents are often reported in local news only, (c) geo-locating is more difficult since there are usually several geo-references in an article[5]. In [3] we address tackling this phenomena in detail.

The *PULS* system[6] is similar in essence to *NEXUS*, but differs in few aspects: (a) it analyzes the entire text of the incoming news articles and performs more in-depth linguistic analysis (e.g., anaphora resolution), which allows us to handle events which may be scattered more widely throughout the article for which not much information has been reported, (b) it is not run on clustered news articles, but attempts to aggregate the extracted facts into groups of related event templates. At the present time, *PULS* analyzes articles in English and French, and is currently being extended to process Russian news.

Both *NEXUS* and *PULS* were evaluated in terms of coverage and precision. As for the border security domain, the precision of slot extraction ranges from 60% to 95% depending on language and slot type, with an average overall slot extraction precision oscillating around 85%. More detailed evaluation figures are given in [3]. In order to improve these figures we are currently exploring the usefulness of cross-lingual information fusion.

4. INFORMATION ACCESS

The main added value to an automated event extraction system is the possibility to clean and combine automatically extracted event descriptions and to visualize them on a map. In order to facilitate this an event moderation component is provided and KML services were set up to visualise the events in earth browsers.

4.1 Event Visualization

For visualisation, events are grouped over time spans, by event classes, source language, and the system that extracted them. This groups events into geo-spacial layers (updated on 10 minute basis) that can then be visualized in tools like Google Earth. The screenshot in Figure 3 shows on the left side the various layers provided from both the automated extraction systems as well as the event moderation system (see 4.2). The events in a given layer are visible on the right side of the window on a map. A selected event (each associated with an icon) shows the instantiated event template in a bubble, with the link to a cluster or single news article from which the information was extracted.

4.2 Event Moderation and Analysis

Event moderation system allows events to be retrieved according to a number of different criteria, including: event type, date of occurrence, language, source (automatically extracted or moderated event), and location. Searches on multiple values are allowed. A screenshot of the user interface to the event moderation system is shown in Figure 4. On the left side 'filtered' events are listed together in different colours according to their source. Selecting an event allows the event template to be visualized (on the right side of the screen) and eventually edited and grouped with other events. The moderation system also provides the possibility to input a new event from say a field report and also

[4]In the context of mining news in different language it is important to note that language-specific processing resources needed to tackle such phenomena might be either non-existing or 'expensive' to obtain.

[5]To event place, country of departure, destination country, places visited on the route, references to countries of origins of the perpetrators, the origin of the means used by authorities, etc.

[6]*http://puls.cs.helsinki.fi*

Figure 3: Visualization of events in Google Earth.

to apply event extraction to any RSS data feed. Furthermore, the event moderation tool is equipped with services to translate part of the event template into any of the 27 EU languages. Finally, 'dynamic ontology' allows to define how the extracted events are interpreted and displayed.

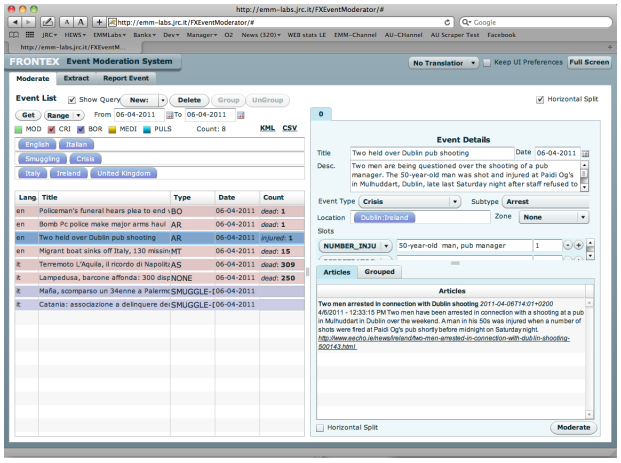

Figure 4: Event moderation: main user interface.

Aggregation of information is the next simple step in a path towards supplying better support to analysts to detect and discover patterns in the pool of gathered information. This is achieved by counting membership of event groups with certain slot values. For instance, grouping by event type, location and time permits the profiling of event type over time and space. An example of this is given in Figure 5, where the aggregation of Civil Unrest and Protest events are shown over time (per week per country since the beginning of 2011 in the Middle East and North Africa - MENA). The foremost peek shows the violence in Egypt followed by Yemen,Bahrain, Libya and finally a high peek from Syria. Clearly this information could be visualized or played over time on a map. A comprehensive description of the event moderation and visualisation can be found in [3].

5. ACKNOWLEDGMENTS

We are greatly indebted to all our colleagues in the OPTIMA action at the Joint Research Centre, who are working on *EMM* and *NEXUS*, and we also would like to the thank our colleagues in the Department of Computer Science of the University of Helsinki, who develop *PULS* system.

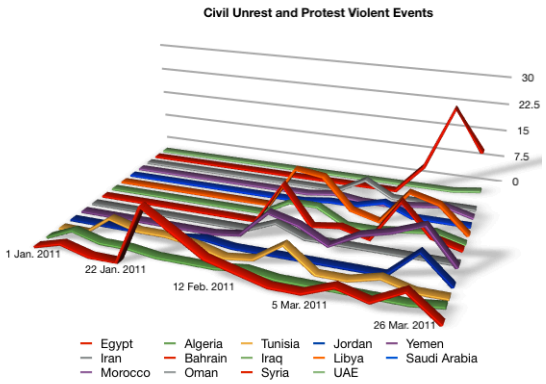

Figure 5: Graph of violent protests in the MENA region (per week basis from the beginning of 2011)

6. REFERENCES

[1] D. Appelt. *Introduction to Information Extraction Technology. Tutorial held at IJCAI-99.*

[2] N. Ashish, D. Appelt, D. Freitag, and D. Zelenko. *Proceedings of the Workshop on Event Extraction and Synthesis.* Held in conjunction with the AAAI 2006.

[3] M. Atkinson, J. Piskorski, R. Yangarber, and E. van der Goot. Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering. In U. K. Wiil, editor, *Open Source Intelligence and Counter-terrorism.* Springer, LNCS, Vol. 2, 2011.

[4] M. Atkinson and E. van der Goot. Near Real Time Information Mining in Multilingual News. In *Proceedings of WWW 2009.*

[5] R. Grishman, S. Huttunen, and R. Yangarber. Real-time Event Extraction for Infectious Disease Outbreaks. In *Proceedings of HLT 2002*, 2002.

[6] G. King and W. Lowe. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders. *International Organization*, 57:617–642, 2003.

[7] J. Li, J. Li, and J. Tang. A flexible topic-driven framework for news exploration. In *Proceedings of KDD 2007*, 2007.

[8] J. Piskorski. ExPRESS – Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of FSMNLP 2007*, 2007.

[9] J. Piskorski, H. Tanev, M. Atkinson, E. van der Goot, and V. Zavarella. Online news event extraction for global crisis surveillance. *Accepted for Transactions on Computational Collective Intelligence*, 2011.

[10] H. Tanev, J. Piskorski, and M. Atkinson. Real-Time News Event Extraction for Global Crisis Monitoring. In *Proceedings of NLDB 2008.*, pages 207–218, 2008.

[11] H. Tanev, V. Zavarella, J. Linge, M. Kabadjov, J. Piskorski, M. Atkinson, and R. Steinberger. Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. *Linguamática: Revista para o Processamento Automático das Línguas Ibéricas*, 2:550–566, 2009.

[12] R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen. Extracting Information about Outbreaks of Infectious Epidemics. In *Proceedings of the HLT-EMNLP 2005*, Vancouver, Canada, 2005.

[13] V. Zavarella, H. Tanev, and J. Piskorski. Event Extraction for Italian using a Cascade of Finite-State Grammars. In *Proceedings of FSMNLP 2008*, 2008.