Automated Event Extraction in the Domain of Border Security

Martin Atkinson[1], Jakub Piskorski[2], Hristo Tanev[1], Eric van der Goot[1], Roman Yangarber[3], and Vanni Zavarella[1]

[1] Joint Research of the European Commission, 21027 Ispra (VA), Italy
{Firstname.Lastname,Firstname.Multi-word-lastname}@jrc.ec.europa.eu
[2] Frontex, Rondo ONZ 1, Warsaw, Poland
Firstname.Lastname@frontex.europa.eu
[3] University of Helsinki, P.O. Box 68, 00014 Helsinki, Finland
Firstname.Lastname@cs.helsinki.fi

Abstract. This paper gives an overview of an ongoing effort to construct tools for automating the process of extracting structured information about border-security related events from on-line news. The paper describes our overall approach to the problem, the system architecture and event information access and moderation.

Keywords: event extraction from on-line news, border security, text analysis, open source intelligence.

1 Introduction

Mining open sources for gathering intelligence for security purposes is becoming increasingly important. Recent advances in the field of automatic content extraction from natural-language text result in a growing application of text-mining technologies in the field of security, for extracting valuable structured knowledge from massive textual data sets on the Web. This paper gives an overview of an ongoing effort to construct tools for Frontex[1] for automating and facilitating the process of extracting structured information on border-security related events from on-line news articles. The topics in focus include illegal immigration incidents (e.g., illegal entry, illegal stay and exit, refusal of entry), cross-border criminal activities (e.g., trafficking, forced labour), terrorist attacks, other violent events (e.g., kidnappings), inland arrests in third countries, displacements, troop movements, man-made and natural disasters, outbreak of infectious disease, and other crisis-related events.

The need of strengthening capabilities for tracking the security situation in the source and target countries for illegal immigration into the EU has been identified and acknowledged by the European Commission (EC). Specifically, the Commission Communication COM (2008) 68 proposes the creation of an

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

P. Daras and O. Mayora (Eds.): UCMedia 2009, LNICST 40, pp. 321–326, 2010.

Integrated European Border Surveillance System (EUROSUR), where step 6 of Policy Option 1 suggests development and deployment of new tools for strategic information to be gathered by Frontex from various sources in order to recognize patterns and analyze trends, supporting the detection of migration routes and the prediction of risks for Common Pre-frontier Intelligence Picture (CPIP). The deployment of open-source intelligence tools for event extraction plays a significant role in this context. There are two end-users of such tools in Frontex, namely, the Risk Analysis Unit (RAU), whose task is to carry out border-security related risk analysis to drive the operational work of the agency, and Frontex Situation Centre (FSC) responsible for providing a constant and short-term picture of the situation, at the EU-external borders and beyond.

Taking into account RAU's and FSC's needs, there are some challenges to be tackled while developing tools for automated event extraction from on-line news. First, for the purpose of risk analysis it is crucial to extract as fine-grained event descriptions as possible, ideally including information on: event's type/subtype, date, precise location, perpetrators, victims, methods/instruments used (if applicable), number and characteristic of people affected, link to the source(s) from which the event was detected, and system's confidence in the automatically generated event description. Second, the capability of providing a live or near-live picture of border-security related events imposes that such tools allow for processing vast amount of news articles in real or almost real-time. Finally, it is crucial to be able to process news articles in many different languages, since a significant fraction of relevant events are only reported in non-English, local news, where Italian, Spanish, Greek, French, Turkish, Russian, Portuguese, and Arabic are the most important ones at the moment.

2 System Architecture

This Section gives an overview of the system architecture, which is depicted in Figure 1.

- First, news articles are gathered by a dedicated software platform for electronic media monitoring, the Europe Media Monitor (EMM)[2] developed at the JRC [1]. EMM currently retrieves 100,000 news articles per day from 2,000 news sources in 42 languages. Articles are classified according to about 700 categories and then scanned to identify known entities. Information about entity mentions is added as meta-data for each article.
- The news articles (harvested in a 4-hour time window) are grouped into clusters according to content similarity. Then each cluster is geo-located, and clusters describing events relevant for the border security domain are selected using keyword-based heuristics. These clusters constitute a small portion (between 1-2% on average) of the stream retrieved by EMM. The clustering process is performed every 10 minutes.
- Next, two event extraction systems are applied on the stream of news articles.

[2] http://press.jrc.it

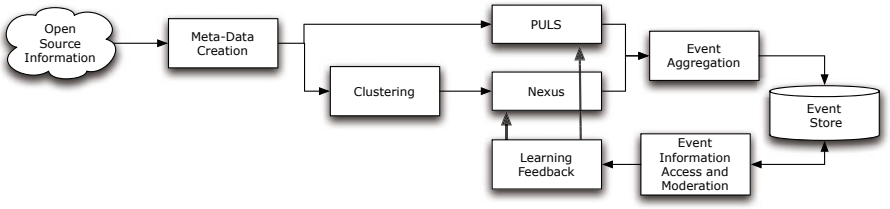

Fig. 1. Event extraction System Architecture

- The first extraction system, *NEXUS*, follows a cluster-centric approach [3,2]. Each cluster is processed by an event extraction engine, which performs shallow linguistic analysis and applies a simple cascade of finite-state extraction grammars[3] on each article in the cluster. The system processes only the top sentence and the title of each article for the following reasons: (a) news articles are written in the "inverted-pyramid" style, i.e., 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, which is hard and requires knowledge-intensive processing; (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. *NEXUS* detects and extracts only the main event for each cluster. It is capable of processing news in several languages, including, i.a., English, Italian, Spanish, French, Portuguese, and Russian. Due to a linguistically light-weight approach *NEXUS* can be adapted to processing texts in a new language in a relatively short time [7,8].
- The second system, *PULS*[4], uses a full-document approach, described in, e.g., [5,6]. This extraction component exploits a similar pattern-based technology to analyze the news articles. It *first* analyses one document at a time, to fill as many event templates as possible for each given document, and then attempts to unify the extracted events into "groups" in a subsequent phase, across different articles. The document-local analysis covers the entire text of each article, which aims to obtain not only the current information, but possibly links to background information—typically reported further down in the article—as well. PULS currently performs processing in English and French, with plans to extend to additional languages in the near future.

[3] The grammars are encoded and processed with *ExPRESS*, a highly efficient extraction pattern matching engine [4].

[4] http://puls.cs.helsinki.fi/medical/

NEXUS follows a cluster-centric approach, which makes it more suitable for extracting information from the entire cluster of topically-related articles. *PULS* performs a more thorough analysis of the full text of each news article, which allows us to handle events for which not much information has been reported.

The two approaches are deployed for event extraction in order to get richer coverage. Also, the two extraction components are (at present) tuned to detect sets of event types that are not entirely overlapping. In particular, *NEXUS* has been mainly deployed for detecting violent incidents and natural and man-made disasters, whereas *PULS* has been primarily customized for the epidemics domain. The combined system is designed to compare and experiment with different approaches to event extraction on the same data.

Although no evaluation of the combined system has been performed yet, information on the current performance levels of *NEXUS* and *PULS* can be found in [3,2,7,9].

3 Event Information Access and Moderation

Event moderation provides the bridge between the automated system and more in-depth situation analysis, which requires clean and reliable data. Moderated event structures also provide feedback to machine-learning components of the automated system. The moderation system enables selection and visualization

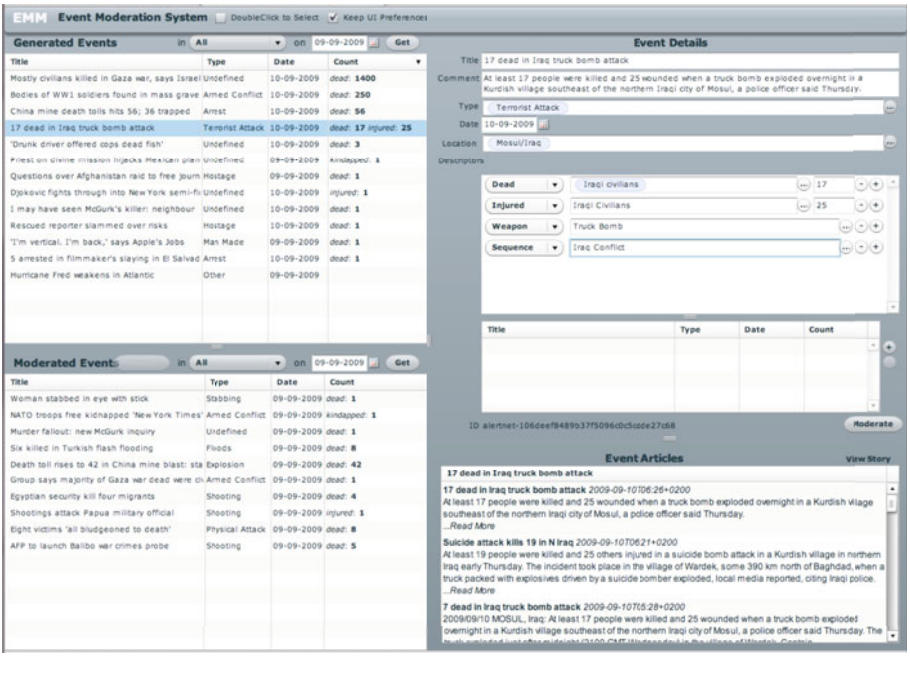

Fig. 2. Event moderation tool

of the generated events, verification of their slot values against the corresponding original document, comparison of new events with previously moderated events, in order to create event chains, and saving in a dedicated information store.

Figure 2 shows a preliminary front-end of the moderation system. The top-left pane shows automatically generated events; the bottom-left pane shows other moderated events from the same time period; the right-hand panes shows the detailed slot values of the selected event, and its link to other events and the original sources from which the events were generated.

In order to provide a near real-time geo-spatial visualisation of the 'current' events (which is important for the creation of the short-term situational picture) the system produces a stream of new event descriptions in Google's KML format every 10 minutes, which is then passed to a *Google Earth* visualisation.[5]

4 Outlook

We have outlined a system architecture based on leading-edge technologies in information extraction applied to the domain of Border Security Intelligence. The main challenge here is to minimize information overload without overlooking weak, but potentially important, signals.

Although automated news surveillance in the domain of border security is new (we are not aware of prior work in this domain), it bears some similarity to previously studied domain of epidemic surveillance, e.g., MedISys and PULS ([9]), HealthMap ([10]), and BioCaster ([11]). In fact, epidemic surveillance is in part (though not entirely) subsumed by the security domain, since the spread of epidemics impacts border security as well. The event schema used in epidemic surveillance is similar to that used in the border security domain, in that it tries to cover the victim descriptions, the cause of harm, etc. In the security domain, the schema is considerably more complex, and requires covering many similar and partly overlapping event types, as described in section 1. It therefore exhibits a higher level of complexity of text analysis.

Because the project is in an early phases, quantitative evaluations are not yet available. We believe that adopting a combination of different approaches to information extraction and aggregating this information via moderation into an information store, will provide an important step toward meeting this challenge. Technical challenges include automated detection of duplicate events, location of event boundaries and linking of event sequences. Thanks to the close collaboration with two groups of end-users at Frontex we will be able to closely monitor the extent to which these challenges are met, as well as the features and usability of the applications as they develop.

[5] A subset of the event descriptions automatically generated by the system is publicly accessible by starting *Google Earth* application with KML: http://press.jrc.it/geo?type=event&format=kml&language=en. For other languages change the value of the language attribute.

References

1. Atkinson, M., Van der Goot, E.: Near Real Time Information Mining in Multilingual News. In: Proceedings of the 18th World Wide Web Conference, Madrid, Spain (2009)
2. Piskorski, J., Tanev, H., Atkinson, M., Van der Goot, E.: Cluster-Centric Approach to News Event Extraction. In: Proceedings of the International Conference on Multimedia & Network Information Systems, Wroclaw, Poland. IOS Press, Amsterdam (2009)
3. Tanev, H., Piskorski, J., Atkinson, M.: Real-Time News Event Extraction for Global Crisis Monitoring. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (eds.) NLDB 2008. LNCS, vol. 5039, pp. 207–218. Springer, Heidelberg (2008)
4. Piskorski, J.: ExPRESS Extraction Pattern Recognition Engine and Specification Suite. In: Proceedings of the 6th International Workshop Finite-State Methods and Natural language Processing 2007 (FSMNLP 2007), Potsdam, Germany (2007)
5. Grishman, R., Huttunen, S., Yangarber, R.: Information Extraction for Enhanced Access to Disease Outbreak Reports. Journal of Biomedical Informatics 35(4) (2003)
6. Yangarber, R., Jokipii, L., Rauramo, A., Huttunen, S.: Extracting Information about Outbreaks of Infectious Epidemics. In: Proceedings of the HLT-EMNLP 2005, Vancouver, Canada (2005)
7. Zavarella, V., Tanev, H., Piskorski, J.: Event Extraction for Italian using a Cascade of Finite-State Grammars. In: Proceedings of the 7th International Workshop on Finite-State Machines and Natural Language Processing, Ispra, Italy (2008)
8. Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., Steinberger, R.: Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. Under submission to the Journal Linguamática: Revista para o Processamento Automático das Línguas Ibéricas
9. Steinberger, R., Fuart, F., Van der Goot, E., Best, C., Von Etter, P., Yangarber, R.: Text Mining from the Web for Medical Intelligence. In: Mining Massive Data Sets for Security. IOS Press, Amsterdam (2008)
10. Freifeld, C., Mandl, K., Reis, B., Brownstein, J.: HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. Journal of American Medical Informatics Association 15(1) (2008)
11. Doan, S., Hung-Ngo, Q., Kawazoe, A., Collier, N.: Global Health Monitor—A Web-based System for Detecting and Mapping Infectious Diseases. In: Proc. International Joint Conf. on NLP, IJCNLP (2008)