Language Processing and Linguistic Data in the CAPER Project

Carlo Aliprandi, Tomas By, Sérgio Paulo

LREC 2012 workshop on "Language Resources for Public Security Applications"
May 27, 2012, Istanbul, Turkey

The EU FP7 CAPER project (Collaborative information Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime) uses state-of-the-art, multi-lingual Natural Language Processing techniques to support analysis and sharing of information obtained through search and monitoring of Web data.

CAPER will allow Law Enforcement Agencies (LEAs) to share informational, investigative and experiential knowledge. A common software architecture for all linguistic processing components allows efficient combination of language-specific and language-independent modules. Cross-lingual search and query expansion is supported by multilingual lexicons and gazetteers. The project includes full support for English, Spanish, Catalan, Italian, and Portuguese, and partial support for French, German, Romanian, Russian, Basque, Hebrew, Arabic, Chinese, and Japanese.

Internet invaded our lives to such an extent that words like e-commerce, e-learning or e-government became familiar to many of us. Moreover, the widespread use of internet resulted in the emergence of highly interconnected societies. From the LEAs point of view, while such societies pose new challenges, they also provide new collaboration opportunities. That is, on the one hand, organised crime can use information technology systems to communicate, work or expand their influence to anywhere in the planet, but current tools for fighting such organisations have shown their limits and reflect the need for developing a scalable tool to track them more efficiently. On the other hand, for languages addressed in the project, internet became a massive repository of both written and spoken data (audio and video files have enjoyed increasing success as a means of information dissemination over the last few years) and, thus, comes up as a priceless source of material to be searched while attempting to detect potentially criminal activities.

CAPER’s objective is to build a common collaborative and information sharing platform for the detection and prevention of organised crime in which the Internet is used (e.g. sale of counterfeit or stolen goods, cyber crime) and which exploits Open Source Intelligence (OSINT).

The benefits of using OSINT for Intelligence Agencies (IA) are well known from previous experiments. LEAs, alike IAs in the past, are increasingly more reliant on information and communication technologies and affected by a society shaped by mass media, but more and more by the Internet and social media. This challenge can also be seen as an opportunity for LEAs. The richness and quantity of information available from open sources, if properly processed, can in itself provide valuable intelligence and help draw inference from existing closed source intelligence. The CAPER project is aimed at giving evidence that OSINT can help LEAs better understand the information they have available: the talk will detail the CAPER platform elements, giving particular evidence from a user-oriented perspective to the exploitation of Open Data Sources and the integration of mass media, closed information sources (LEA internal systems) and Social Web data sources, like Linked Data, Wikipedia, Geonames, or Yago.

A typical use case for the CAPER system is when an investigator uses the Internet as a reference source to extract structured and unstructured information of relevance to an ongoing criminal case. The investigator can establish mechanisms to automate the daily work; i.e. periodic access to predetermined sites, downloading predefined content, manage users and passwords, set alerts, identify individuals with multiple identities in the net etc. The system supports cross-lingual search and indexing, and will be linked to third-party machine-translation services for on-demand translation of texts and documents.

The project is not focused on developing new technology, but on the fusion and real validation of existing state of the art to solve current bottlenecks faced by LEAs. So the design methodology will be based on an iterative development lifecycle, with several integration steps that will be carried out.

The 6 technology pillars of the CAPER platform are:

- Open and Closed Data Sources: TV, Radio, and Information in closed legacy systems are the data sources to be mined and evaluated by CAPER, in addition to Open Internet data sources and Semantic Web data collections (Linked Data, for example, like Geonames, DBpedia or Yago).
- Data Acquisition: Depending on the information source type, different acquisition patterns will be applied to ensure acquired information is the richest possible and has a suitable format for analysis.
- Information Analysis: Each analysis module is geared towards a specific content type, i.e. Text, Image, Video, Audio and Speech or Biometric data. These modules interact with a 'Semantic mash-up' component, to interlink Semantic Web data.
- Information and Reference Repositories: source data and mined information will be stored in these repositories, separated by content type. Repositories will also store the reference images, text, keywords, biometric data etc. of interest to the LEAs.
- Interoperability and Management Application: This is the end users' workbench, the main Human Computer Interface. It will allow LEAs to collaborative create and configure their monitoring requests and analysis petitions. Through this HCI, Law Enforcement Officers will be able to create and configure their monitoring requests and analysis petitions. It will allow a structured collaboration between LEAs, who will also be able to configure their own internal closed information sources and control how and to whom the data is shared.
- Visual Analytics (VA) and Data Mining (DM): VA and DM will provide the intelligence necessary to support the output of the system. They will allow LEAs to effectively mine processed data both from Closed and Open Sources, and to further relate it to Semantic Web sources when required.

The CAPER system will be designed from a linguistically neutral point of view. This design methodology will allow linguistic analysis and speech recognition components for any language to be added in the future. LEA users will provide reference images, keywords, biometric data, and define concepts to be used in information acquisition.

The design methodology will leverage standardization of data and interchange of tools: once a data format is accepted as a standard, tools can be adapted and shared with little effort. CAPER will also standardize the processing of language sensitive information by adopting and extending KAF (Knowledge Annotation Format), a multi-layered XML format for linguistic and semantic annotation of unstructured documents that has been proven to be suitable for the purpose of Information Processing.

CAPER aims at extending data representation standards to also cope with multimedia and structured data, particularly focusing on the analysis and exploitation of Social Web Content and Semantic Web Data. To collect and integrate these different kinds of online information sources, after a thorough analysis of social and semantic data from the point of view of the CAPER users, we will develop specific components for Named Entity Recognition and Word Sense Disambiguation in the target Open Sources.

The domain-specific linguistic resources used in CAPER include extensions to general lexicons and ontologies, for normal words that have special senses in certain contexts, such as the drugs trade, as well as slang expressions not normally used in the language. There will also be interpretation/conversion modules for the special jargon and spelling conventions used in online chat rooms and similar environments.

CAPER will allow investigators to seamlessly search and integrate information in foreign languages (among those supported by the system), and thus more effectively assess

threats and interpret intelligence. There is much information readily available on the Internet and in the Mass Media, of relevance to investigations but that is missed because investigators are limited to their own language, and may not be aware of related investigations in other countries. CAPER can thus help investigators better understand the information they have available.