The EUMSSI project – Event Understanding through Multimodal Social Stream Interpretation

Jens Grivolla and **Yannick Estève** and **Eelco Herder**
Nam Le and **Kay Macquarrie** and **Raúl Marín**
Sylvain Meignier and **Maite Melero** and **Jean-Marc Odobez** and **Susanne Preuß**

Abstract. Journalists, as well as users at home, face increasing amounts of data from a large variety of sources, both in professionally curated media archives and in the form of user-generated-content or social media. This provides a great opportunity at the same time as a great challenge to use all of this information, which EUMSSI approaches by providing semantically rich analysis of multimedia content, together with intuitive visual interfaces to explore the data and gain new insights. The goal of the EUMSSI project is to provide a complete system for large-scale multimedia analysis, including a wide range of analysis components working on different media modalities (video, audio, text). Additionally, we have developed two example applications (demonstrators) that build upon this platform with the goal to showcase the platform's potential, but also to lead towards a commercial exploitation of the project outcomes. The first is a tool to help journalists when writing an article or preparing a report, while the second is aimed at viewers of TV or streaming content, providing background information and entertainment.They are currently running and publicly accessible online, incorporating hundreds of thousands of videos and news articles, as well as almost 10 million tweets.

1 Introduction

Nowadays, a multimedia journalist has access to a vast amount of data from many types of sources to document a story. In order to put information into context and tell their story from all significant angles, they need to go through an enormous amount of records with information of very diverse degrees of granularity. Manually searching through a variety of different unconnected sources and relating all the disperse information can be time-consuming, especially when a topic or event is interconnected with multiple entities from different domains.

At a different level, many TV viewers are getting used to navigating with their tablets or iPads while watching TV, the tablet effectively functioning as a second screen, often providing background information on the program or interaction in social networks about what is being watched. However, this again requires an important effort either from the provider of a specific second screen application that includes curated content, or from the viewer at home who needs to identify potentially relevant sources of background knowledge and perform appropriate searches.

In the EUMSSI (Event Understanding through Multimodal Social Stream Interpretation) project[1] we have developed a system that can help both the journalist and the TV viewer by automatically analyzing and interpreting unstructured multimedia data streams and, with this understanding, contextualizing the data and contributing with new, related information.

The huge amounts of textual content available on the web and in news archives have been "tamed" over the last years and decades to some degree through the development of efficient, scalable search engines and improved ranking algorithms, and more recently by providing the user with more directly usable insights and answers, in addition to the traditional search result lists.

Tackling multimedia data in a similar way is a complicated endeavor, requiring the combination of many different types of analysis to bridge the gap from raw video recordings to semantically meaningful insights. It is now becoming computationally feasible to analyze large amounts of media content, and the EUMSSI project leverages the project partners' expertise in speech recognition, audio and video based person identification, text analysis or semantic inference to provide an integrated platform for large-scale media analysis and exploration.

2 Project objectives

In EUMSSI we have developed methodologies and techniques aiming at identifying and aggregating data presented as unstructured information in sources of very different nature (video, image, audio, speech, text and social context), including both online (e.g. YouTube, Twitter) and traditional media (e.g. audiovisual repositories, news articles), and for dealing with information of very different degrees of granularity.

This is accomplished thanks to the integration of state-of-the-art information extraction and analysis techniques from the different fields involved (image, audio, text and social media analysis) in a UIMA-based multimodal platform. The multimodal interpretation platform continuously analyzes a vast amount of multimedia content, aggregates all the resulting information and semantically enriches it with additional metadata layers.

The EUMSSI platform can be potentially useful for any application in need of automatic cross-media data analysis and interpretation, such as intelligent content management, recommendation, real time event tracking, content filtering, etc.

The project has built two demonstrators on top of the platform aimed at showing its exploitation potential: a Computer Assisted Storytelling tool, and a Second-screen application.

Through the **Storytelling** demonstrator, we expect EUMSSI to boost productivity of the overall news production workflow (such

[1] FP7-ICT-2013-10: http://eumssi.eu

as the one used by Deutsche Welle's journalists), by providing an efficient manner to automate and integrate monitoring, gathering and filtering tasks in the news article creation lifecycle. Hence, while the journalist is editing a story, he will automatically be presented with a series of related content, suggestions, hot topics, from many sources ranked in turn by relevance, up-to-dateness, reliability, category, etc., by means of an array of visualization devices that will help him explore the data in an interactive and principled way, and discover hidden relevant information in large document collections.

With its **Second-screen** demonstrator, EUMSSI has the potential to enhance the media watching experience of the European citizens by providing them with personalized visualizations and games, aimed to further engage them and to allow them to investigate background information and related aspects, as well as share their findings in social media.

Figure 1 shows how both applications build on a common base of multimedia analysis and content aggregation/recommendation algorithms. Both demonstrators are described in more detail in section 5.

3 Project consortium

EUMSSI being a multimodal technology integration project, with a well defined product-oriented application, is well served by a consortium formed by 5 research centers, each providing their own specialised expertise in the different multimodal technologies involved, plus a public service broadcaster and a company providing solutions for the media industry. This is the EUMSSI consortium:

- The Computational Linguistics Research Group (**GLiCom**), at the Universitat Pompeu Fabra [2], has acted as project coordinator and developers of the multimodal platform, as well as working some of the text analysis tasks, such as entity linking, opinion mining and social media analysis.
- The Computer Science Laboratory (**LIUM**) the University of Maine [3] has worked on the audio analysis tasks, including speech detection and recognition, and speaker segmentation and diarization.
- The **L3S** Research Center [4], associated with Leibniz Universität Hannover, has defined the multimodal analytics common semantic framework, and has developed the Second-screen demonstrator.
- The **IDIAP** Research Institute [5], affiliated with the Swiss Federal Institute of Technology, has directed the work on the audio-visual recognition of people.
- The Institute for Applied Information Sciences (**IAI**)[6] of Saarbrücken has taken care of some of the text analysis tasks, such as keyphrase extraction and enrichment, quote detection and ASR-name normalisation.
- **VSN** Innovation and Media Solutions [7], as a provider of multimedia content production and management solutions is interested in EUMSSI's potential market and business possibilities, and has integrated part of EUMSSI's results into their own workflow tool.
- **DW** [8], as a public service broadcaster, has provided the ideal setting for reaching to professional users using the Storytelling tool for journalists.

[2] https://portal.upf.edu/web/glicom
[3] http://www-lium.univ-lemans.fr/en/
[4] http://www.L3S.de
[5] http://www.idiap.ch
[6] http://www.iai-sb.com/
[7] https://www.vsn-tv.com/en/
[8] http://www.dw.com/

4 Multimodal analytics and semantic enrichment in EUMSSI

The EUMSSI system currently includes a wide variety of analysis components (many of which leverage and improve upon existing open source systems), such as automatic speech transcription (ASR), person identification (combining voice and face recognition, OCR on subtitles for naming, and Named Entity Recognition and Linking), and many natural language processing approaches [2], applied to speech transcripts as well as original written content or social media, e.g. NER (Stanford NLP [5]), Entity Linking (DBpedia Spotlight [6]), keyphrase extraction (KEA), quote extraction, topic segmentation, sentiment analysis, etc.

4.1 Audio-visual person diarization and identification

As the retrieval of information on people in videos are of high interest for users, algorithms indexing identities of people and retrieving their respective quotations are indispensable for searching archives. This practical need leads to research problems on how to identify people presence in videos and answer 'who appears when?' or 'who speaks when?'. To this end, a video must be segmented in an unsupervised way into homogeneous segments according to person identity, like speaker diarization, face diarization, and audio-visual (AV) person diarization [1]. Combined with names extracted from overlaid text thanks to OCR techniques, AV person diarization makes it possible to identify people in videos [3].

4.2 Automatic speech recognition system

There are two main issues for the components dedicated to automatic speech recognition (ASR): one is the accuracy of the transcription, the other addresses the computation time. In the framework of the EUMSSI platform, we take benefits from interactions between multi-layered analysis components. Audio and audiovisual documents processed by the ASR component are associated to textual meta-data: title of the document, keywords, ... These textual data are processed through natural language processing algorithms to extract relevant information like named entities (cf. section 4.3). Such information can also be extracted from video analysis, for instance by applying OCR to images in order to extract relevant named entities (cf. section 4.1). By improving multi-threaded algorithms to better share GPU, CPU, RAM and disk space to process huge amount of audio for the EUMSSI platform, and by modifing the multi-pass architecture to take into account improvements provided by DNNs [4], we have succeeded in accelerating our ASR system with a factor 27 [9]. Advancements in this area in the framework of the EUMSSI project have awarded LIUM with the Innovation Radar Prize 2016 in the in the Industrial and Enabling Tech category.

4.3 Text analysis

The main goal of the text analysis component in EUMSSI is to extract structured information from unstructured textual sources, such as news, micro-blogs, audio transcriptions and video captions, and generate annotations, which are stored as metadata, and can be used by the demonstrators. In the Storytelling tool, text annotations are used for answering the big five questions journalists are interested in, namely: WHO did WHAT, WHERE, WHEN and WHY; while in the

[9] This means that the ASR system works at about $0.025 \times$ *real time*.

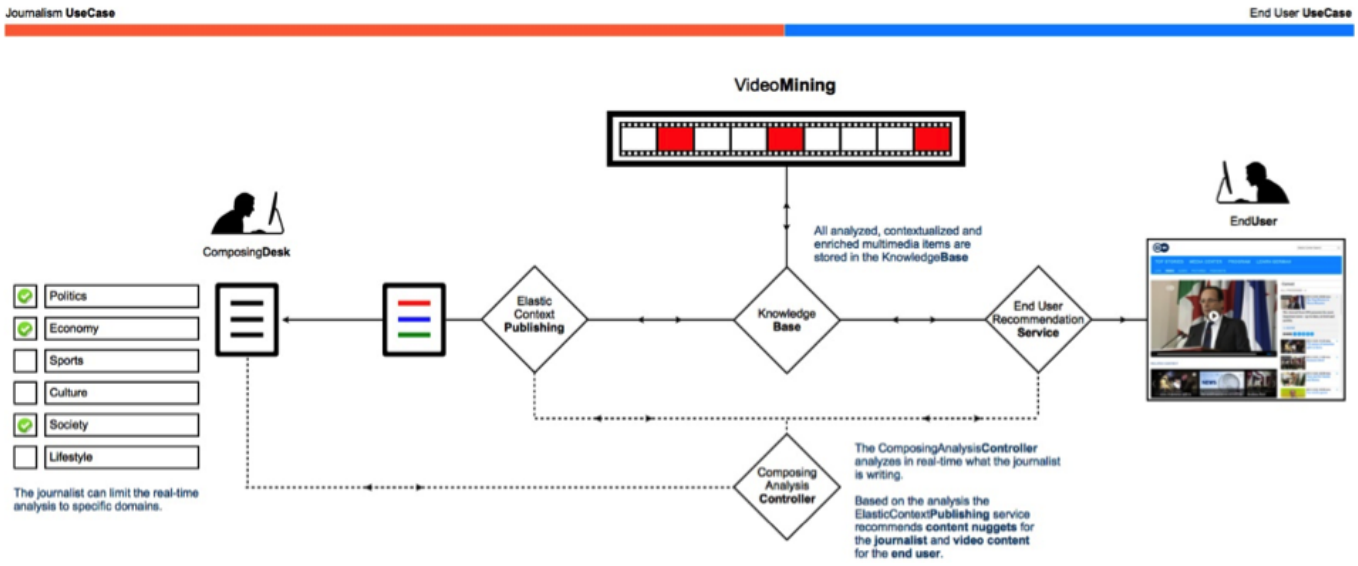

Figure 1. EUMSSI multimodal platform catering both for the journalist and the end-user's use-cases

Second screen, they are used to provide additional information and to generate quiz games. Input to this component is text from well-edited news, user-generated social-media content and transcriptions. Particular attention has been given to processing of the output of the ASR component, consisting of an uninterrupted chain of low-case words, which need to be segmented into sentences, and suitably capitalized. In order to normalize transcribed text we use a monolingual statistical machine translation system, translating from ASR output to a text which includes punctuation and capitalization. A different problem posed by ASR output are badly transcribed proper names. We have developed a normalisation module that annotates corrections of incorrectly transcribed names and adds name annotations to definite name descriptions using entity linking techniques. The module makes use of the fact that news broadcasts usually adhere to the best practice of mentioning the function or office title of important people in the immediate vicinity of the name, thus the properly recognized office title or function serves to detect ill-recognized names. Another text analysis tasks is keyphrase extraction, in which we also annotate keyphrases that, while not appearing verbatim in the text, are good descriptors of the text content thanks to enrichment patterns, such as replacing verbal and adjectival expressions with nominal expressions. Since opinions of people are of particular interest to journalists, we also detect and analyse sentiments in attributed quotations.

5 Use-cases demonstrators

Using the data stored in the system, and made available through Solr indexes, as well as on-demand text analysis services, as explained in detail in "EUMSSI: Multilayered analysis of multimedia content using UIMA, MongoDB and Solr" in this volume, a variety of applications can be built that provide access to the content and information. Two very different applications were built within the EUMSSI project, serving as technology demonstrators as well as having real-world use.

5.1 Storytelling tool for journalists

The first demonstrator is aimed at journalists, providing context and background information while working on a news article or report. This storytelling tool provides a web interface with a rich editor that allows a journalist to work on an article. Automatic on-demand analysis of the text the journalist is writing is then used to provide relevant background information.

Potentially interesting entities, such as persons, places, organizations, but also other concepts such as "Syrian civil war" or "Nuclear power" are highlighted in the editor and provide direct access to the respective Wikipedia pages. More importantly, they can also be used to find related content in the archives (including content from outside and social media sources).

It is also possible to manually define search queries, searching for specific keywords in any part of the indexed content, including text articles, associated metadata, but also the automatically generated audio transcripts, OCR output, etc. Additionally filters for languages, sources, date ranges and more can also be defined by the user.

A variety of graphical widgets then allow to explore the content collection, finding relevant video snippets, quotes, or presenting relevant entities and the relations between them. Figure 2 shows the main page with the editor on the left and an overview of visualization widgets on the right.

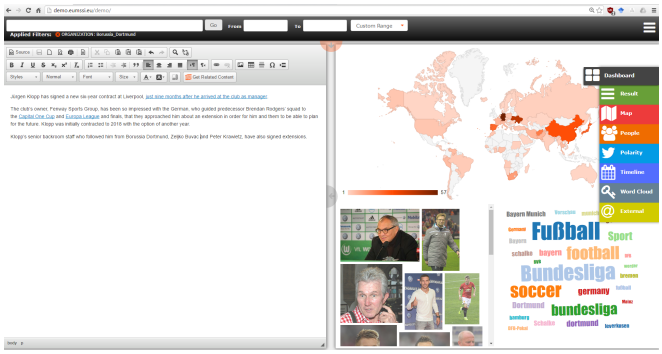

Figure 2. The storytelling web application

The different visualizations then calculate statistics over these documents to generate results such as "What persons (or countries) are mentioned most (and therefore are likely most important) in articles and videos talking about the Syrian civil war?". The widgets then al-

low the user to interact with them by, e.g., filtering the results by a specific entity, viewing the Wikipedia page, or including the visualization in the article the journalist is writing.

5.2 Second screen application

The second screen application provides both 'passive' content that provides additional information and 'active' content that invites the user to interact with it. The most basic form of additional information is the provision of text-based background information on politicians, locations, historical figures, events and locations that are featured in a news broadcast or other programs. In order to provide such content, we use the entities extracted and linked to DBpedia and enrich them using DBpedia links and relations, as well as other linked databases, such as Geonames[10]. Enrichment of extracted entities allow automatic generation of questions. DBpedia triplets such as "*Germany dbo:capital dbr:Berlin*" may be translated into the informative sentence "*The capital of Germany is Berlin*" or lead to the question "*What is the capital of Germany*".

Generating multiple choice questions also involves finding *distractors* (wrong answers that are close enough to be considered as true) [7], for example: for the question "which city is the capital of Germany", we can either search for capitals of neighbouring countries or neighbouring cities in the same country – in both cases by finding the top-n cities or countries with the smallest distance (in terms of longitude and latitude) from the current country or city - for example Schwerin, Potsdam and Magdeburg. Figure 3 shows a typical usage scenario for the second screen.

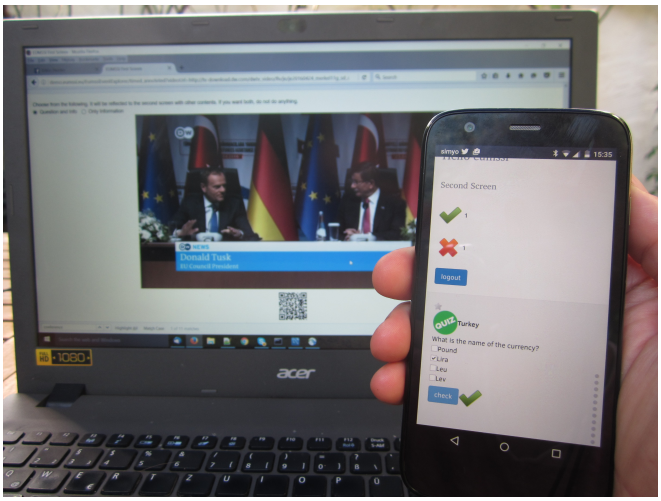

Figure 3. The second screen application running on a smartphone, showing a question related to the video running on the laptop.

5.3 Perceptive evaluations: the User Days

How were the EUMSSI prototype and demonstrators received by the audience and what feedback did the users share? The most important evaluations took place during two joined User Days with the EU projects MultiSensor and EUMSSI, organised in Bonn (2015) and Barcelona (2016). Bonn focussed on the evaluation of the features and functionalities of the prototype. In Barcelona the demonstrators "Second Screen Tool" and "Storytelling Tool" were presented and evaluated. All feedback was used for iterative optimisations of the prototype and the demonstrators.

The first User Day in Bonn aimed to collect feedback for the EUMSSI demonstrator, which was a common prototype for both use cases within EUMSSI (storytelling tool and second screen). The focus was on providing and testing functionality using a common framework and widgets rather than a nice and attractive design. The EUMSSI prototype showing various widgets for content visualisation was generally found very interesting and promising. An open panel finished the user day and lead to fruitful discussions for upcoming developments of the demonstrators.

The storytelling tool was evaluated in July 2016. Ten journalists used the demonstrator and provided valuable feedback. The demonstrator was overall perceived very well. Most of the journalists found the tool very useful, easy to use and would recommend it to a colleague or a friend. Nevertheless, the testers came up with a list of over seventy issues or remarks, which were prioritized and evaluated in the course of the following months. The storytelling tool significantly benefited from these user insights.

During a joined User Day with the EU project SAM in Bonn in July 2016, the second screen prototype was presented and evaluated. The majority of the users and testers found the prototype promising. Six user provided feedback within a questionnaire. The feedback provided was used to optimise the second screen tool in many respects. One of the most important comments addressed the need to develop an editing layer. Authors would than be able to modify and edit the automatically derived items.

The second and final joined User Day in Barcelona showcased the EUMSSI demonstrators: the second screen tool and the storytelling tool. The user day took place in September 2016 and attracted a joined audience of roundabout fifty visitors including participants from the two projects. Ten users filled out the questionnaire for the second screen demonstrator. For the storytelling tool fifteen user participated in the test and questionnaire session. Both demonstrators were well received and earned high scores in the important categories of usefulness and usability.

Based on the feedback received from the evaluation sessions both the storytelling tool and the second screen tool were able to attract and inspire users. In both questionnaires most of the testers marked, that they found the tools useful, easy to use and that they would recommend the demonstrators to a colleague and a friend.

6 Conclusion

We present here the outcome of the successful EUMSSI project, which has built a complete system for large-scale multimedia analysis, including a wide range of analysis components working on different media modalities (video, audio, text).

The different analysis modalities were developed by separate research groups, and significant improvements were achieved in different areas over the course of the project, as detailed in the respective sections. All technical results of the project are accessible to the community, both research and industrial under an open source licensing scheme, allowing a commercial exploitation of the developed resources and tools and also an unrestricted use for research purposes. The source code of the platform and many of the analysis components is publicly available at `https://github.com/EUMSSI/`. Additional documentation can be found in the corresponding wiki at `https://github.com/EUMSSI/EUMSSI-platform/wiki`.

Two example applications (demonstrators) have also been devel-

[10] `http://www.geonames.org/`

oped on the EUMSSI platform with the goal to showcase the platform's potential, but also to lead towards a commercial exploitation of the project outcomes. The first is a tool to help journalists when writing an article or preparing a report, while the second is aimed at viewers of TV or streaming content, providing background information and entertainment. They are currently running and publicly accessible online, incorporating hundreds of thousands of videos and news articles, as well as almost 10 million tweets. Links to the publicly accessible demonstrators can be found on the EUMSSI web page at http://eumssi.eu/.

REFERENCES

[1] G. Dupuy, S. Meignier, P. Deléglise, and Y. and Estève, 'Recent improvements on ilp-based clustering for broadcast news speaker diarization', in *Odyssey 2014: The Speaker and Language Recognition Workshop*, Joensuu (Finland), (16-19 jun. 2014).

[2] Richard Eckart de Castilho and Iryna Gurevych, 'A broad-coverage collection of portable nlp components for building shareable analysis pipelines', in *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pp. 1–11, Dublin, Ireland, (August 2014). Association for Computational Linguistics and Dublin City University.

[3] P Gay, S Meignier, p Deleglise, and J.-M. Odobez, 'Crf based context modeling for person identification in broadcast videos', *Frontiers in ICT*, **3**, (2016).

[4] Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin, 'Word embedding evaluation and combination', in *10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož (Slovenia), (23-28 May 2016).

[5] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky, 'The stanford corenlp natural language processing toolkit.', in *ACL (System Demonstrations)*, pp. 55–60, (2014).

[6] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer, 'Dbpedia spotlight: Shedding light on the web of documents', in *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pp. 1–8, New York, NY, USA, (2011). ACM.

[7] Dominic Seyler, Mohamed Yahya, and Klaus Berberich, 'Generating quiz questions from knowledge graphs', in *Proceedings of the 24th International Conference on World Wide Web*, pp. 113–114. ACM, (2015).