Multimedia Semantic Analysis in the PrestoSpace Project

Valentin Tablan, Hamish Cunningham, Cristian Ursu

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S1 4DP
Sheffield, UK

Abstract

PrestoSpace is a European-funded research project that aims at addressing the problem of decaying audio-visual archives throughout Europe by means of digitisation for preservation and access. One of the work areas within the project is Metadata Access and Delivery (MAD) which employs innovative methods of generating metadata for the digitised media in order to enhance the resulting archives and to ease access to the stored material. One such method is the use of automatic semantic analysis using natural language processing techniques in the process of creating analytical metadata for the preserved essence.

1. Introduction

Europe has a long-standing tradition of museums, archives and libraries for preserving its cultural heritage represented by paintings, sculptures, printed material or photographs. The 20th Century, through the advent of audio-visual technology, has started producing new types of media that need to be preserved films and several types of magnetic tapes for both audio and video material. Key events were recorded, and audio-visual media became a new form of cultural expression. These new types of material have also started to be preserved using traditional methods, by storing copies on shelves in large preservation facilities. The size of these archives is considerable The UNESCO estimates the size of the world audio-visual holdings to about 200 million hours, out of which around 50 million are in Europe. It has soon become apparent that this solution is not ideal because these new types of media suffer from chemical and physical decay (some films produce acetic acid vinegar syndrome, while all types of magnetic tapes become demagnetised over time). Another problem faced by the archives is technical obsolescence, there are fewer and fewer machines still capable of playing the older formats and keeping those functioning is becoming more and more expensive. In some cases even finding operators who are still qualified to operate those machines is becoming a problem as older personnel retires and new one is only trained for newer types of devices.

Although one possible solution would be to copy the legacy material onto newer storage formats, these operations would lead to loss of quality which is inherent to analogue processes. It is now widely accepted that the best available solution given the technical possibilities of today is to digitise the contents of the archives thus stopping the process of deterioration and freezing the quality levels at their current state. Starting from the digital copy, further transfers to new types of media will be possible with no loss of quality.

Throughout Europe large audio-visual archives, such as those managing the holdings of large broadcasting organisations, have already started the process of digitisation for preservation. This is an expensive process, the average cost for transfer from old to new media using the most cost-effective current technology is around 500 euros/hour a finding of the now ended Presto project. Budgetary restrictions mean that the current rate of transfer to digital for the most archives is not fast enough to ensure the preservation of the entire back-catalogue before it falls prey to decay. While an increase in budget would solve the problem, expecting that would be unrealistic. This is why the PrestoSpace project is addressing the issue starting from the other end by finding a way to lower the costs associated with the preservation process.

Better preservation and access also leads to increased reuse potential for the legacy audio-visual material, enabling media organisations to

extract more value from their holdings. This extra value can be returned as extra investment for preservation activities speeding up the digitisation effort and thus helping to save even more material from being deprecated and forgotten.

The next sections of the paper provide an overall view of the organisation of the PrestoSpace project, a more detailed view of the Metadata Access and Delivery work-area of the project and then it centres on the work done for automatic semantic analysis.

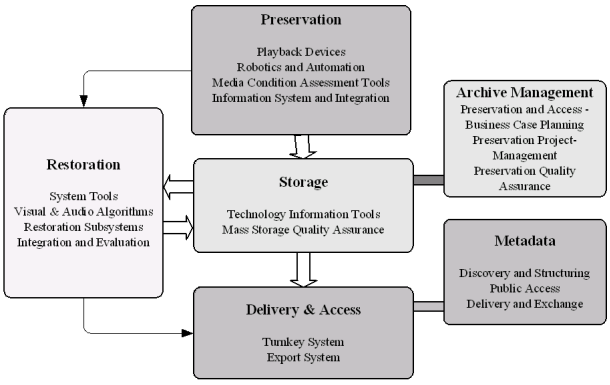

Figure 1: Overall organisation of the PrestoSpace project

2. The PrestoSpace Project

Audiovisual archiving is a complex and multidisciplinary domain spanning such diverse fields as chemistry, physics, signal processing, robotics and artificial intelligence. The challenge is to integrate partners of all domains representing the variety of competencies needed. The Project therefore brings together participants including 8 archive institutions, most of them representing the archives as well as their R&D departments, 3 applied R&D institutions, 6 academic institutes and 15 industrial partners.

The partners have analysed the different steps of preservation work towards access according to archives practises and to the required skills and technologies. The main production chain is the migration from analogue to digital material, including stock evaluation, identification and selection, the digitisation process and its control, the restoration, the storage and the production of content information (metadata) allowing for access and delivery.

Figure 1 depicts the projects work areas as well as the way they interact through the general work-flow. The Preservation work area is at the start of the chain and deals with the digitisation of the analogue media. All further processing is then performed on the digital copy. This area is concerned with robotics, hardware and software facilities dedicated to automating the process of digitisation to the highest possible level with a view to reducing the associated costs.

The next work area is Storage and Archive Management which aims to supply archives of all sizes with the required information and management tools so they can plan their own preservation process and keep track of their assets and the costs involved in moving from an analogue to a digital storage solution.

The Restoration work area provides an integrated restoration system that will be capable of analysing the digitised material, identify defects and apply the most appropriate software algorithms for correction. This will be a scalable system aimed at high throughput for a good enough quality at a low cost.

Metadata Access and Delivery MAD provides solutions to the problem of finding and making accessible the material preserved in the archives. This entails first generating metadata information describing the audio-visual items, by transferring the existing legacy metadata from the old analogue archives and by generating new information as a result of various content analysis processes and semantic analysis. Once the metadata exists, efficient retrieval methods are provided that combine the power of traditional information retrieval techniques with novel search methods based on conceptual search over the semantic metadata.

In order to help reducing the preservation costs, a factory approach is taken when the overall work-flow is designed. The various work areas interact creating a preservation chain that provides high throughput and good quality at a cost as low as possible. Human interaction is avoided wherever it can, being replaced by robotics and algorithms that can take decisions based on the setup of the system and the set of requirements.

2.1. The MAD Documentation Platform

Digital material can only be effectively accessed if metadata describing it is available in some sort of cataloguing system. Production of such metadata currently requires manual annotation by an archivist, a time consuming and hence costly task. The MAD platform is responsible for automating the documentation process as much as possible by employing state of the art algorithms for content analysis and semantic analysis based on human language technologies (HLT) in order to derive metadata. Depending on the level of detail required for the resulting metadata, some human intervention may still be necessary but that is kept to a minimum and the automated processing is still employed as a helping tool even when a human archivist is authoring the metadata.

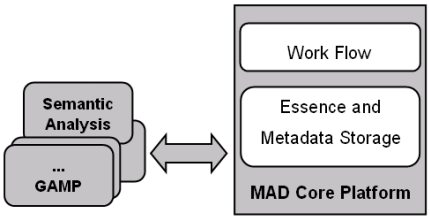

Figure 2: The MAD work-area architecture.

The architectural organisation of the MAD platform is illustrated in Figure 2. The system comprises a core element (the MAD core platform) and a set of configurable Generic Activity MAD Processors (GAMPs). The core platform handles the work and data flow through the system and provides services for storage of the essence and metadata files. The essence is stored as a file containing the digitised version of the audio-visual item. Several other representations such as a low-resolution preview version or separate audio channels or video track can be derived as required by the processes applied. The metadata is stored as XML files using a schema centred on the concept of Editorial Object (EDOB) which can represent either a programme or a unitary section of one. All temporal decompositions of EDOBs such as time-aligned speech transcripts or visual analysis metadata are represented using MPEG7. The storage for metadata files provides versioning support through SourceJammer, a CVS-like, open-source Java system, wrapped up as a web service. This provides some sort of transactional support by allowing rollbacks for failed operations that need to be rerun.

All the processing within the MAD platform is performed by the various GAMPs which implement algorithms for metadata creation or provide services to the other GAMPs such as multimedia de-multiplexing or the generation of automated speech-to-text transcripts. The two main metadata creating GAMPs are the Audio-Visual Content Analysis one which identifies keyframes, scene or shot boundaries and produces other technical metadata and the Semantic Analysis GAMP which generates conceptual metadata starting from the speech transcript or other textual sources available (such as subtitles or closed captions). A web-based interface allows the operator to configure the work-flow and the individual GAMPs as well as to monitor the state of the system at any point and to intervene for solving any problems arising.

3. The semantic analysis process

The Semantic Analysis GAMP uses textual sources such as automatic speech recognition or subtitles or the output of a video OCR process running over the titles or the credits section in order to derive conceptual information about a multimedia item. This metadata can then be used to perform new types of searches within the archives allowing the retrieval of material based on conceptual queries using semantic entities like person names, geographical locations or commercial organisations and the relations between them.

The main challenge that needs to be addressed is posed by the poor quality of the text that results form speech recognition of OCR. The process of automatic extraction of semantic meta-data from text is a complex one and needs good quality text as input in order to provide usable results. While some analysis can be performed on the type of text that is obtained from the multimedia file, the amount of information that can be extracted is rather limited. One way to solve this problem is to try and find better textual sources relating to the multimedia material. One source of good quality text is the Internet and, especially in the

case of news broadcasts, one could hope to find web pages that are closely related to the events mentioned in the media item. This is the hypothesis that our system is based on.

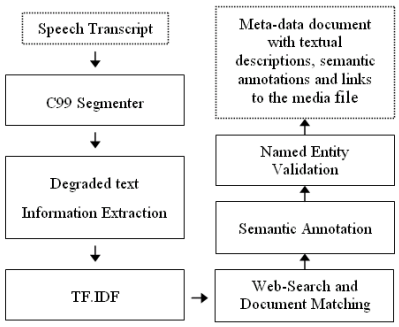

Figure 3: Architecture of the semantic analysis pipeline.

Figure 3 shows the pipeline used for performing semantic analysis. The speech transcript is first segmented using the C99 segmenter (Chaisorn et al., 2003) which uses lexical similarity to decide where a segment split should be inserted. For each of the resulting segments a customised Information Extraction system is used that is capable of extracting some entities from degraded text. Next keyphrases are extracted from each of the segments using the TF.IDF technique. These are then used to perform web searches that find pages that might be related to the content in the multimedia file. From all the candidate pages, the system selects the ones that exhibit a level of similarity with the transcript that is higher than a pre-defined threshold. The pages thus obtained represent high quality text associated with the multimedia item being analysed and can be used to extract semantic meta-data in the form of named entities and relations. This is done using the KIM platform (Popov et al., 2004) that was designed especially for processing web pages.

Once the semantic entities in the related web pages have been detected, a method for merging and assigning confidence scores for these results back in the transcribed text is required. The idea is to augment the entities found in the ASR transcript with the information extracted from the corresponding entities identified by KIM. Firstly, the stemmed entities from the ASR transcription are matched against the stemmed content of the ones in the related web document. If more than half of their content is found among the one of the entities found by KIM, the highest confidence score is assigned to both entities. The semantic information carried by the web entity, is then transferred to the one in the transcript obtaining both temporal and conceptual accuracy. Secondly, the remaining KIM entities are matched against the stemmed content of the ASR transcript and for every match, the semantic content of the KIM entity is transferred to the topical segment containing the text region of the match.

In order to browse and validate the results, a simple user interface has been implemented that can display the media content, the ASR transcript, the links to the related web pages and details about the entities discovered. A screenshot of that interface is presented in Figure 4.

Figure 4 shows the results of running the system over a news broadcast from 2002. One of the leading stories at the time involved a person named “Paul Burrell”. What is interesting about him is that because of the non-standard surname, the speech system failed consistently to recognise the name – despite the fact that it is mentioned 4 times in the story it is never recognised correctly. The semantic analysis system however, manages to get the correct named entity because it is extracted from the web page where it is spelt correctly and is then matched to the partial entity *paul* extracted from the transcript. The interface shows the transcript containing the text “paul bar all” while the entity details pane shows the spelling “Paul Burrell”.

4. Acknowledgements

The research for this paper was conducted as part of the European Union Sixth Framework Program project PrestoSpace (FP6-507336). We would like to thank the BBC archives for providing information about their annotation process and for making broadcast material available to us.

5. References

Chaisorn, T., L.and Chua, C. Koh, Y. Zhao, H. Xu, H. Feng, and Q. Tian, 2003. A two-level multi-modal approach for story segmentation of large news video corpus. In *Proceedings of the TRECVID Conference*.

Figure 4: User interface used to validate results.

Popov, B., A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, 2004. KIM – Semantic Annotation Platform. *Natural Language Engineering*.