SEMANTIC ANALYSIS FOR TOMORROW'S AUDIO-VISUAL DIGITAL ARCHIVES

Cristian Ursu Valentin Tablan Hamish Cunningham

Department of Computer Science,

University of Sheffield

Sheffield, S1 4DP, UK

+44 114 22 21800

{cursu, valyt, hamish}@dcs.shef.ac.uk

Keywords: audio-visual archives, semantic annotation, natural language processing

Abstract

PrestoSpace[1] 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 audiovisual media became the 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 public broadcasters, 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/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 better reuse of the past audio-visual material, enabling large to small media businesses 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 media 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

[1] http://www.prestospace.org

of the Metadata Access and Delivery work-area of the project and then it centres on the work done for automatic semantic analysis within this work area.

2. The PrestoSpace project

Audiovisual archiving is a complex and multi-disciplinary 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 university institutes and 15 industrial partners.

The partners have analysed the different steps of preservation work towards access according to archives practices 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.

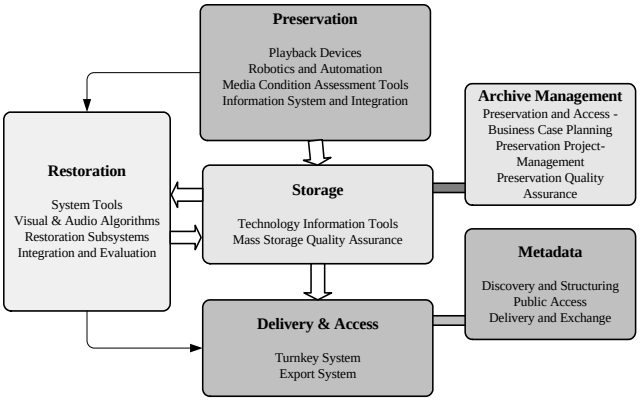

Figure 1: Overall structure of the PrestoSpace project

There is a strong motivation to achieve this work in a continuous way - for technical and economic reasons and to limit, when possible, human intervention. Thus it is expected to collect all information available during the process, including assessment of equipment and technical quality. This approach minimises poor playback, material damage and any limitation on later use of the results.

Figure 1 depicts the project's work areas as well as the way they interact through the general workflow. The Preservation work area is at the start of the chain and deals with the digitisation of the analogue media. All further processing is the performed on the digital copy. This area is concerned with robotics, hardware and software facilities dedicated to automating the process of digitization 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 workflow 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.

3. 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.

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 re-run.

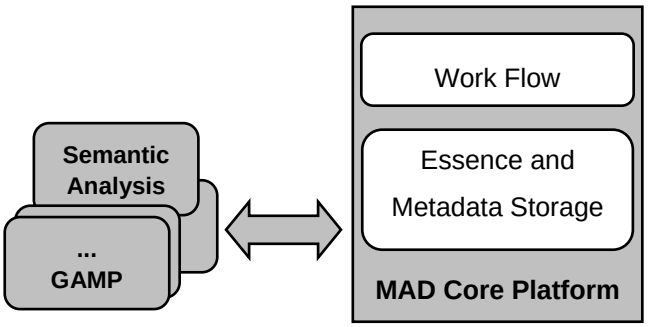

Figure 2: The MAD architecture

The workflow engine is based on OpenFlow – an open-source tool that runs inside the Zope content management system.

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 workflow 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.

4. The Semantic Analysis processor

The Semantic Analysis GAMP uses textual sources such as automatic speech recognition or subtitles in order to derive conceptual information about a multimedia item. This type of 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.

While this methodology will be used for all types of material processed by the MAD platform, the first prototype was developed for news broadcasts. This choice was motivated by the availability of test material provided by the BBC archives (one of the partners in the PrestoSpace project) and by the higher level of existing expertise for performing natural language processing on this type of texts. Once the best practice has been crystallised, tested and proven on news, the same principles will be applied to new types of material.

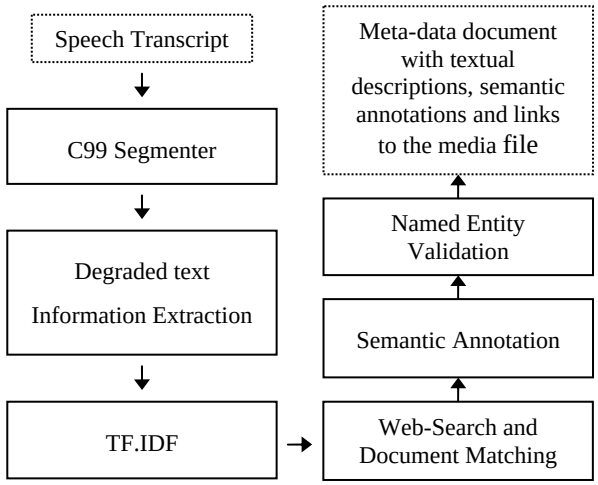

Figure 3: Architecture of Semantic Analysis processor

The Semantic Analysis system can be divided into six modules, as shown in Figure 3. These modules must execute sequentially as each builds on the output of the previous one. The initial components are firstly a segmentation module that divides the broadcast into segments corresponding to individual news stories, secondly a customized annotation engine adapted to work over speech transcripts, thirdly a module that finds key words for each story, and a fourth module that finds web pages that report the same story as that in the broadcast. The next module performs information extraction and semantic annotation on the text of the web pages, thus allowing the named entities in the broadcast story to be correctly identified. The final component matches the entities found by the semantic annotator with the ones found by the first module or with pieces of text from the transcript. The output comprises a list of entities mentioned in each story, a headline and a short summary for each segment where those could be extracted from a matching web page.

Although Blinkx and Google have recently launched television search engines, their systems rely heavily on simple full-text-search, and do not use the inherent structure of broadcasts to aid in the retrieval process. Previous work has adopted similar information extraction technologies to those used here (see for example [16]), but our work is novel in both the use of web-based content augmentation and in the use of semantic annotation [14].

4.1. Topical Segmentation

In order to produce useful descriptions and indexes of particular topics, a good segmentation of the initial broadcast is required. Such a task is far from being trivial because there are no reliable markers of boundaries in the broadcast material. For the audio and video material there is potential for exploiting audio and video cues to help in segmentations.

However, it is also possible to segment based on the language that is contained in the media. Many approaches to topical segmentation have used textual cues but often they have been in conjunction with cues from non-textual sources.

<s> thousands of local people have been protesting at the way the authorities handle the operation <SIL> can marshal reports from the coastal village of mitch a <SIL> crash patches of oil has started to perk up and dalglish encased <SIL> are the main body of the thick blue is several miles offshore <s> dozens of volunteers working on a beach in which at <SIL> having to use a blade to carve up the thick

Figure 4: Example of the Speech Recognizer's Output. (The story, from BBC Radio 4 news, reports an oil spill.)

The system described by Chaisorn et al [2] segments television news using a wide range of cues, such as analysis of the television picture, and the captions that appear on it. These cues supplement the segmentation which is based on the analysis of a transcription produced using a speech recognizer. However there are also some drawbacks to this approach as the visual cues are only available for the visual media. It is uncertain how well the system would perform on news in a different format or how great is the cost of retraining the system. For these reason, a system that doesn't need visual cues or training and produces acceptable results, would be preferable to use.

A simple segmentation technique applicable to the ASR transcriptions would be to identify certain words or phrases which tend to occur at the boundaries of the topical segments. There are some indicative key phrases such as "*and [name of reporter] thank you*" or "*back to the studio*", but, because of recognition errors, such cues do not occur reliably, suggesting that such an approach would not produce good results.

There is a considerable amount of published work in the area of textual segmentation. Franz et al [6] trained a decision tree model which looked at the distribution of nouns on both sides of candidate boundaries, in order to find words and bigrams that were indicative of segment boundaries. Mulbregt et al [13] describe a system that used HMMs to detect boundaries, in which the hidden states of the HMM represented topics, and the observations with respect to the HMMs were words or sentences. This system was trained on a 15 million word corpus, in which the topic boundaries were marked. This is problematic in the case where no training corpus is available. Kehagias et al [11] used product partition models to achieve text segmentation. Their system was also trained on a corpus in which topic boundaries were marked, so that the parameters of the model could be set.

A totally different approach, was taken by Kan et al [10], who used the concept of lexical chains to locate boundaries between topics. Chains were identified between all the occurrences of repeated noun-phrases. This method has a big advantage over the techniques described above, in that it does not require training data, meaning that it can be deployed even when no training data is available.

The majority of topical segmentation techniques use measures of lexical cohesion to determine which sections of the text are about the same topic. Such techniques require predefined input segments, which would typically be sentences or paragraphs. Stop words (the most common words in the language) are then removed from the input segments, and the remaining words are usually stemmed (any affixes are removed). A comparison is then made of the extent to which neighbouring segments contain the same stems. Where there is a high degree of overlap between neighbouring segments, it is unlikely that they will be about different topics, but when there is little similarity they probably should be placed in different segments. Such methods segment based on an analysis of the text as a whole, and so should be relatively robust even when words at topic boundaries have been misrecognised. It was therefore decided to proceed using such an approach, and the specific segmentation algorithm used was the C99 segmenter [2].

Kehagias et al [11] report that C99's performance was not greatly below that of their own segmenter, which relied on training data, and which they claimed achieved the highest performance of any segmenter reported in the literature. (C99's performance on the test corpus used by Kehagias et al was 13.0% in terms of Beeferman's Pk metric [1], compared to 5.38% for their own algorithm. Lower Beeferman scores indicate higher performance.) Therefore there seems to be little justification for using a system that requires training, when comparable results can be achieved without the need for training data.

C99 calculates the similarity between input segments using the cosine measure (see for example Jurafsky and Martin [9]). This measure gives high scores to segments that contain mainly the same words, and in which those words occur with similar frequencies, and lower scores to documents that contain different words, or in which the words occur with different frequencies. Initially, all of the input segments are grouped together in one big super-segment, and segmentation then proceeds by breaking this initial segment into successively smaller super-segments. C99 can decide when the optimal segmentation has been achieved, and so there is no need to specify how many topical segments should be created.

C99 has been found to work well on the BBC news programs, though it often fails to create separate topical segments for very short stories (which are often covered in one or two sentences). Headlines also create a problem, as C99 will often break these into topical segments in a fairly arbitrary manner, usually resulting in several stories appearing in each topical segment. We will see below that the document matcher can compensate for such errors.

4.2. Degraded text Information Extraction

Although the output of the speech recogniser contains many errors (the average error rate on the BBC news broadcasts was found to be 30% but it can vary from 10% to 90% for localised segments) some basic Information Extraction processing can be performed nevertheless. This is useful as it

identifies candidate entities that can later be matched with the ones reliably identified by the more complex Semantic Annotation module providing an anchoring in the timeline for the entities found in the web pages. We used the MUSE Information Extraction engine [20] which was previously reported to achieve accuracy close to 50% on speech data. Figure 4 shows a fragment of the speech transcript processed by the system, demonstrating that while the transcription is generally intelligible, its quality is poor, and it would not form an acceptable basis for semantic annotations intended to be used for conceptual retrieval.

4.3. Key-phrase Extraction

Once the segmenter has segmented the ASR transcript so that we have a section of text for each story in the original broadcast, the next stage is to try to find key words or phrases that are representative of the story. These key phrases can then be used as the basis of a search on the Web for pages that report the same story. Therefore the aim of the key phrase extractor component is simply to extract several phrases that are likely to occur on relevant web pages, and which are unlikely to appear on unrelated web-pages. Whether or not these phrases are coherent as far as human readers are concerned is irrelevant.

There is a significant literature on the subject of key-phrase extraction, and the closely related topic of title generation. Jin and Hauptmann [8] describe an algorithm that automatically generates titles for transcribed broadcast news (the titles could be seen as a kind of key phrase), and Turney [19] describes a system that aims to extract key-phrases for use in indexing or summarizing documents. However, both these systems have a major, drawback, in that they require training on large collections of documents on which key phrases or titles have already been marked. No such collection was available for the BBC data, and, in general, NLP systems do not perform well if the data on which they are trained is not similar in topic and structure to the data on which they are applied.

Both Jin and Hauptman and Turney's systems used *term frequency inverse document frequency* (TF.IDF) as a central part of their mechanism for selecting key-phrases. This method looks for phrases that occur more frequently in the text under consideration than they do in the language as a whole. This is likely to find phrases that are characteristic of the text, while ignoring phrases that occur frequently in the text simply because they are common in the language as a whole. It requires training data in order to determine how common each phrase is, but this training data need not be marked up with any annotations, and so the ASR transcripts of the broadcasts could themselves be used as the training data. TF.IDF is really a family of methods, because there are several different formulas that can be used to calculate TF.IDF scores for each phrase, and various criteria for deciding what constitute candidate phrases. The chosen method was the same as that used for KEA, and reported by Frank et al [6].

Firstly, any sequence of words up to length six was considered to be a 'phrase', except that phrases that began or ended on stop words were ignored. The transcripts of 13,353 news broadcasts were used for collecting phrase frequency data. Each word was stemmed, and how many times each phrase occurred in the training data was determined. However, because the number of phrases up to length six occurring in the training data was so large, once more than 300,000 distinct phrases had been observed, those with the lowest frequencies were removed until there were less than 100,000 remaining. This process was repeated every time the number of phrases stored exceeded 300,000.

Key phrases were extracted for each topical segment. Firstly the frequency of each stemmed phrase in the topical segment was found, and if it occurred two or more times, its TF.IDF score was calculated using Equation (1), in which N is equal to the number of transcripts in the training data, n is the number of documents in which the phrase occurs, t is the frequency of the phrase in the topical segment, and p is the number of candidate phrases in the current document. (n would be zero for phrases not recorded in the phrase frequency data.)

$$\text{TF.IDF Score} = \log\left(\frac{N}{n+1}\right) \times \frac{t}{p} \qquad (1)$$

The four phrases with the highest TF.IDF scores were then taken to be key-phrases for the topical segment. (Sometimes fewer than four key-phrases would be extracted, as fewer than four phrases would occur at least twice in the segment.) This technique was usually successful in finding appropriate key phrases, although often there would be fragments that could not really be described as phrases, such as '*minister nick*' instead of '*local government minister nick raynsford*', or inappropriate phrases were returned, sometimes including mis-recognized words, or words that would not help to identify the topic of the story. However, we will see below that such errors do not seriously affect the accuracy of the meta-data produced as the final product of the annotation system.

4.4. Search of the Related Web Documents

The purpose of extracting key-phrases was so they could be used to search for web pages reporting the same story on the Web. Searches were conducted using Google, which was accessed via the Google Web API[2]. Searches were restricted to the news section of the BBC, Times, Guardian and the Telegraph websites. An attempt was made to restrict the search only to the day of the original broadcast, or the day before, by adding a term specifying either of these dates in the format that they appear on each of the websites. In the case of the BBC for example, something like: *"1 December, 2004" OR "30 November, 2004"*. The dates of the broadcasts were known, as they were always input to the speech recognition component of the system. This technique was usually successful in restricting the dates of the web pages returned,

[2] See http://www.google.com/apis/

but sometimes the search would also return web pages containing references to events that happened on those dates, which were sometimes published years later.

Up to five searches were performed for each topical segment. The first would include the two key phrases with the highest TF.IDF scores, while the other four would each search with just one of the four key-phrases extracted. For each search, the first three URLs returned by Google were retrieved..

Examination of the results showed that the first URL returned often pointed to the web page that most closely matched the story, and when it did not, then often the second or third URL returned did. In the cases where no appropriate URL was returned for a story, this was most usually because the segment contained two separate stories, or corresponded to a part of the news broadcast containing headlines, in which case no single web page would be appropriate. However, even when a correct URL was returned, a procedure was still needed for determining which of the URLs it was. This was achieved by the addition of a document matching component.

The document matching component loads the documents found by Google, starting with those found using the first two key-phrases, and then subsequently those found using the first, second, third and finally the fourth key-phrase. The text of each web page was then compared to that of the input segment, and if they were sufficiently similar then the web page would be associated with the topical segment, and no more web pages would be considered for that segment.

4.5. Semantic Annotation

Up to the present point, the metadata we have created for news stories has been in a textual format. This could allow searches for stories whose metadata contains particular text, which would work in much the same way as an ordinary search engine. However, it would be better if it were possible to perform more specific searches, which could make reference to specific unique entities, such as people or countries. For this purpose, the KIM knowledge and information management platform [15] was used. KIM, in common with most of the other components of the system, is based on the GATE natural language processing framework [4]. It produces metadata for the Semantic Web in the form of annotations with respect to a basic upper-level ontology called PROTON[3] encoded in OWL[4]. These annotations can be associated with particular words or phrases in the documents.

KIM identifies entities in texts using a number of techniques. Firstly, and most simply, text is looked up in gazetteers (lists of particular types of entity, such as names of cities, or days of the week). More complex approaches make use of a shallow analysis of the text, and pattern matching grammars. KIM combines the results of all these methods in order to produce more annotations, and more accurate annotations, than could be extracted using a single method alone. Kiryakov et al [12] report results showing that KIM achieves a meta average F1 score of 91.2% when identifying dates, people, organizations, locations, percentages, and money in a corpus made up of UK, international, and business news.

The most important difference between KIM and the majority of the other information extraction systems is that KIM can identify unique entities, and it can annotate all occurrences of the same named entity with the same URI. This will assign the entities to a position in an ontology that is a formal model of the domain of interest. Each entity will also be linked to a specific instance in the knowledge base for that entity (which will be added if it does not exist at that point), and it will also be linked to a semantic entity description. KIM will also try to determine when the same named entity appears in two different formats, so that, for example, *New York* and *N.Y.* would both be mapped to the same instance in the knowledge base. For a more detailed description of the information extraction capabilities of KIM, see Popov et al [14].

An inspection of the annotations produced by KIM when run over the related web pages and a comparison between them and the corresponding broadcast recordings reveal that many of the named entities found by KIM actually do occur in the broadcasts, and that we find more relevant named entities by annotating the web pages than we do by annotating the transcripts themselves. An example of a web page that has been annotated by KIM is shown in Figure 5 where we can see the features associated with the document, recording details such as its headline and the media file. We can also see a part of the text on the page, in which two organization and one person annotation have been marked.

4.6. Named Entity Verification

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. This module implements an algorithm that performs this matching, assigning confidence scores in the process.

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. At this moment, the link between the media file and the semantic information repository is complete. An example of such a link can be found in

where the transcript contained the text "*.,.[SIL] paul bar all had told him…*". The phrase "*bar all*" is a typical mistake made by the ASR for the name "*Burrell*" due to the similarity in sound of the two. Using the related web stories processed by KIM, the ASR entity "paul" with a link in the media file,

[3] http://proton.semanticweb.org

[4] http://www.w3.org/TR/owl-features/

has been matched and enriched with semantic information pointing to the actual person involved in the news, Paul Burrell which was correctly recognised by KIM on the web page.

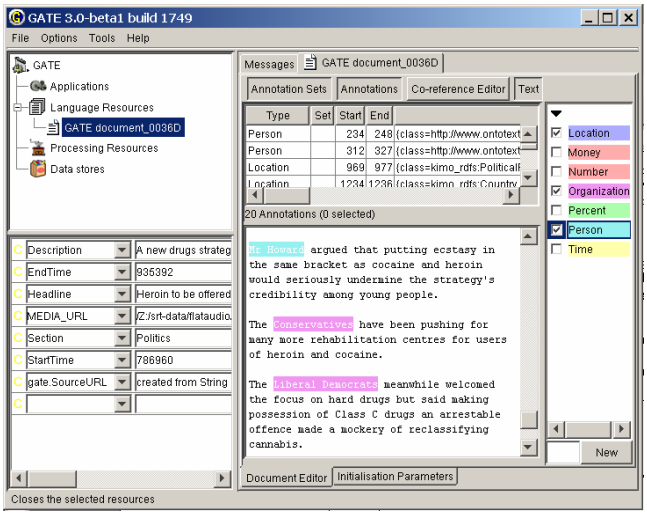

Figure 5: An example of a Story Index Document that has been annotated by KIM, displayed in the GATE GUI.

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.

5. RichNews interface

In order to be able to inspect the output of the semantic analysis system we have developed a simple web-based browsing interface. This was named *RichNews* and a screenshot is presented in

.

The results produced by the Semantic Analysis GAMP are exported to XML and then an XSL style sheet is applied in for converting that to HTML that is displayable in a normal web browser.

The interface comprises four interconnected panels. The upper left corner shows a media player for the audio-visual document used as input. The upper right corner contains the segmented ASR transcript where alternate background colours mark the different segments. The annotated entities are displayed in different highlight colours according to their type. The lower left corner shows the entity profile for the current entity as contained in the KIM knowledge base. Finally, the lower right pane displays summary information about the current topical segment or story. All these panes are synchronised and mouse clicks inside different segments or over highlighted entities cause the media player to jump to the right time in the media file and the other panes to be repopulated with information.

6. Evaluation

The evaluation of the system's performance was conducted by first playing nine broadcasts, and noting the stories that occurred in each. The programs used in the evaluation were BBC Radio 4's *The World at One*, which is a half hour long daily national news programme.

Figure 6: The media rich interface used to browse and validate the results produced by RichNews

	Correct	Incorrect	Precision (%)	Recall (%)	F1
Strict	25	2	92.6	37.9	53.8
Lenient	27	0	100	40.1	57.2

Table 1: Results of the Evaluation on 66 News Stories.

Table 1 shows the results under two conditions. In the first condition, *strict*, annotation was only considered successful if the correct story was matched, but in the second, *lenient*, it was considered correct if a closely related story was matched. The nine broadcasts considered (making a total of approximately four and a half hours of material), contained a total of 66 news stories. The results of the evaluation show that the system achieved very high precision, but that recall was much lower. This is by design as the system is intended to work mainly in a fully automated manner which makes it important that the output is correct even if that means the coverage is not the best. Most archives will employ human specialists for annotating the material they see as the most valuable but the high cost of this means that this will only be true for a small fraction of their holdings. This system will help add some annotation in a fully automated manner to some 40% of the remaining content which can be a significant amount of data for big archives.

7. 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.

8. References

[1] D. Beeferman, A. Berger and J. Lafferty. "Statistical models for text segmentation", *Machine Learning*, **Volume 34**, pp. 177-210, (1999).

[2] L. Chaisorn, T. Chua, C. Koh, Y. Zhao, H. Xu, H. Feng and Q. Tian. "A Two-Level Multi-Modal Approach for Story Segmentation of Large News Video Corpus." Presented at TRECVID Conference, (Gaithersburg, Washington D.C), (2003).

[3] F. Choi. "Advances in domain independent linear text segmentation", *In Proceedings of NAACL*, (Seattle, USA), pp. 26-33,(2000).

[4] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan. "GATE:A framework and graphical development environment for robust NLP tools and applications", *In proceedings of ACL* (Philadelphia, USA), (2002).

[5] N. Dimitrova, J. Zimmerman, A. Janevski, L. Agnihotri, N. Haas, D. Li, R. Bolle, S. Velipasalar, T. McGee and L. Nikolovska. "Media personalisation and augmentation through multimedia processing and information extraction", *Personalized Digital Television*, pp. 201-233, Kluwer Academic Publishers, Dordrecht, Netherlands, (2004).

[6] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin and C. G. Nevill-Manning. "Domain-specific keyphrase extraction", *In Proceedings of IJCAI*, (Stockholm, Sweden), pp. 668-673, (1999).

[7] M. Franz, B. Ramabhadran, T. Ward and M. Picheny. "Automated transcription and topic segmentation of large spoken archives", *In Proceedings of Eurospeech* (Geneva, Switzerland), pp. 953-956, (2003).

[8] R. Jin and A. G. Hauptmann. "A new probabilistic model for title generation", *In proceedings of COLING* (Taipei, Taiwan), (2002).

[9] D. Jurafsky and J. H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Prentice Hall, Upper Saddle River, NJ, (2000).

[10] M. Kan, J. L. Klavans, and K. R. McKeown. "Linear segmentation and segment significance", *In Proceedings of the 6th International Workshop on Very Large Corpora*, (Montreal, Canada, August), pp. 197-205, (1998).

[11] A. Kehagias, A. Nicolaou, V. Petridis and P. Fragkou. "Text Segmentation by Product Partition Models and Dynamic Programming", *Mathematical and Computer Modelling*, **Issues 2-3**, pp. 209-217, (2004).

[12] A. Kiryakov, B. Popov, I. Terziev, D. Manov. and D. Ognyanoff. "Semantic annotation, indexing and retrieval", *Journal of Web Semantics*, **2, Issue 1**, (2005).

[13] P. V. Mulbregt, I. Carp, L. Gillick, S. Lowe and J. Yamron. "Text segmentation and topic tracking on broadcast news via a hidden Markov model approach", *The 5th international conference on spoken language processing* (Sydney, Australia), (1998).

[14] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov and A. Kirilov. "KIM – a semantic annotation platform for information extraction and retrieval", *Natural Language Engineering*, **10, Issues 3-4**, pp. 375-392, (2004).

[15] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, Al. Kirilov and M. Goranov. "Towards semantic web information extraction", *In proceedings of ISWC* (Sundial Resort, Florida, USA), (2003).

[16] M. Przybocki, J. Fiscus, J. Garofolo and D. Pallett. "1998 HUB-4 information extraction evaluation", *In Proceedings of the DARPA Broadcast News Workshop* (Herndon, VA), pp. 13-18, (1999).

[17] T. Robinson, D. Abberley, D. Kirby and S. Renals. "Recognition, indexing and retrieval of British broadcast news with the THISL system", *In Proceedings of Eurospeech*, (Budapest, Hungary), pp. 1067-1070, (1999).

[18] T. Robinson, M. Hochberg and S. Renals. "The use of recurrent networks in continuous speech recognition", *Automatic speech and speaker recognition – advanced topics*, Kluwer Academic Publishers, Boston, pp. 233-258, (1996).

[19] P. D. Turney. "Coherent keyphrase extraction via web mining", *In Proceedings of IJCAI* (Acapulco, Mexico), pp. 434-439, (2002).

[20] D. Maynard, K. Bontcheva and H. Cunningham. "Towards a semantic extraction of named entities", *Recent Advances in Natural Language Processing*, (Bulgaria), (2003).