

# Semantically Enhanced Television News through Web and Video Integration

Mike Dowman<sup>1</sup>, Valentin Tablan<sup>1</sup>, Cristian Ursu<sup>1</sup>,  
Hamish Cunningham<sup>1</sup>, Borislav Popov<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK  
{mike, valyt, c.ursu, hamish}@dcs.shef.ac.uk  
<http://www.gate.ac.uk>

<sup>2</sup> Ontotext Lab, Sirma AI EAD, 135 Tsarigradsko Chaussee, Sofia 1784, Bulgaria  
borislav@sirma.bg  
<http://www.ontotext.com>

**Abstract.** The Rich News system for semantically annotating television news broadcasts and augmenting them with additional web content is described. On-line news sources were mined for material reporting the same stories as those found in television broadcasts, and the text of these pages was semantically annotated using the KIM knowledge management platform. This resulted in more effective indexing than would have been possible if the programme transcript was indexed directly, owing to the poor quality of transcripts produced using automatic speech recognition. In addition, the associations produced between web pages and television broadcasts enables the automatic creation of augmented interactive television broadcasts and multi-media websites.

## 1 Introduction

This paper describes the Rich News annotation system that semantically annotates television news broadcasts using news websites as a resource to aid in the annotation process. The chief obstacle to semantic annotation of television is that transcripts produced through automatic speech recognition (ASR) are of poor quality, so existing semantic annotation systems do not perform well on such data. Therefore Rich News bases its semantic analysis on related web pages, rather than on the ASR transcripts themselves. Previous work has used named entity extraction in conjunction with television broadcasts [23, 19], but Rich News is innovative both in the use of web-based content augmentation and in the use of semantic annotation [21].

Rich News is essentially an application-independent annotation system, there being many potential uses for the semantic annotations it produces. The first use proposed for the Rich News system was to automate the annotation of BBC news programmes [9]. For more than twenty years the BBC have been semantically

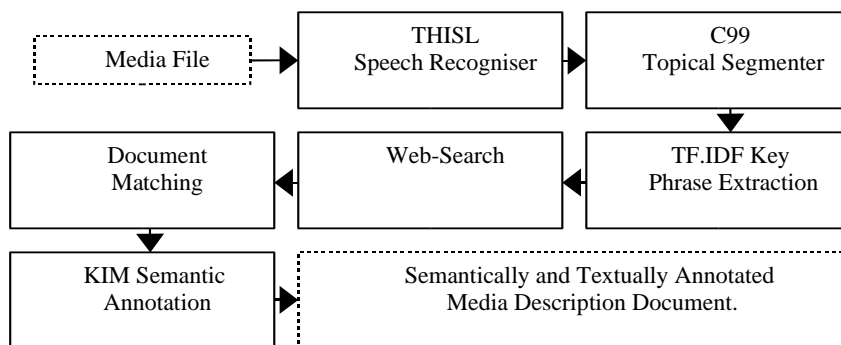
---

<sup>1</sup> This research was supported by the European Union through the SEKT, Knowledge Web and PrestoSpace projects. We would also like to thank the BBC archives for providing information about their archiving process, and for making broadcast material available to us.

annotating their news in terms of a taxonomy called *Lonclass*, which was derived from the Universal Decimal Classification system commonly used by libraries to classify books [3]. Lonclass has 300,000 subject terms, and so allows for an extremely precise form of annotation, but owing to limited archiving resource, 90% of the BBC's output is annotated at only a very basic level. While Rich News cannot annotate with the same level of precision as a human annotator, its performance is sufficient that it provides access to much material that would otherwise be difficult to retrieve owing to no description of it being available.

Besides being used in television archiving, Rich News could also be used to enhance the television experience for the viewer [10]. At present, electronic programme guides, which give information about the times of television programmes, and descriptions of them, are available on many television systems, enabling the viewer to access this information through their television set. Semantic electronic programme guides that give information about the entities referred to in particular broadcasts could be created, which would allow the user to view details of forthcoming programmes in terms of this information. If Rich News was incorporated into a digital video recorder, then the annotation process and generation of electronic programme guides could be performed by the digital video recorder in users' homes, rather than by the television company prior to transmission.

The third mode in which the annotations produced by Rich News can be used is to provide enhanced mixed-mode interactive television [8, 10]. In the UK, television news broadcasts now commonly include additional textual material that can be accessed by pressing a button on the remote control<sup>2</sup>. This textual content typically covers stories in more depth than is possible in the ordinary broadcast. Rich News can find relevant material automatically, and, if deployed on a digital video recorder, gives the viewer rather than the television company control over how that material is selected.



**Figure 1.** Architecture of Rich News Annotator

Rich News was developed using the GATE natural language processing architecture [7] which facilitated re-use of code from previous NLP systems, and hence rapid development of the system. The overall annotation system can be

<sup>2</sup> <http://news.bbc.co.uk/2/hi/entertainment/3974523.stm>

divided into the six modules shown in Figure 1. It takes a media file as input and produces a GATE document containing meta-data for each news story in the input file. The rest of this paper discusses each component in turn, and then shows one way in which the data produced by Rich News has already been used, in a semantic television search engine.

## 2 Speech Recognition

Speech recognition was achieved with the THISL speech recognition system [25, 24], which uses the ABBOT connectionist speech recognizer [26]. This system was optimized specifically for use on BBC news broadcasts, by customizing the pronunciation dictionary, and training the language model on over 100 million words of news text. Figure 2 shows an example of the speech recognizer's output.

one of the few things that could undermine that <SIL> as being a <SIL>  
heavy duty criminal activity and squire of class he dropped the more  
damage and roads that are <SIL> that could have brought her <SIL> from  
everything handing back their own making <s> i think today's action  
<SIL> has given a community confidence that does not want to be the  
case <s> and presumably bob

**Figure 2.** Example of the Speech Recognizer's Output. The story reports police operations against drug dealers, <s> and <SIL> mark short and long silences respectively

## 3 Segmentation into News Stories

Topical segmentation was used to divide the broadcast into the individual news stories that it contains. Most approaches to topical segmentation of media have been based on an analysis of the language used in the broadcast, but sometimes this has been in conjunction with cues from non-textual sources, such as an analysis of the television picture and the captions that appear on it [4, 18].

Most segmentation systems are trained on large corpora in which the story boundaries are marked [12, 20, 16], but no such corpus of BBC news programmes was available for training. Therefore, it was decided to segment the broadcasts using a measure of lexical cohesion, in which sentences in which the same words are repeated are grouped together. Several such systems have been developed [12, 15], but the one used in Rich News was C99 [5], which calculates the similarity between parts of a text using the cosine measure (see for example [14]), and which can automatically decide how many segments a text contains. Published evaluations show that C99's performance is not greatly below that of systems that rely on training data [16].

Key Phrases were then extracted from each story using *term frequency inverse document frequency* (TF.IDF) [11]. 'Phrases' (that is any sequence of one to three words), that occurred at least twice in a story, and which are not commonly seen in

the language as a whole, were taken to be key-phrases. More sophisticated key-phrase extraction systems have been reported in the literature [13, 28], but typically they require large collections of marked-up training documents, so there seems to be little advantage in using such systems when the present system works satisfactorily.

## 4 Web Search

The Google API<sup>3</sup> was used to search the news section of the BBC website, and the on-line versions of the *Times*, *Telegraph* and *Guardian* newspapers for web pages reporting the same stories as those in the broadcast. The searches were restricted to the day of broadcast, or the next day, by including these dates in the search query. For each story, the web sites were searched first using the best two key-phrases together, then each of the best four key-phrases individually, and in each case the first five URLs that Google returned were retrieved. An example of a complete search term that was passed to Google is given in (1).

(1) site:guardian.co.uk "3 December, 2002" OR "2 December, 2002" "smallpox " "vaccination "

A document matching component then loads the pages found by Google and finds which one is the most similar to the ASR text, in terms of matching words. If this page matches sufficiently closely, it is associated with the story, and used to derive title, summary and section annotations. An evaluation of an earlier version of Rich News [9] found that 92.6% of the web pages found in this way reported the same story as that in the broadcast, while the remaining ones reported closely related stories. That version of Rich News searched only the BBC website, and was successful in finding web pages for 40% of the stories, but the addition of multiple news sources, and an improved document matching component, can be expected to have raised both the precision and recall of the system, though no formal evaluation of the current system has been conducted.

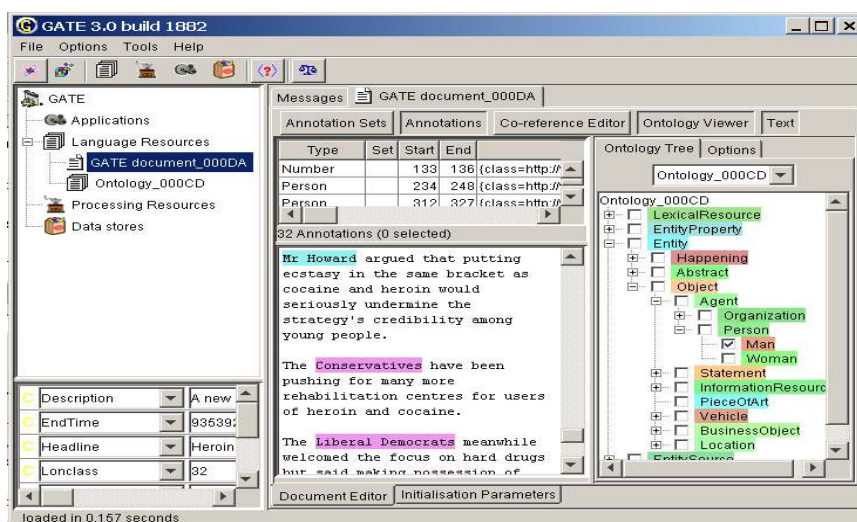
## 5 Semantic Annotation

An index document was created that described each story for which a matching web page had been found. These documents contained the main content text of the web page, its URL, and summary and section information extracted from it, as well as the start and end times of the story in the original broadcast. The text of the documents was then semantically annotated using the KIM knowledge and information management platform [22]. KIM produces meta-data for the Semantic Web [1] in the form of annotations with respect to a basic upper-level ontology called PROTON<sup>4</sup>. This ontology consists of three modular layers that contain categories for the most common types of entity (such as people, companies, and cities – in the Top ontology

<sup>3</sup> See <http://www.google.com/apis/>

<sup>4</sup> <http://proton.semanticweb.org>

module), as well as more specific ones (in the Upper ontology module). KIM identifies entities in texts both by looking them up in predefined lists, and by making shallow analyses of the text [21].



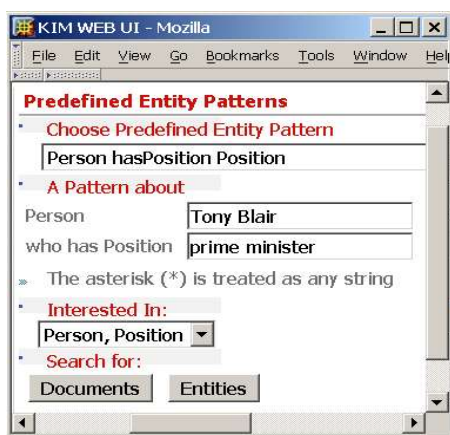
**Figure 3.** An Example of a Story Index Document that has been annotated by KIM, displayed in the GATE GUI

Because of the poor quality of the ASR transcripts, KIM was not able to annotate them effectively, so it was applied to the associated web pages instead – a medium on which the performance of KIM exceeds 90% (measured in terms of average F1 score) [17]. The ASR text was then searched to see if it contained text that matched any part of each entity on the web page, and if so a high confidence score was assigned to that entity. Often, even when most of an entity was missing or incorrect in the ASR transcript, it was still possible to match part of it using this technique. Entities in the web page that could not be matched to the ASR transcript were given confidence scores depending on how often they occurred, because those entities appearing most frequently in the web page were most likely to also occur in the broadcast. This is a step additional to the system presented in [9] and has the role of filtering out entity mentions that appear only in the web pages. This ensures a better match between the broadcast and the associated list of entities that is indexed. An example of a story index document that has been annotated by KIM is shown in Figure 3.

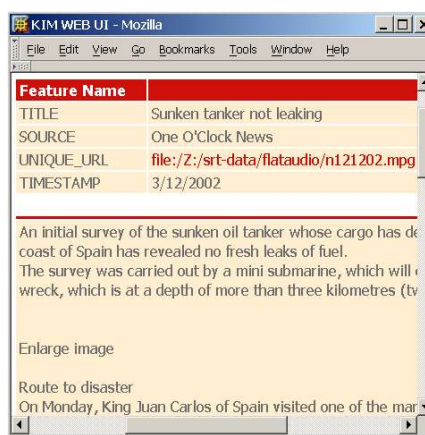
## 6 Search and Retrieval of Broadcasts

One of the applications of Rich News is to enable access to television news reports via a search interface. This is similar to the recently launched Blinkx and Google

television search engines<sup>5</sup> and Maybury’s Broadcast News Navigator [19], but those services do not semantically index the named entities in terms of an ontology. Rich News uses KIM’s annotation and Web UI (Figure 4) components to allow semantically enhanced searches. So, for example, if we wanted to search for a person whose last name was *Sydney*, we could specify that only entities annotated as *person*, or as some ontological sub-class of person (such as *woman*) were to be considered. This would prevent references to the city of Sydney being returned, which are far more numerous than references to people called Sydney. Figure 5 shows the result of a search. Clicking on the *UNIQUE\_URL* hyperlink opens a media player window to play the news story.



**Figure 4.** One of the predefined search patterns for entities in the KIM Web UI



**Figure 5.** A Story Found by the KIM Web UI

## 7 Future Developments

The major limitation of Rich News at present, is that it is not able to find web pages for all news stories. The most common reason that Rich News fails to find a related web page is because the broadcast has been incorrectly segmented, therefore a significant improvement might result if the segmentation algorithm was improved either by using more sophisticated semantic analysis techniques such as Latent Semantic Analysis [6, 2], or by using visual cues [4, 18]. A more accurate transcription could be obtained by using teletext subtitles (closed captions) instead of the ASR transcript when these are available, an approach used by [18] and Google’s television search.

At present the system only semantically analyses the web page that is the best match to the broadcast, but previous work [27] has shown that merging redundant results from multiple sources improves performance, so future work will investigate this possibility. So far Rich News has only been used to allow search of television broadcasts, as described in section 6, so future work will investigate its use for the

<sup>5</sup> blinkx.tv and video.google.com

other purposes outlined in section 1, such as to automatically create semantic electronic programme guides, or to deliver web material as part of television broadcasts. Rich News could also be extended so that it could be applied in non-news domains.

## 8 Conclusion

Rich News shows how web pages can be used to aid in the annotation of television news broadcasts. Applying semantic annotation tools such as KIM directly to ASR transcripts of television broadcasts produces poor results, but applying them to related web pages, and then matching the entities found in the web pages to the ASR transcript produces much better results. An additional benefit of this system is that the web pages can then be delivered as additional content in interactive television broadcasts, or used to produce new web pages integrating television content.

## References

1. Berners-Lee, T., Hendler, J. and Lassila, O. The semantic web. *Scientific American*, 284, Issue 5 (2001), 34-43.
2. Brants, T., Chen, F. and Tsochantaridis, I. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM* (McLean, VA, USA, November 2002), 211-218.
3. British Standards Institution. *Guide to the Universal Decimal Classification (UDC)*. British Standards Institution, London, 1963.
4. Chaisorn, L., Chua, T., Koh, C., Zhao, Y., Xu, H., Feng, H. and Tian, Q. A Two-Level Multi-Modal Approach for Story Segmentation of Large News Video Corpus. Presented at *TRECVID Conference*, (Gaithersburg, Washington D.C, November 2003). Published on-line at <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.
5. Choi, F. Y. Y., Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, (Seattle, USA, April, 2000), 26-33.
6. Choi, F. Y. Y., Wiemer-Hastings P. and Moore, J. Latent semantic analysis for text segmentation. In *Proceedings of EMNLP* (Pittsburgh, USA, June 2001), 109-117.
7. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. GATE: A framework and graphical development environment for robust NLP tools and applications. In *proceedings of ACL* (Philadelphia, USA, July 2002).
8. Dimitrova, N., Zimmerman, J., Janevski, A., Agnihotri, L., Haas, N., Li, D., Bolle, R., Velipasalar, S., McGee, T. and Nikolovska, L. Media personalisation and augmentation through multimedia processing and information extraction. In L. Ardissono and A. Kobsa and M. Maybury (Eds.), *Personalized Digital Television*, 201-233, Kluwer Academic Publishers, Dordrecht, Netherlands, 2004.
9. Dowman, M., Tablan, V., Cunningham, H. and Popov, B. Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. In *proceedings of 14th International World Wide Web Conference* (Chiba, Japan, May, 2005).
10. Dowman, M., Tablan, V., Cunningham, H. and Popov, B. Content Augmentation for Mixed-Mode News Broadcasts. In *proceedings of 3rd European Conference on Interactive Television: User Centred ITV Systems, Programmes and Applications*. (Aalborg , Denmark, March-April, 2005).

11. Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C. and Nevill-Manning, C. G. Domain-specific keyphrase extraction. In Proceedings of IJCAI, (Stockholm, Sweden, July-August, 1999), 668-673.
12. Franz, M., Ramabhadran, B., Ward, T. And Picheny, M. Automated transcription and topic segmentation of large spoken archives. In Proceedings of Eurospeech (Geneva, Switzerland, September 2003), 953-956.
13. Jin, R. and Hauptmann, A. G. A new probabilistic model for title generation. In proceedings of COLING (Taipei, Taiwan, August, 2002).
14. Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, 2000.
15. Kan, M., Klavans, J. L., and McKeown, K. R. Linear segmentation and segment significance. In Proceedings of the 6th International Workshop on Very Large Corpora (Montreal, Canada, August, 1998), 197-205.
16. Kehagias, A., Nicolaou, A., Petridis, V. and Fragkou, P. Text Segmentation by Product Partition Models and Dynamic Programming. *Mathematical and Computer Modelling*, 39, Issues 2-3, (January 2004), 209-217.
17. Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2, Issue 1, (2005).
18. Maybury, M. Broadcast News Understanding and Navigation. In proceedings of Innovative Applications of Artificial Intelligence (Acapulco, Mexico. August, 2003), 117-122.
19. Merlino, A., Morey, D. and Maybury, M. Broadcast News Navigation using Story Segmentation. In proceedings of ACM International Multimedia Conference (Seattle, WA, USA, November 1997), 381-391.
20. Mulbregt, P. V., Carp, I., Gillick, L., Lowe, S. and Yamron, J., Text segmentation and topic tracking on broadcast news via a hidden Markov model approach. The 5th international conference on spoken language processing (Sydney, Australia, November 1998).  
Published on-line at <http://www.shlrc.mq.edu.au/proceedings/icslp98/WELCOME.HTM>.
21. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D. and Kirilov, A. KIM – a semantic annotation platform for information extraction and retrieval. *Natural Language Engineering*, 10, Issues 3-4, (September 2004), 375-392.
22. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. Towards semantic web information extraction. In proceedings of ISWC (Sundial Resort, Florida, USA, October, 2003).
23. Przybocki, M., Fiscus, J., Garofolo, J. and Pallett, D. 1998 HUB-4 information extraction evaluation. In Proceedings of the DARPA Broadcast News Workshop (Herndon, VA, February, 1999), 13-18.
24. Renals, S., Abberley, D., Kirby, D. and Robinson, T. Indexing and Retrieval of Broadcast News. *Speech Communication*, 32, Issues 1-2 (September 2000), 5-20.
25. Robinson, T., Abberley, D., Kirby, D. and Renals, S. Recognition, indexing and retrieval of British broadcast news with the THISL system. In Proceedings of Eurospeech, (Budapest, Hungary, September 1999), 1067-1070.
26. Robinson, T., Hochberg, M. and Renals, S. The use of recurrent networks in continuous speech recognition. In C. H. Lee, K. K. Paliwal and F. K. Soong (Eds.), *Automatic speech and speaker recognition – advanced topics*, 233-258, Kluwer Academic Publishers, Boston, 1996.
27. Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O. and Wilks, Y. Multimedia indexing through multisource and multilingual information extraction; the MUMIS project. *Data and Knowledge Engineering*, 48, (2003), 247-264.
28. Turney, P. D. Coherent keyphrase extraction via web mining. In Proceedings of IJCAI (Acapulco, Mexico, August, 2002), 434-439.