

Content Augmentation for Mixed-Mode News Broadcasts

Mike Dowman*

Valentin Tablan*

Hamish Cunningham*

Borislav Popov[†]

*Department of Computer Science,
University of Sheffield
Sheffield, S1 4DP, UK
+44 114 22 21800

[†]Ontotext Lab, Sirma AI EAD,
135 Tsarigradsko Chaussee,
Sofia 1784, Bulgaria
+359 2 9768 310

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

borislav@sirma.bg

ABSTRACT

Rich News, a system that augments news broadcasts with textual content, is described. The system identifies individual stories in news broadcasts, and annotates them with related content from the World Wide Web. The web content is subsequently semantically analysed, and used to produce summary information for each news story. This content can then be delivered to users as part of an interactive television broadcast, or used to create semantically enhanced electronic programme guides. It also enables sophisticated search and browsing of news stories via a web interface. Rich News could be deployed either by broadcasters, or on digital video recorders in viewers' homes, and allows the creation of new personalized media experiences that integrate television and web content into one unified viewing experience.

Keywords

Semantic television, interactive television, electronic programme guides, multi-media, natural language processing, automatic speech recognition, topical segmentation, key-phrase extraction, Semantic Web.

1. Introduction

This paper demonstrates a system that allows enhanced modes of consuming television news, that have the potential to become widespread within 5-10 years. It describes the Rich News system that augments video content with material from the World Wide Web, allowing for the automatic creation of sophisticated semantically enabled electronic programme guides, the integration of web and video formats, and facilitating on demand access to television news. Media organizations such as the BBC have already introduced interactive news bulletins, delivered via digital television¹, and the Rich News system can automatically create meta-data suitable for use in these services. Furthermore, Rich News produces semantic meta-data that allows for sophisticated search and browsing, improving access to on-demand television services.

The Rich News system could be deployed either by a news broadcaster, and the meta-data then delivered to home users along with the television broadcast, or it could run in digital video recorders or set-top boxes in viewers' homes. The latter solution has the advantage of creating more potential for personalization, and does not have the privacy disadvantage of exposing sensitive personnel data to commercial interests, as would be the case if personalization was performed by the broadcaster. In home deployment of the system would also enable on demand access, without the need for high bandwidth connections to users homes, as the broadcasts could be played from recordings on a digital video recorder.

Rich News, annotates television news automatically, by associating documents found on the World Wide Web with each of the stories in a broadcast. This aids the creation of textual descriptions and summaries for the news stories, and allows for the creation of semantic annotations that can potentially form part of the Semantic Web [1]. In addition, the automatic linking of web and multimedia content enables a new model of mixed-mode media consumption [9], as viewers can find more in depth information about a story by viewing the associated web content. Previous work has adopted similar information extraction technologies (see for example [20]), but our work is novel in both the use of web-based content augmentation and in the use of semantic annotation [18].

The annotation process starts by first performing automatic speech recognition to achieve a rough transcript for each programme, and then analyzing this transcript to determine the boundaries between the various news stories that it describes. Rich News then associates pages from the BBC web site with each news story. These web pages are a form of meta-data in themselves, but they can also be used to create titles and classifications for individual news stories. The KIM information

The research for this paper was supported by a European Union Sixth Framework Programme grant, and is part of the PrestoSpace project. We would like to thank the BBC archives for providing information about their annotation process, and for making broadcast material available to us.

extraction system is used to find entities in the web pages, which are then annotated with semantic classes, allowing the stories to be indexed and queried in much more flexible ways than if text search alone were used. Most of the system, including parts of KIM, was developed using the GATE natural language processing architecture [8], which allowed rapid development, because many pluggable components were already available, and which facilitated the modular design of the system, making it easy to develop and maintain.

The overall annotation system can be divided into the seven 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. At present Rich News is used to allow on demand access to individual news stories, via a web-based search and browsing interface, and the stories are played from locally stored recordings. However, a television company could send the web content as part of an interactive broadcast, or use it to create electronic programme guides with enhanced search and browsing capabilities.

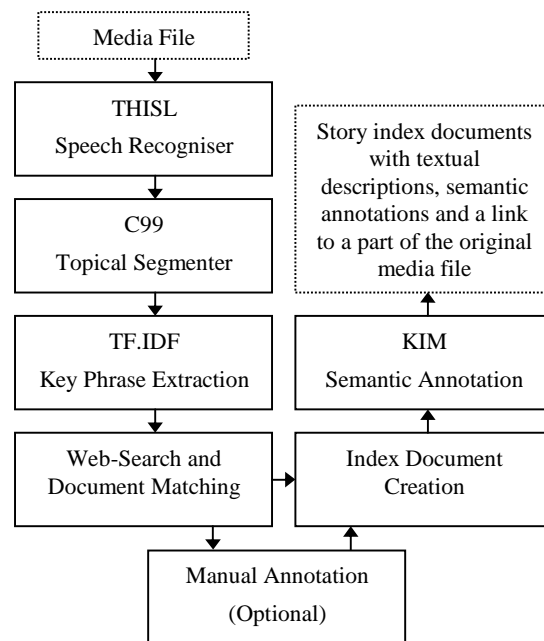


Figure 1. Architecture of Rich News Annotator.

2. Speech Recognition

Speech recognition was achieved with the THISL speech recognition system [22, 21], which uses the ABBOT connectionist speech recognizer [23]. 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. The speech recognizer does not mark punctuation in the transcripts it produces, and all words are output in lower case. However, it marks pauses as <s> or <SIL>, depending on length, and these are a good indicator of sentence breaks.

<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 2. Example of the Speech Recognizer's Output. (The story reports an oil spill.)

3. Topical Segmentation

Once an ASR (automatic speech recognition) transcript has been produced, the next stage is to try to segment this into individual news stories. There is a considerable literature concerning methods for segmenting both textual documents and audio and video material by topic. Most approaches to topical segmentation of media have been based on an analysis of the

¹ See <http://news.bbc.co.uk/2/hi/entertainment/3974523.stm>

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 [5].

Segmentation based on the language occurring in a broadcast can be achieved by identifying text that is indicative of boundaries between stories (such as 'back to the studio'), or by comparing the lexical similarity of nearby parts of the transcript. (Sections about the same story will tend to repeat the same words). However, most such approaches are trained on large training corpora (several million words), in which story boundaries are marked [11, 17, 15]. This created a problem for the case of BBC news, because no such corpus of BBC news programmes was available for training and NLP systems do not generally 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.

Therefore it was decided to attempt segmentation with a technique that did not require training data, as this would make the segmentation system easier to develop, and would not restrict it to one particular genre or programme. Several segmentation systems have been developed that segment using measures of lexical cohesion [14, 13]. A comparison is made of the extent to which neighbouring parts of the text contain the same words, which is a good indicator of whether they are about the same story. Such methods segment based on an analysis of the text as a whole, rather than just the text at story boundaries, 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 [6], which calculates the similarity between parts of a text using the cosine measure (see for example Jurafsky and Martin [13]), and which can automatically decide how many segments a text contains. Kehagias et al [15] 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 P_k metric [1], compared to 5.38% for their own algorithm. Lower Beeferman scores indicate higher performance.)

C99 has been found to work well on the BBC news programmes, though it often fails to create separate topical segments for very short stories (which are sometimes covered in only 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. However, we will see below that the document matcher can compensate for some such errors.

4. Key-phrase Extraction

Once the segmenter has segmented the ASR transcript, the next stage is to find key words or phrases that are representative of each story. Most of the key-phrase extraction software reported in the literature [12, 25] relies on genre-specific training data in which suitable key-phrases are already marked. However, no such training corpus was available for BBC news, and so a simple system based on the *term frequency inverse document frequency* (TF.IDF) measure of Frank et al [10] was implemented. This identifies words and phrases that occur more commonly in the stories than they do in the text as a whole, and was customised for BBC news by training on ASR transcripts of 13,353 news broadcasts.

5. Search of the Web for Related Documents

The purpose of extracting key-phrases was so they could be used to search for web pages reporting the same story on the BBC web site. Searches were conducted using Google, which was accessed via the Google Web API². Searches were restricted to the news section of the BBC web site, by adding the term *site:news.bbc.co.uk* to each search. The searches were restricted only to the day of broadcast, or the day before, by adding a term specifying either of these dates in the format that they appear on BBC news web pages, for example "*1 December, 2004*" OR "*30 November, 2004*". Besides the site and date terms, key phrases were also added to the queries. (2) gives one example of a complete search term, which concerned a story about UK government preparations for a possible terrorist attack involving smallpox. Multiple searches were performed for each story using different key phrases in order to maximise the chances of obtaining a matching web pages, and for each search, the first three URLs returned by Google were retrieved.

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

A document matching component then loads the pages found by Google and compares their text to that of the ASR transcript of the story until one is found that matches sufficiently closely. This web page is then associated with the story, and a title, summary and section extracted from *meta* HTML in the web page. Sections of the broadcast that correspond to headlines, or which contain more than one news story, will not match any web pages, and so will usually not be associated with any meta-data, hence compensating for errors made by the topical segmenter, and improving the overall precision of the system.

² See <http://www.google.com/apis/>

6. Manual Annotation

Rich News is not always successful at finding web pages for all stories, and sometimes the segmentation is not completely accurate, so Rich News allows corrections to be made, and missing meta-data added, using the ELAN linguistic annotator [4], shown in Figure 3. Such correction could be undertaken by a media company prior to broadcast (so long as the broadcast is not live), but is clearly not suitable if the system is deployed in viewers' homes. However, so long as the viewer is prepared to tolerate some inaccuracy or omissions in the final system, this stage can be omitted.



Figure 3. Editing Annotations on a Television Broadcast.

7. Story Index Document Creation

For each story in the broadcast for which a matching web page has been found, a GATE document is created, containing the main content text of the web page, its URL, and the summary and section information extracted from it. In addition, the URL of the media file containing the original broadcast, the start and end times of the story, and information about the channel and date on which the programme was broadcast, are added. These documents can then be used to index the broadcast in search or browsing systems or to create a detailed electronic programme guide, or could provide additional content during the broadcast.

8. Semantic Annotation

In order to allow more sophisticated search and browsing of the news stories, the textual meta-data produced by earlier components was enhanced through the addition of semantic annotations. This was achieved using the KIM knowledge and information management platform [19]. KIM produces meta-data for the Semantic Web [2] in the form of RDF annotations, and organises these annotations with an ontology that contains categories for the most common types of entity, such as people, companies and cities. KIM identifies entities in texts both by looking them up in predefined lists, and by making shallow analyses of the text [17]. Because of the poor quality of the ASR transcripts, KIM was not able to effectively annotate them, so instead it was applied to the associated web pages, a medium on which KIM's performance exceeds 90% (measured in terms of average F1 score) [16]. An example of a story index document that has been annotated by KIM is shown in Figure 4.

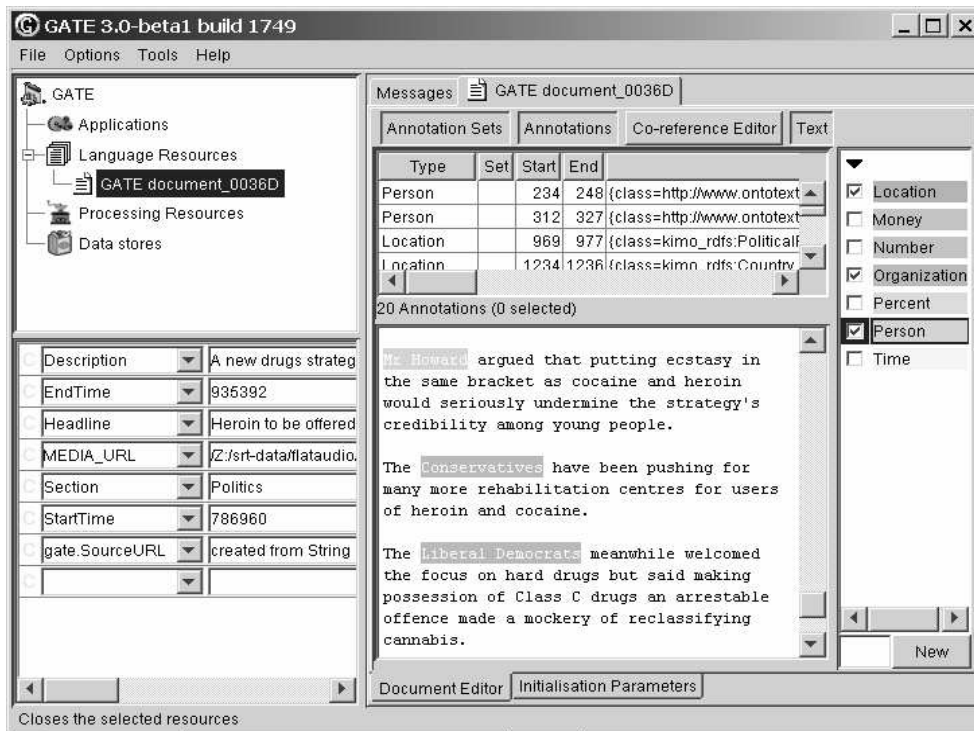


Figure 4. An Example of a Story Index Document that has been annotated by KIM, displayed in the GATE GUI.

9. 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 was achieved using the KIM Web UI (Figure 5), which allows not only simple text searches, but also semantically enhanced ones. 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. Searching for an Organization with the KIM Web UI.

Figure 6 shows the result of a search, in which a description of the news story is displayed. Clicking on the *UNIQUE_URL* hyperlink will open a media player window to play the news story. These story description documents could also be used to make a single HTML document describing a news broadcast, from which a viewer could choose to play individual news stories in any order that he or she chose. Use of Rich News in this way requires that the news programmes be recorded and saved locally on a computer disk, so that they can be played on demand. This functionality could be produced by incorporating Rich News into a digital video recorder, or could be provided over the internet, subject to sufficient bandwidth being available for the transmission of video data.



Figure 6. A Story Found by the KIM Web UI.

10. Evaluation of the Rich News Annotator

An evaluation of Rich News Annotator with respect to four and a half hours of BBC news output was performed. This comprised nine half hour news broadcasts, which contained a total of 66 news stories. It was found that 92.6% of the web pages found reported the same story as that in the broadcast, and the remaining ones reported closely related stories. However, web pages were found for only 40% of the stories, but work is in progress to improve on this by complementing the search of the BBC web site with searches of other on-line news sources, increasing the probability of finding a related web page. The stories that were missed by the annotator were usually those that consisted of only one or two sentences, and they were not always reported on the BBC web site, suggesting that these stories were of lesser importance.

11. Future Developments

While Rich News is a complete, fully functional system at present, it will remain under development for some time. While the individual components it contains all function well enough for acceptable results to be produced, all of them could clearly be improved. The biggest single improvement in the overall system would result if the quality of the speech recognition could be improved. However, the speech recognition component used is already state of the art, and it is not clear that significantly better speech recognition systems will be available in the near future, so it is more likely that improvements will have to be made in other areas.

The story segmentation component could be improved by basing the segmentation on a semantic analysis of the meaning of the words in the ASR transcripts, rather than on the words directly. This can be done using *Latent semantic analysis*, and has been shown to considerably improve segmentation performance [7, 3]. As many of the failures of the present system result from segmentation errors, any such improvement to the segmentation component could be expected to make a significant improvement to the performance of the system as a whole. A similar semantic analysis procedure might also help to improve the key-phrase extractor and the document matcher components.

In previous work on annotating sports video [24] we found that merging redundant results from multiple sources improves performance, so future work will apply this approach in Rich News by using on-line newspapers in addition to the BBC web site. This would also provide an enhanced experience for the viewer, as they would then receive news from multiple sources. Furthermore, the system could be developed to automatically integrate multiple news broadcasts and then deliver them to the viewer as a unified whole, in much the same way as meta-media systems do at present for on-line news services³. The addition

³ See for example <http://www.google.co.uk/news>

of personalization to such a service would allow viewers to create custom news channels focussing on topics of interest, by combining relevant stories from all available news broadcasts.

12. Conclusion

Rich News addresses the problem of how to produce meta-data for augmented interactive news broadcasts, and how to improve on-demand access to news stories. At present, interactive news services typically contain very little additional content, due largely to the high cost of producing such material, and the need to produce it in time for the broadcast. There is also generally no potential to personalise such services. Rich News solves these problems in a fully automatic way, and facilitates the integration of web and television content. The hardware necessary for the deployment of Rich News is already available, so it is likely that such systems will come into widespread use within in the next few years.

13. References

- [1] Beeferman, D. Berger, A. and Lafferty, J. Statistical models for text segmentation. *Machine Learning*, Volume 34 (1999), 177-210.
- [2] Berners-Lee, T., Hendler, J. and Lassila, O. The semantic web. *Scientific American*, 284, Issue 5 (2001), 34-43.
- [3] 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.
- [4] Brugman, H. and Russel, A.. Annotating multi-media / multi-modal resources with ELAN. In *proceedings of LREC* (Lisbon, Portugal, May, 2004), 2065-2068.
- [5] 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>.
- [6] Choi, F. Y. Y., Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, (Seattle, USA, April, 2000), 26-33.
- [7] 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.
- [8] 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).
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Jin, R. and Hauptmann, A. G. A new probabilistic model for title generation. In *proceedings of COLING* (Taipei, Taiwan, August, 2002).
- [13] 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.
- [14] 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.
- [15] 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.
- [16] Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. Semantic annotation, indexing, and retrieval. *Web Semantics*, (in press).
- [17] 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>.

- [18] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. KIM -- semantic annotation platform. *Natural Language Engineering* (to appear).
- [19] 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).
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] Turney, P. D. Coherent keyphrase extraction via web mining. In Proceedings of IJCAI (Acapulco, Mexico, August, 2002), 434-439.