The Cambridge University Multimedia Document Retrieval Demo System

A.Tuerk†, S.E. Johnson†, P. Jourlin‡, K. Spärck Jones‡ & P.C. Woodland†

†Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {at233,sej28,pcw}@eng.cam.ac.uk

‡Cambridge University Computer Laboratory
Pembroke Street, Cambridge, CB2 3QG, UK.
Email: {pj207,ksj}@cl.cam.ac.uk

http://svr-www.eng.cam.ac.uk/research/projects/Multimedia_Document_Retrieval

1 System Description

The CU-MDR Demo [3] is a web based application that allows the user to query a database of automatically generated transcripts of radio broadcasts that are available on-line. The system downloads the audio track of British and American news broadcasts from the Internet once a day and adds them to its archive. The audio, which comes in RealAudio format, is first converted into standard uncompressed format from which a transcription is produced using our large vocabulary broadcast news recognition engine. This yields a collection of text and audio documents which can be searched by the user.

The recogniser is similar to the system running in 10 times real time described in [2]. This system gives a word error rate of 15.9% on the 1998 Hub4 broadcast news evaluation data. On the Internet audio used here the run-time is increased due to reduced audio quality but the general level of transcription accuracy remains high at approximately 20% word error rate on NPR data.

The information retrieval engine used in the CU-MDR demo is the benchmark system described in [1]. A windowing/recombination system is used for audio data for which story boundary information is not available. Semantic posets are not automatically applied in searching. Instead they are exploited to suggest new words to the user for addition to the original query. Interactive relevance feedback is also available. This allows the user to mark the documents that contain relevant information and have the system suggest additional query words that distinguish those from the non-relevant documents.

2 User Interface

The user can either query the audio/text database interactively, submitting a request by typing a query into the search field or he/she can specify a set of standing queries that represent his/her long-term interests. The retrieval engine is run on these queries only on login. If the retriever finds a document that matches one of these queries and which has not been seen by the user before, the query is highlighted. This feature effectively allows the user to filter the incoming broadcasts.

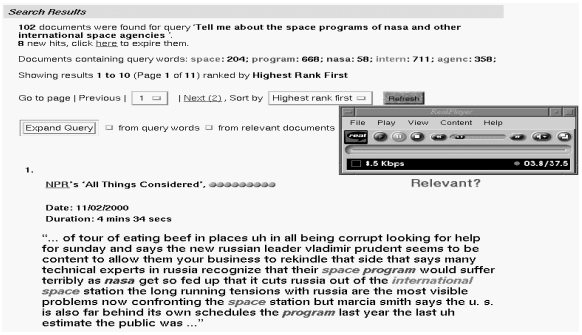

Figure 1: MDR demo interface

Once the retriever has returned the results for a particular search, a list of extracts from the returned text documents is created with the query words highlighted. Each extract is supposed to represent that part of the broadcast document that is most relevant to the query. The user can listen to the part of the sound source that corresponds to the extract. The whole automatic transcript can be accessed on a separate web page where the user can listen to selected parts of the transcription by highlighting them.

Acknowledgements

This work is partly funded by EPSRC grant GR/L49611. We would also like to thank Ben Timms and Richard Wareham for their contribution to the demo interface.

References

[1] S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland. Spoken Document Retrieval for TREC-8 at Cambridge University. To appear. In *Proc. TREC-8*, NIST Gaithersburg, MD, 2000.

[2] J. J. Odell, P. C. Woodland, and T. Hain. The CUHTK-Entropic 10xRT Broadcast News Transcription System. In *Proc. DARPA Broadcast News Workshop*, pages 271–275, Herndon, VA, 1999.

[3] A. Tuerk, S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland. The Cambridge University Multimedia Document Retrieval Demo System. In *Proc. RIAO 2000, Content-Based Multimedia Information Access*, volume 3, pages 14–15, Paris, France, 2000.

To appear in SIGIR 2000, July 2000, Athens, Greece