Multilingual Video and Audio News Alerting

David D. Palmer, Patrick Bray, Marc Reichman, Katherine Rhodes, Noah White
Virage Advanced Technology Group
300 Unicorn Park
Woburn, MA 01801
{dpalmer, pbray, mreichman, krhodes, nwhite}@virage.com

Andrew Merlino, Francis Kubala
BBN Technologies
50 Moulton St.
Cambridge, MA 02138
{amerlino, fkubala}@bbn.com

Abstract

This paper describes a fully-automated real-time broadcast news video and audio processing system. The system combines speech recognition, machine translation, and cross-lingual information retrieval components to enable real-time alerting from live English and Arabic news sources.

1 Real-time Video Alerting

This paper describes the Enhanced Video Text and Audio Processing (eViTAP) system, which provides fully-automated real-time broadcast news video and audio processing. The system combines state-of-the-art automatic speech recognition and machine translation components with cross-lingual information retrieval in order to enable searching of multilingual video news sources by a monolingual speaker. In addition to full search capabilities, the system also enables real-time alerting, such that a user can be notified as soon as a word, phrase, or topic of interest appears in an English or Arabic news broadcast.

The key component of the news processing is the Virage VideoLogger video indexer software package (Virage 2003). The VideoLogger processes an incoming live satellite feed, encodes the video as a digital file, and manages the video and audio processing components. The individual components integrated into the VideoLogger platform currently include the audio processing and machine translation systems described in Section 2, as well as face ID, broadcaster logo ID, and scene change analysis.

The video and audio processing components produce textual metadata that is time-stamped to enable synchronization with the encoded video file. All data is indexed and stored for retrieval by a cross-lingual information retrieval engine. Figure 1 shows the EViTAP cross-lingual search and alerting interface, with real data from the system. The list of relevant video clips matching an alerting profile or a user search is shown on the left, with broadcast source and time, most-frequent named entities, and a relevancy score. Note that the English query "bin laden" resulted in the display of relevant stories in both English and Arabic. The center of the screen contains the video playback window, with clip navigation and keyframe storyboard. The right of the interface contains the transcripts from the ASR and MT engines; video playback is synchronized with the transcripts such that words are highlighted as they are spoken in the video.

2 Real-time Spoken Language Processing

The real-time audio processing in the eViTAP system is performed by the BBN AudioIndexer system, described in detail in (Makhoul *et al.* 2000). The AudioIndexer system provides a wide range of real-time audio processing components, including automatic speech recognition, speaker segmentation and identification, topic classification, and named entity detection. All audio processing is carried out on a high-end PC (dual 2.6 GHz Xeon CPU, 2 GB RAM). The real-time speech recognition system produces a word error rate of roughly 20-30% for English and Arabic news sources.

Figure 1: Multilingual alerting and search interface, with alert list, synchronized video playback, Arabic speech recognition output, Arabic-to-English machine translation output.

The Arabic words produced by the speech recognition system, including all ASR errors, are processed by an Arabic-to-English machine translation system that also operates in real time (on a separate high-end PC). The eViTAP system currently processes Arabic sources using statistical machine translation systems from IBM (Al-Onaizan 2003) and Language Weaver (Benjamin *et al.* 2003).

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency and monitored by SPAWAR under contract NBCHD030007.

References

Y. Al-Onaizan, R. Florian, M. Franz, H. Hassan, Y. S. Lee, S. McCarley, K. Papineni, S. Roukos, J. Sorensen, C. Tillmann, T. Ward, F. Xia, "TIPS: A Translingual Information Processing System," In *Proceedings of HLT-NAACL 2003 Demonstrations*, Edmonton, 2003.

B. Benjamin, L. Gerber, K. Knight, D. Marcu, "Language Weaver: The Next Generation of Machine Translation," In *Proceedings of MT Summit IX*, New Orleans, Louisiana, September 23-27, 2003.

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and Language Technologies for Audio Indexing and retrieval," In *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338-1353, 2000.

Virage VideoLogger Factsheet (2003)
http://www.virage.com/files/products/VL_DS_lores.pdf