Intelligent Multimodal Stream Processing

Mark Maybury
Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
mavbury(dlmitre. org
www, mitre, org/resources/centers/it

Abstract

This poster describes methods to enable intelligent access to multimodal information streams. We illustrate these methods in two integrated systems: the Broadcast News Editor (BNE) which incorporates image, speech, and language processing and the Broadcast News Navigator (BNN) which provides search, visualization and personalized access to broadcast news video. BNN enables users to perform keyword and named entity search, temporally and geospatially visualize entities and stories, cluster stories, discover entity relations, and obtain personalized multimedia summaries. By transforming access from sequential to direct search and providing hierarchical hyperlinked summaries, BNE and BNN enable users to access topics and entity news clusters nearly three times as fast as direct search of video.

1. Intelligent News on Demand

Figure 1 illustrates the BNE and BNN systems that integrate components from MITRE, Oracle, Carnegie Mellon University and Lincoln Laboratory to process imagery, audio, and text to enable news on demand. Key elements of BNE include scene change detection and classification from imagery, silence and speaker change detection from audio, and named entity extraction from speech or closed caption transcripts. A correlation component extracts cues from across these streams to detect the start and end of broadcasts, commercials, and stories. Subsequently, stories are classified and summarized, including selection of key frames, named entities, and representative sentences. Finally, the user can search and explore video stories, provide relevancy feedback, specify media preferences, and obtain personalized displays using the BNN web browser as shown in Figure 4 below. As detailed in [Maybury 2003], BNE and BNN transform video access from sequential to direct search, providing novel navigation and discovery mechanisms such as topical and entity-specific news clusters. Empirical evaluations have shown that users can find stories and answer questions nearly three times as fast as searching digital video with no loss in precision and recall.

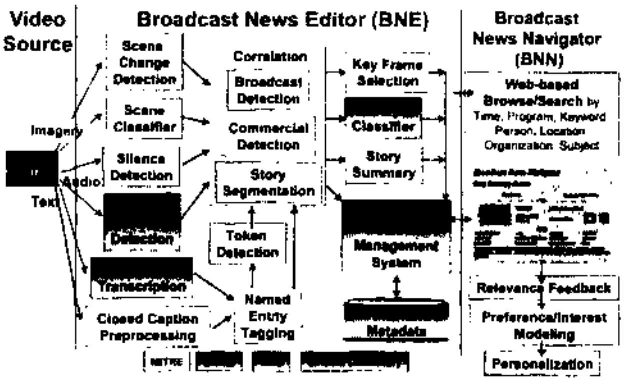

Figure 1. Multistream Broadcast News Understanding: System Architecture

2. Intelligent Multimodal Segmentation

BNN exploits text, audio, and imagery streams and associated cues to detect story and commercial segment boundaries and to select media elements to use for summaries and multimodal displays. Underlying BNN is a set of machine-learned, time-enhanced finite state automata modeling news structure that take into account the above cues and probabilistic, temporal models of event occurrence [Boykin and Merlino 2000]. Program start/end, anchor/report shots, commercials, and/or story shifts are inferred from cross media cues in text (e.g., frequent weather or sports terms, funding and/or copyright notices), discourse cues (e.g., "coming up next"), music (e.g., characteristic jingles), silence (indicating breaks to commercial), and visual cues (e.g., logos, anchor vs. reporter shots). For example, as shown in Figure 2, our frequency analysis of news anchor sign on terms (e.g., "hello", "welcome") or sign off terms (e.g., "that's all", "thanks for watching") occurring minutes from program start enables the creation of pattern detectors for such events. We have discovered analyzing months worth of news, for example, that sign off* terms occur in 97% of news programs.

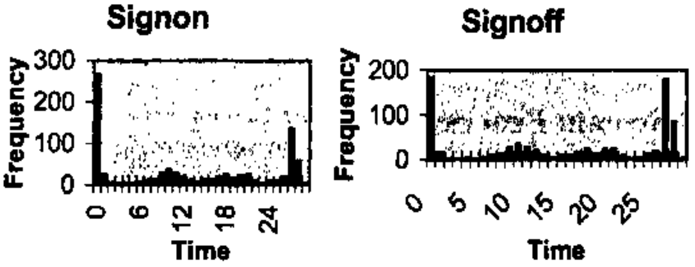

Figure 2. Text Analysis of Program Start and End Terms

Analogously, as shown in Figure 3, BNE classifies 15 frames per second into images associated with program start (e.g., logos), commercial breaks (e.g., blackframes), single and double anchor shots, and reporter shots.

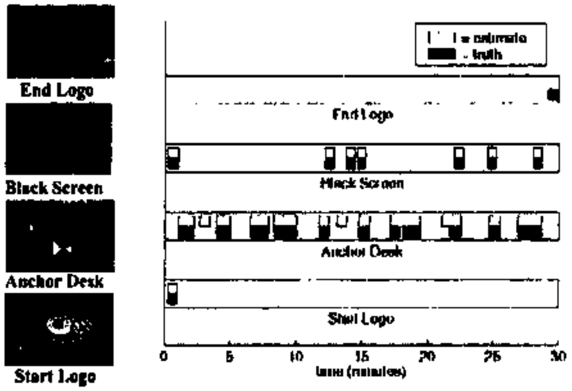

Figure 3. Imagery Analysis

3. Search, Retrieval, and Visualization

After BNE segments stories with multimodal cues, BNN supports the retrieval of stories by source, date range, keyword, named entity, or topic query. As exemplified in the upper left of Figure 4, after selecting news programs (e.g., CNN NewsNight, ABC World News Tonight) and indicating date ranges (e.g., Februrary 6-9, 2003), a user types in keywords in the text box. If a user is unfamiliar with retrieval terms, they can select from an alphabetic listing of all the named entities extracted for the time period and from the programs of interest, such as shown in the person, organization, and location lists in the upper left of Figure 4 (note "Hans Blix" is highlighted). This retrieves 14 matching stories and displays them as a "Story Skim", the source and date, the top 3 named entities in each story, and a representative key frame from each selected segment. Finally, by selecting a story, the user is presented with "Story Details", including all named entities, a text summary, and access to video and transcript sources. In empirical studies this mix of extracted media and hypertext organization enables analysts to perform searches three times faster than with video source.

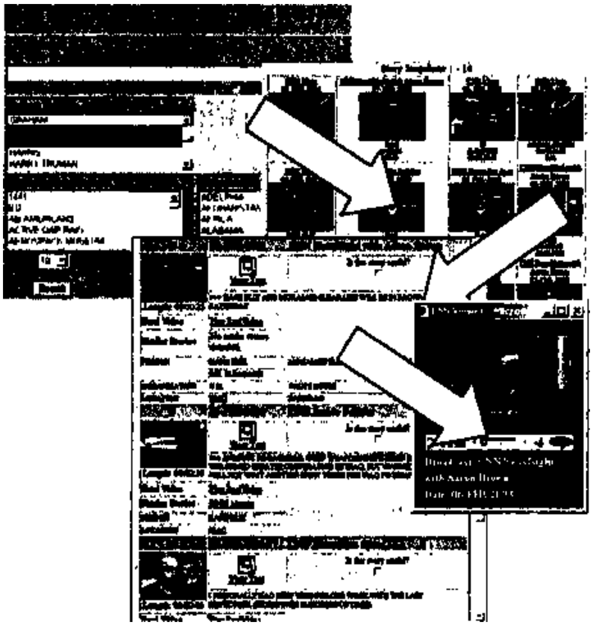

Figure 4. Text and Named Entity Search Menus, "Story Skim" Results and "Story Detail" Results

Figure 5 illustrates a user visualizing named entity frequencies and animating story occurrences associated with geospatial regions over time enabling information discovery.

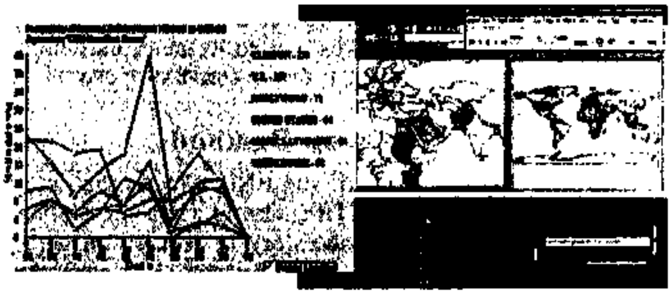

Figure 5. News Visualization

Acknowledgments

This research involved collaboration among Stanley Boykin, Andy Merlino, Warren Grieff, and Chad McHenry, Chris Clifton, John Aberdeen, John Burger, David Day, Lynette Hirschman, and Rod Holland.

References

[Boykin and Merlino, 2000] Boykin, S. and Merlino, M. Feb. 2000. A. Machine learning of event segmentation for news on demand. *Communications of the ACM.* Vol 43(2): 35-41.

[Maybury 2003] Maybury, M. 2003. Broadcast News Understanding and Navigation. Innovative Applications of AI. Acapulco, Mexico. 12-14 August, 2003.