Broadcast News Understanding and Navigation

Mark Maybury

Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
www.mitre.org/resources/centers/it

Abstract

The Broadcast News Editor (BNE) and Broadcast News Navigator (BNN) are fully implemented systems that exploit integrated image, speech, and language processing to support intelligent access to broadcast news video. This paper summarizes the integration of a range of AI techniques within these systems to provide intelligent segmentation, extraction, search, summarization, visualization, and personalized multimedia generation. 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. We report an empirical evaluation that demonstrates users can find stories and answer questions nearly three times as fast as searching digital video with no loss in precision and recall.

News Understanding

Motivated by the need to potentially search hundreds of thousands of programs broadcast each week globally, we have explored several fundamental tasks supporting the automated processing and exploitation of video news from North American and foreign sources including:

- *Segmentation* – Automatically decomposing news into subparts of stories, commercials, interviews, etc.
- *Extraction* – Extraction of video subcomponents from audio (e.g., jingles, music, speaker ID), imagery (e.g., anchor or reporter scenes, representative key frames, faces), or text (e.g., significant terms or named entities such as people, organizations, and locations from speech transcripts or closed captions).
- *Search* – Retrieval of stories based on user query, either keyword, named entity, or subject; use of relevancy feedback to refine and enhance information need specification and satisfaction.
- *Discovery* – Learning relationships among entities (e.g., name co-occurrence in stories and sources)
- *Summarization* – Automated selection and compression of text, audio, and/or imagery segments.
- *Generation* – Selection, structuring, ordering and layout of extracted video components to synthesize presentations organized around sources, times, topics, or named entities.
- *Visualization* – Temporal and geospatial visualization of video content.
- *Navigation* – Organization into linked and/or hierarchical subdocuments to support the rapid browsing and discovery of content.
- *Personalization* – Generating news sensitive to the content, media, and layout preferences of the user.
- *Multilinguality* – Foreign language news processing.

The remainder of this paper describes our exploration of these areas. To conclude, we report results from controlled evaluations and user studies.

Intelligent Multimedia Segmentation

The Broadcast News Editor (BNE) analyzes news by cross media processing of text, audio, and motion imagery. BNE uses multimodal cues to robustly segment video by detecting story states in a temporally indexed, state transition model (Boykin and Merlino 2000).

Text Processing

Using a corpus based, machine learning approach (Aberdeen et al. 1995) operating on speech transcriptions or closed caption text (when available), we discover rule patterns such as "Good evening, I'm <person-name>", indicating the start of a news program. These discourse cue patterns signal news state transitions such as anchor to reporter handoffs (e.g., "We go now to <person-name> in <location-name>"), the end of a reporter segment (e.g., "I'm <person-name> reporting from <location-name>"), and story termination and segment cataphora (e.g., "coming up next, <person-name> reports on <topic-x> from <location-name>"). In BNE evaluations, we found that story segmentation using these discourse cues performs better than lexical topic segmentation or (incomplete and errorful) closed-captioned cues indicating speaker change (">>") or topic change (">>>"). We use all evidence, however, to overcome both missing cues and 10-15% word error rates in captions and transcripts. For example, machine learned terminology distribution

models help characterize segment class occurrence (e.g., weather, sports) as shown in Figure 1. Similarly, we have found blank lines in closed caption text correlate with advertisements.

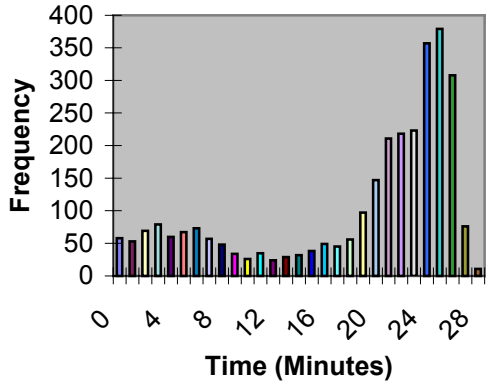

Figure 1. Frequency in minutes after program start of weather terms from one month of CNN Prime News™

Audio Processing

From the audio stream, BNE detects silence (e.g., at least 0.7 seconds are indicative of commercial start/end in CNN), music logos (indicative of program start, e.g., Leher News Hour), and speaker change. Carnegie Mellon University's SPHINX-II provides speech transcription and Lincoln Laboratory software detects speaker change.

Imagery Processing

Visual elements of broadcast news represent an equally rich source of information regarding the structure and content of news. 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.

Segmentation Learning

BNE is founded on a state-based model of broadcast news that assumes news is structured into intentionally sequenced segments such as scenes of news overviews, anchors, reporters, and advertisements. BNE utilizes a range of multimodal cues, content analysis, and temporal knowledge of likely event occurrences and durations to detect shifts between states. Segmentation was originally controlled by hand crafted, temporally-constrained, finite state automata (T-FSA).

Figure 2 exemplifies the T-FSA for CNN Prime News. After broadcast classification (state #1) and initiation (#2), the highlights of the day (#3, 4) are typically followed by anchor story segments (#5, 6, 7), reporter segments (#22), advertisements (#8, 9, 13, 14, 15), and program end (#18-21). Note, for example, the transition from detecting an advertisement segment (from states #13-15 to state #8) is recognized by detecting "silence" in the audio stream, "blank" lines in the close caption stream, and "black" frames in the video stream. Cues shown in Figure 2 such as visual logo or black key frame, detection of a person name, or audio silence of a specified duration are the primary ones used for detecting each state. However, the T-FSA is robust enough to detect state transitions even when major cues are missing.

This manually crafted model was extended using hidden Markov modeling to automatically learn the most predictive combination of cues to classify transitions (Boykin and Merlino 2000). Transition cue combination and segment sequence learning is induced from marked-up gold standard examples of hundreds of news programs. The hand crafted T-FSA segmentation performance over a range of broadcast sources (e.g., CNN, MS-NBC, and ABC) averaging across all cues is 38% precision and 42% recall. In contrast, performance for the best combination of multimodal cues rises to 53% precision and 78% recall (Maybury 2002). When visual anchor booth recognition cues are specialized to a specific program (e.g., British ITN Evening News), performance rises as high as 96% weighted precision and recall.

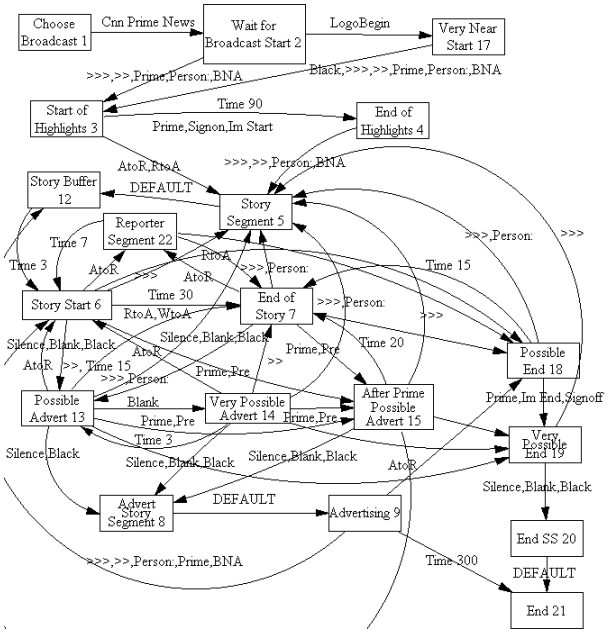

Figure 2. Temporally enhanced finite state news model for CNN Prime News™

Media Extraction

Just as BNE processes cross media cues to support segmentation, it extracts text (e.g., statistically significant terms or named entities such as people, organizations, locations), audio (e.g., audio logos), and imagery (e.g., anchor or reporter segments, representative key frames) for subsequent use. For example, the named entity extractor achieves 88% accuracy (balanced precision and recall) on reference transcriptions with word error rates (WER) of 0%. More consistent with our operational

data, in the DARPA HUB-4 evaluation, the system achieves 71-81% accuracy extracting named entities on broadcast news speech data with a wide range of WERs ranging from 13% to 28% (Palmer et al. 1999).

Importantly, detected news structure is exploited during extraction. For example, the sentence within a story segment that contains the most frequently occurring named entities is extracted as the story summary. Related, BNE exploits news structure and visual properties to extract visual keyframes are most representative of the story. Specifically, for an anchor shot BNE selects a frame at the beginning of the segment where the story topic is often visually introduced with a graphic overlay. In contrast, for a report from the field BNE chooses the middle of the segment where the content of the segment is being presented. Selection heuristics take into account those keyframes that contain text overlays that provide additional identification of speaker, location, or topic.

Search/Retrieval

The front end to the BNE news analysis system is the Broadcast News Navigator (BNN). After selecting sources and date ranges, BNN supports the retrieval of stories based on user query by keyword, named entity, or subject. As shown in the top of Figure 3, a user can type in keywords as in a classical search engine (e.g., "Korean weapons of mass destruction"). If a user is unfamiliar with retrieval terms, they can display an alphabetic listing of all the named entities extracted from programs of interest for selected time periods, as shown in Figure 3. As discussed in the previous section, note the information extraction errors such as "reserve" under the scrollable list of organizations or "English" under locations.

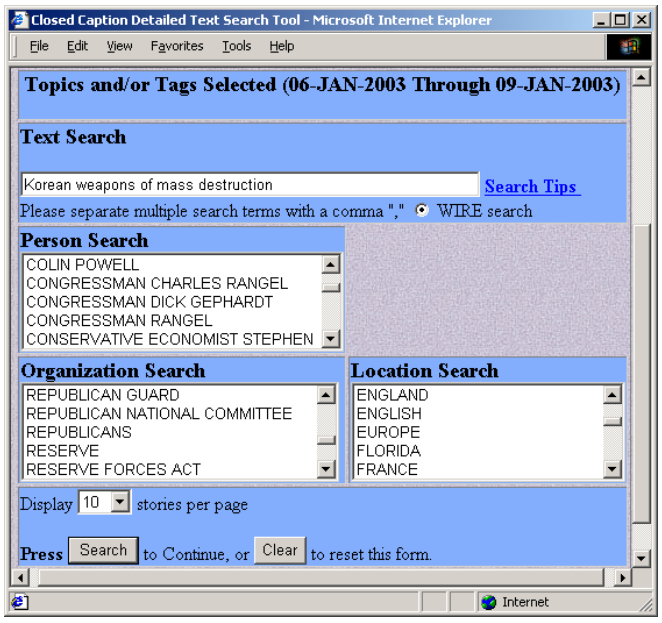

Figure 3. Automatically Generated Named Entity Menus

Figure 4. January 2003 North Korean "Story Skim"

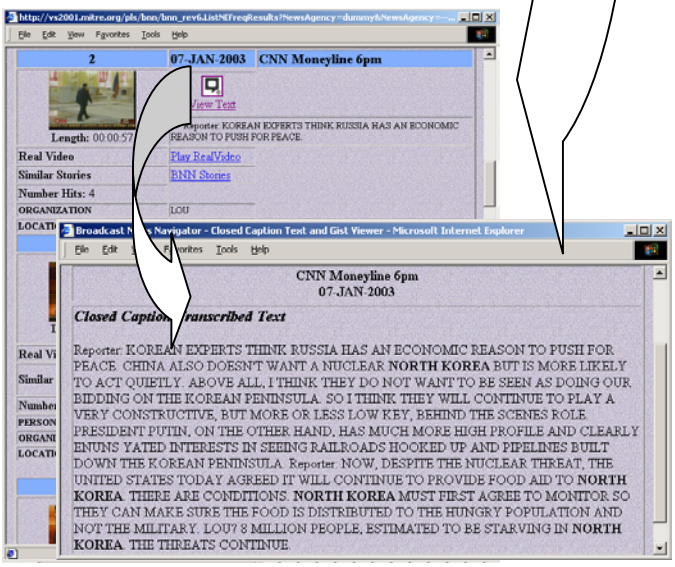

Figure 5. North Korean "Story Details" and Transcript

If the user scrolls and selects "North Korea" in Figure 3, they are provided access to the 39 stories detected from 6-9 January 2003 that contain this named entity. The results are displayed as a "Story Skim", shown in Figure 4. In this case the system shows a representative key frame from each selected segment, the source and date, and the top 3 named entities in each story. Note the stories range from several sources (e.g., CNN, CNN Money Line, CNN NewsNight with Aaron Brown), dates, and times of day. If the user is interested in China and North Korea and select the second story, the "Story Detail" in

Figure 5 is generated, from which they can navigate to a story transcript (displayed with highlighted entities), video source, or related stories.

Query Expansion, Relevancy Feedback

In order to enhance the precision and recall of user query, Local Context Analysis (LCA) (Xu and Croft 2000) is used both to expand the user's original query to the most related terms in the video corpus as well as to allow the user to provide interactive feedback to increase the relevancy of selected news stories. For example, if a user types the query "North Korea", based on an analysis of our corpus of broadcast news stories, LCA expands these two words into the rank ordered list of query expansion terms "nations", "monster", "iran", "security", "accounts", "pop", "kim", etc.

Figure 6 illustrates the user providing relevancy feedback indicating that they are interested in stories that talk about both North and South Korea. LCA finds co-occurring terms in selected documents. In empirical evaluations of a collection of 600 news stories from October 2002 (culled out of tens of thousands of stories from several years) from multiple program sources we found that iterative relevance feedback can enhance precision by nearly 20%, although the most dramatic gains (nearly 30% from a simple query baseline) occurred when users down-selected specific terms from expansion sets. While LCA query expansion terms helped users increase precision, they selected on average 10 relevant documents following their initial query which required time and cognitive effort.

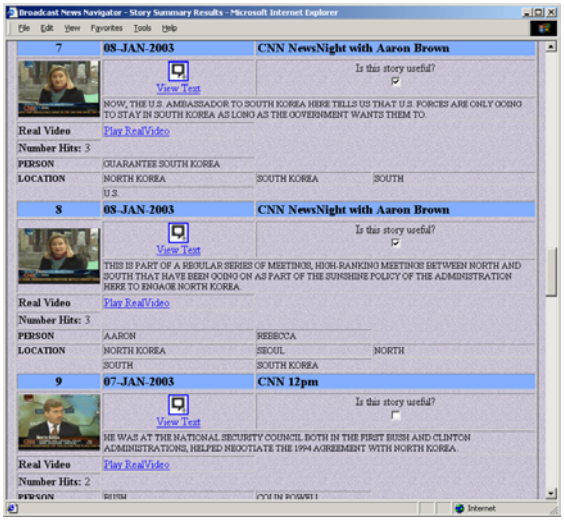

Figure 6. Stories about North and South Korea

Discovery

In addition to term expansion discovery, Figure 7 illustrates how BNN displays (and links to stories about) the most frequent names in any specified time period or source, in this case the first week of January 2003.

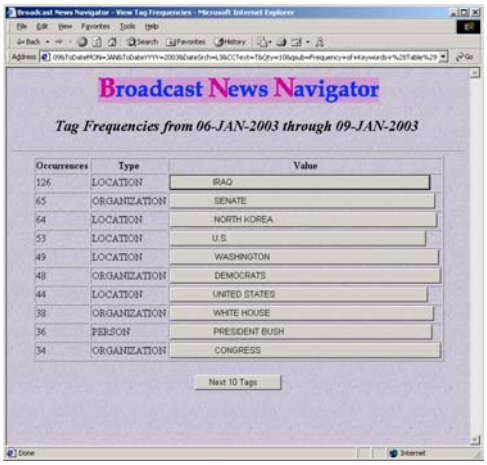

Figure 7. Most frequent January 2003 names

Beyond mention frequencies, a user may be interested in correlation among people, places, and organizations. Accordingly, we have extended association-rule mining to discover named entity classes co-occurrences (Tsur et al. 1998). As shown in Figure 8, users can interactively explore entity relationships within specified sources and date intervals. This enables a user to identify correlations among concepts in a news story corpus, identify topics in a collection of news articles, and discover two and three way entity relationships (e.g., the people most frequently mentioned with North and South Korean).

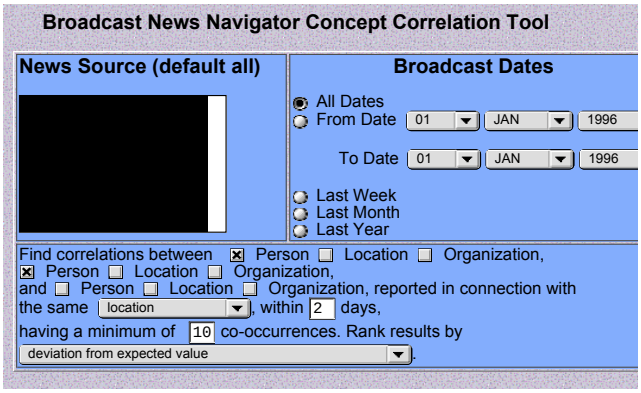

Figure 8. Association Rule Mining

Multimedia Summarization

BNN exploits BNE output to represent stories in a compact form using a range of devices. As described in section 3 and illustrated in Figure 4 and 5, stories are summarized by selecting the most frequent named entities, the sentences within a story that contain the most frequently occurring named entities, and the keyframes that are most representative of the story. As we detail in empirical evaluations below, providing search followed by story skim summaries followed by story details provides a hierarchy of increasingly detailed summaries, enabling rapid information discovery.

Media and Browsing Personalization

Because BNE segments stories and extracts media elements therein, BNN can not only retrieve content that is relevant to a user's information need, but it can also reassemble story components or browsing structure personalized to user preferences. Figure 9 illustrates the media and browsing preference profile in which the user can select what media elements to present when stories are retrieved (e.g., story length, video keyframe, text story summary, people, organizations or locations mentioned in the story). The user also can select what sources are available via hyperlinks (e.g., video, transcript, similar stories). Finally, the user can control if story details are presented directly following a query or if a skim summary (Figure 4) is presented first before navigation to story details (Figure 5). We are currently exploring the use of interface logging and machine learning to automatically learn these preferences by looking at associations among selected broadcasts, dates, and queries and the media and navigation selections users make during search and exploration.

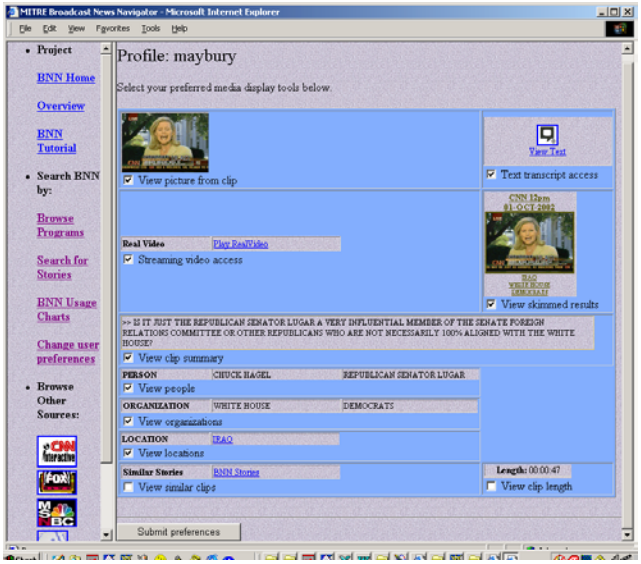

Figure 9. Personal Media and Browsing Preferences

Visualization

In addition to multimedia display of story content, temporal and geospatial visual displays were created to assist in trend and metadata analysis. Figure 10 illustrates a frequency over time trend analysis display of named entities and a geospatial display of story frequency occurrence. Both of these displays are dynamic. In the geospatial case, the location most associated with the story (e.g., in the headline or most frequently mentioned in the story body) is used to locate the story on a digital map. Also, stories are clustered exploiting named entity co-occurrences (Tsur et al. 1998). The user can then use a VCR-like controller to both manipulate the map as well as to move forward and backward in time to display story or cluster occurrences.

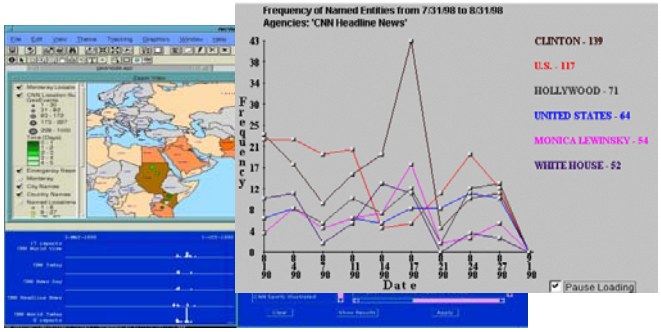

Figure 10. Entity Geospatial Display and Trend Analysis

Foreign Language Broadcasts

We have constructed models of several programs from foreign news to generalize our multimodal stream based approach to news understanding. Manual annotation of fourteen distinct foreign broadcasts including Indian (4), British (3), Chinese (5), and Russian (2). Because foreign broadcasts lack captions, audio and image classification is essential. Notably, we found 68% of foreign stories start and finish with an anchor shot.

System Evaluation and Findings

To test the value of the system, we evaluated video story retrieval as well as question answering tasks on a mix of 20 technical and non-technical users on 20 questions, measuring performance (precision and recall) and time, and then interviewing to assess usability. Varying sequence, we compared use of various combinations of the media elements described in Figure 9. Figure 11 shows average retrieval performance across all users for different media types for the story retrieval task. Discrepancies between the F-score and the precision and recall values in Figure 11 are due to the fact that when users did not record a response, the recall was scored as 0 and the F-score was not computed. We found:

- No difference in task accuracy between most mixed media presentations and original video.
- Providing less information to the user enabled them to discover content quicker without loss of accuracy.
- Precision was higher for high quality versus errorful sources, e.g. video (.94) versus closed captions (.80).
- Key frames have poor recall.

We also evaluated question answering from video, a task in which the user must extract some fact to answer one of ten questions. We found that, as hypothesized:

- Key frames alone are poor for question answering
- Story skim, all named entities, 3 named entities, topic, and summary presentations result in average answer miss rates (from 55% to 65%).

- Text, full details, story details, and video result in very low miss rates (from 0% to 10%).
- Watching source video takes the most time to view; story details, text, full details, skim, and topic presentation methods take moderate viewing time; key frame, all named entities, 3 named entities, and summary presentation methods take the least amount of time.

Summary Type	Precision	Recall	F-Score	Time (Secs/-Story)
Story Details	0.93	0.97	0.94	10.23
Full Video Source	0.94	0.94	0.93	24.53
Key Topics	0.95	0.73	0.92	6.41
Full Story Details	0.91	0.96	0.92	8.28
Summary Text	0.88	0.77	0.86	3.57
Full Text Transcript	0.80	0.81	0.84	13.58
3-Named Entities	0.88	0.60	0.82	2.08
Story Skim	0.93	0.74	0.81	3.69
All Named Entities	0.81	0.68	0.78	3.54
Key Frame	0.84	0.21	0.54	3.42
Mean	0.82	0.70	0.78	6.92
Standard Deviation	0.25	0.25	0.22	6.66

Figure 11. Retrieval Performance Evaluation

After each test, users subjectively assessed their preferences for presentation methods on a Likert scale of 1 (not preferred) to 10 (most preferred) for both story retrieval and question answering tasks. For story retrieval, the average rating for the story details and full detail methods (7.8) was higher than the average rating for other presentations (5.2). For question answering from video, on average, users preferred story detail or full detail presentation methods about twice as much as other methods (8.2 versus 4.0). On average answering questions took twice as long to perform as story identification.

In summary and most significantly, users of the best combination of BNN presentations performed nearly three times as fast as users of the source digital video. As a consequence, the default user interface in the deployed system hyperlinks story skims to story details.

Deployment and Transition

In addition to controlled studies, the prototype system has been available on line at The MITRE Corporation for several years and a version was deployed to a government organization responsible for information awareness and dissemination. Servers run on Windows NT and Sun Solaris accessible via a standard browser, using analog MPEG as input, storing extracted information in an Oracle RDBMS and Video Server, delivering output as RealVideo or MPEG and indexing video streams in 2.5x real-time. Qualitative user assessments suggest that:

- BNN is an effective system for browsing days or weeks of news to find stories or answer questions.
- BNN is valuable for alerting users and superiors to unexpected stories/relationships for further analysis.
- Trend analysis, story clustering, and geospatial displays provide useful overviews.
- Named entity performance degradation as a result of speaker transcription or closed captioning errors is tolerable by algorithms and users.
- There is a need to cluster temporally and topically neighboring stories to address oversegmentation.

Over the past several years, the system has been demonstrated to several national broadcasters who have embraced the concept, primarily for news reuse. Finally, several of the underlying algorithms have found their way into a commercial video processing product.

Acknowledgments

BNE/BNN contributors include Stanley Boykin, Andy Merlino, Warren Grieff, David Palmer, Chad McHenry, John Aberdeen, John Burger, David Day, Lynette Hirschman, Marc Light, Chris Clifton, and Rod Holland.

References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. 1995. Description of the Alembic System Used for MUC-6. In Proceedings of the Sixth Message Understanding Conference, 141-155. ARPA/ITO, 6-8 November Columbia, MD.

Boykin, S. and Merlino, M. Feb. 2000. Machine Learning of Event Segmentation for News on Demand. *Communications of the ACM*. Vol 43(2): 35-41.

Maybury, M. 2002. News on Demand. In Rahman, S. M. (ed.) *Multimedia Networking: Technology, Management and Applications*. Idea Group Pub.: Hershey, PA, 126-133.

Merlino, A. and Maybury, M. 1999. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. Mani, I. and Maybury, M. (eds.) *Automated Text Summarization,* MIT Press.

Palmer, D., Burger, J. and Ostendorf, M. 1999. Information Extraction from Broadcast News Speech Data. DARPA Broadcast News Workshop, 28 February – 3 March 1999, Herndon, VA.

Tsur, D., Ullman, J., Abiteboul, S., Clifton, C., Motwani, R., Nestorov, S. and Rosenthal, R. 1998. Query Flocks: A Generalization of Association-Rule Mining. In ACM SIGMOD, 1-12.

Xu, J. and Croft, W.B. 2000. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems* 18(1):79–112.