Preface

Mark Maybury

Significant advances have been achieved over the past decade in language processing for information extraction from unstructured multilingual text (see, e.g., trec.nist.gov). However, the advent of increasingly large collections of audio (e.g., iTunes), imagery (e.g., Flickr), and video (e.g., YouTube) together with rapid and widespread growth and innovation in new information services (e.g., blogging, podcasting, media editing) is driving the need not only for multimedia retrieval but also for information extraction from and across media. Scientists and engineers are innovating in new challenges such as multimodal sentiment analysis, multimodal summarization, and collaborative multimedia editing. While largely independent research communities have addressed extracting information from single media (e.g., text, imagery, audio), to date there has been no forum focused exclusively on cross media information extraction.

Objectives

The AAAI Fall Symposium presents a unique opportunity to move toward an integrated view of media information extraction. The language processing, speech processing, image/video processing and spatial/temporal reasoning communities have much to offer and much to learn from one another. The purpose of this symposium is to bring together researchers and practitioners to report on current advances in multimedia information extraction algorithms and systems and their underlying theories, to foster scientific interchange among these individuals, and as a group to evaluate current efforts and make recommendations for future investigations. An edited collection to include extended versions of the best papers is planned.

Organizing Committee

We have been fortunate to have a highly qualified and diverse program committee which spans the many disciplines involved in multimedia information extraction with representation from the government, industry, academia and non profits including:

- Kelcy Allwein, Defense Intelligence Agency, Kelcy.Allwein@dia.mil
- Elisabeth Andre, University of Augsburg, andre@informatik.uni-augsburg.de
- Thom Blum, Muscle Fish, a division of Audible Magic thom@musclefish.com
- Shih-Fu Chang, Columbia University, sfchang@ee.columbia.edu
- Bruce Croft, University of Massachusetts, croft@cs.umass.edu
- Alex Hauptmann, Carnegie Mellon University, alex+@cs.cmu.edu
- Mark Maybury (chair) The MITRE Corporation, maybury@mitre.org
- Andy Merlino, Pixel Forensics, amerlino@pixelforensics.com
- Ram Nevatia, University of Southern California, nevatia@iris.usc.edu
- Prem Natarajan, BBN, prem@bbn.com
- Kirby Plessas, Open Source Works, kirbyp@open-source-works.org
- David Palmer, Virage, dpalmer@virage.com
- Mubarak Shah, University of Central Florida, shah@cs.ucf.edu
- Rohini K. Shrihari, SUNY Buffalo, rohini@cedar.buffalo.edu
- Oliviero Stock, Istituto per la Ricerca Scientifica e Tecnologica, stock@fbk.eu
- John Smith, IBM T. J. Watson Research Center, jsmith@us.ibm.com
- Rick Steinheiser, DNI/Open Source Center, RickS@rccb.osis.gov
- Sharon Walter (co-chair) AFRL, Rome NY, Sharon.Walter@rl.af.mil

We thank each for their important contributions in shaping the Symposium.

Approach

The Committee took advantage of the AAAI Symposium format to go beyond a traditional series of presentations to address issues that cut across disciplines to address processing multiple media (e.g., text, speech, maps, imagery, video) and human perceptual modalities (e.g., audition, vision). There are few if any multidisciplinary venues to integrate the multiple subcommunities working in this important

emergent area. The organizing committee hopes this forum will accelerate progress by helping form and shape this new community. Three invited speakers were chosen to illustrate cross media representation and reasoning challenges and promising technical approaches. Contributions included invited talks, papers, posters, demos, and roadmap statements that serve as input to the group technology strategy discussions. Selected presentations were organized into sessions addressing specific challenges such as language extraction, graphics extraction, video extraction, and multimedia authoring. Breakout sessions and discussions are planned to address cross cutting topics such as (cross)media data sets, multimedia machine learning algorithms, innovative architectures, transmedia applications, and evaluation methods.

Roadmapping

While prediction is always challenging, to increase the lasting value of this unique event, a road map of future technology developments over time was created. Building on a set of road map contributions contributed by organizing committee members, the participants in the symposium elaborated and expanded upon these during the discussions. The group captured capabilities and challenges into a unified roadmap containing lanes (e.g., multimedia data, methods, and applications) and used these to stimulate iterative brainstorming "roadmap" sessions throughout the symposium to further the outcome. Beyond the technical content, a primary purpose of this is social: to create a shared view on common challenges and directions and to stimulate trust, relationship and community building.

State of the Art

Individual media extraction performance varies broadly. Text extraction is the most mature are with extraction of many classes of objects (e.g., people, places, things, time), many relations (e.g., part of, property of, related to), and some events (e.g., physical, economic, political, social) being possible with commercial tools, increasingly in languages beyond English (e.g., Spanish, Chinese, Arabic). Typically these entities, with references back to the original text sources, are captured in an independent data or knowledge base. Spoken language transcription, affected by the quality of the environment, channel, and speaker(s), can support similar if slightly degraded extraction. Analysis of non speech audio can support the extraction of sounds with specific acoustic (e.g. pitch, duration, loudness) or perceptual properties (e.g., scratchy, buzzy or tinny sounds). With respect to still imagery, extraction of objects (e.g., people, faces, animals) and their relations (e.g., next to, above) are also within the state of the art. Video extraction is typically limited to object detection and tracking (e.g., people, animal, moving object) where complex event detection (e.g., exchange of an object between two people, aircraft landing, building implosion) is typically restricted to special purpose crafted systems. Relatively few systems explicitly address cross media processing and extraction. In addition to establishing common capabilities (e.g., multimodal biometrics for identity), higher level social phenomena (e.g., social relations, sentiment, deception) may be better enabled with cross media processing and extraction.

Research is needed in all aspects of multimedia information extraction, including, but not limited to the following fundamental problems:

- Object, attribute, and relation extraction from media (e.g., text, audio, maps, imagery, and video)
- Simple and complex event detection and extraction from text, audio, imagery, and video
- Integrated speech, language, and image processing methods for cross media information extraction (i.e., transmedia information extraction)
- Emotion and sentiment detection and tracking from media
- Tailoring multimedia information extraction to particular users, tasks, and contexts
- Intra- and inter-media representation languages and cross media ontologies
- Architectures for multimedia information extraction
- Constraints and capabilities of IE components and their integration
- Psychoperceptual and cognitive issues in multimodal information extraction
- Multimedia browsing/visualization tools and cross-media query
  (e.g., visual, linguistic, and auditory)
- Studies and analyses of multimedia corpora
- Annotation schemes and tools for multimedia information extraction
- Evaluation methods and metrics for multimedia information extraction
- Innovative machine learning approaches to multimedia information extraction

Grand Challenges

This area presents several grand challenges which could motivate joint research efforts

- ***Video understanding*** – automated extraction of content from broadcast, satellite, cable video, mobile phone or videoconferencing to enable search, retrieval, visualization, and summarization.

- ***Audio understanding*** – automated transcription, characterization, extraction and summarization of audio content.
- ***Image and Graphics understanding*** - automated extraction, analysis, and summarization of imagery and graphics.
- ***Integration*** of media extraction streams and algorithms to support cross media cueing, understanding, and fusion. For example, the parallel application of rhetoric structure theory, music theory, and cinematography might be used to understand complex multimedia artifacts. Modality integration can address mutual disambiguation, that is the integration of multiple ambiguous messages to resolve ambiguity. It also possible for one modality to enhance another dominant modality, for example to use of gesture to emphasize speech. Finally a message conveyed by one modality can modulate or alter the content of a message conveyed by another modality, for example, the way a facial expression can influence the meaning of an utterance.
- ***Automated multimedia presentation design*** - the automated selection of content, structuring and sequencing of coordinated multimedia elements to create an integrated effect possibly employing affective communication

Potential Benefits

The Symposium also aims to better understand if not quantify and motivate the benefits of this technology.

- ***Performance Enhancement***: Because of various maturities of extraction technologies across various media, where artifacts are multimodal, it is possible that the performance of less mature extractors can being boosted using cross media processing. For example, visual analysis of lips can enhance speech recognition from a noisy channel. Similarly, acoustic analysis has been used to enhance segmentation of broadcast video.
- ***Quality***: The completeness and quality of extracted media output and processing can be enhanced by exploiting redundancy and complementarity of media. Media streams that are redundant can be used to increase the accuracy of recognition and complementary modalities can each convey only part of a message but their integration results in a complete message.
- ***Utility***: The practical implications of cross media processing for a variety of uses from analysis, to design (using cross media artifacts), to education and training
- ***Discovery***: Cross media processing and phenomena may enable new insights (e.g., multimodal biometrics for identity management, cross media analysis for more robust bias and/or deception detection)
- ***Enjoyment***: Cross media presentations may enhance the level of richness and engagement over unimodal artifacts.

Conclusion

Multimedia information extraction is an exciting area, both from a research and practical perspective. Our lives can be encroached on or enriched by multimedia. Our collective challenge remains to advance the underlying algorithms that can enable information producers and consumers alike to enhance the quality and value of media they produce and use. Properly done, it may also unlock new unknown value in collections and communities.