Conversation Clusters: Grouping Conversation Topics through Human-Computer Dialog

Tony Bergstrom and Karrie Karahalios
University of Illinois at Urbana Champaign
{abergst2, kkarahal}@cs.uiuc.edu

ABSTRACT

Conversation Clusters explores the use of visualization to highlight salient moments of live conversation while archiving a meeting. Cheaper storage and easy access to recording devices allows extensive archival. However, as the size of the archive grows, retrieving the desired moments becomes increasingly difficult. We approach this problem from a socio-technical perspective and utilize human intuition aided by computer memory. We present computationally detected topics of conversation as visual summaries of discussion and as reference points into the archive. To further bootstrap the system, humans can participate in a dialog with the visualization of the clustering process and shape the development of clustering models.

Author Keywords

Clustering, Meeting Archival, Visualization

ACM Classification Keywords

H5.3 [Information interfaces and presentation]: Group and Organizational Interfaces - Synchronous interaction, Computer supported cooperative work.

INTRODUCTION

This paper explores human augmented computer visualization as a means to leverage human understanding in recalling past conversation. Specifically, we look at co-located meetings of groups where, traditionally, a designated scribe records minutes of key meeting points. Our target meetings have an expectation of being re-examined in the future; people refer to these minutes to review what was said.

Many current computational approaches to meeting archival mirror the function of a courtroom stenographer; they document the entire discourse of the event. Searching through these archives requires a lookup of key words, while understanding the entire discourse requires reading the entire record. The lookup problem is made more difficult as automated speech recognition transcripts are not always accurate. We mitigate these effects by mimicking human memory to forget the actual verbal exchange but retain a generalized understanding through the use of key words. By focusing on the ideas conveyed and forgetting some speech, our memories better recall the most recent and relevant moments in conversation [11].

CHI 2009, April 3 - 9, 2009, Boston, MA, USA.

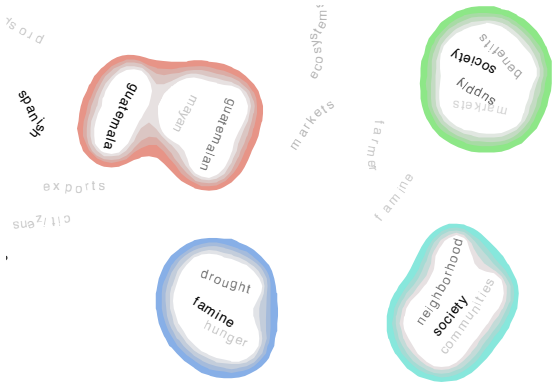

Figure 1. Conversation is visualized as clusters of words grouped by topics. Each bubble indicates words that had been grouped together. Singular words were from other potential topics float freely before they disappear or form new clusters.

The key to making our approach work is combining the strengths of man and machine. The computer has the ability to store large archives; people have the ability to augment their judgment based on the context at hand. We present two archival visualizations implemented to summarize interaction: a topic view and a history view. The topic view serves to see the topics discussed over the course of the meeting while the history view allows the viewer to see how the meeting progressed by mapping the evolution of topics over time.

Leveraging People and Computers

Understanding conversation is hard. People can readily understand speech in spite of mumbling, self correction, and ambiguous references; for computers this is more difficult. The best speech recognition system to date can produce a 3% word error rate on transcripts, though more common conditions incur 20-30% error rates [12]. Aside from detecting the words to transcribe, proper understanding of speech requires understanding semantic meanings and context, a task that is difficult even with a perfect transcription of the most formal language [1]. Natural Language Processing (NLP) can help in the analysis and understanding of language, but the techniques are often limited to understanding and detecting statistical patterns in language in specified domains. Altering the conversation's topic, word choice, or context from that of the training set used to create the NLP model drastically

reduces automated comprehension for the generalizable situation [15].

Conversation Clusters attempts to bridge the verbal language barrier by using humans and machines. In the domain of meetings, people understand conversations better than computers. They can infer the context of social situations, humor, sarcasm, body language, etc. Computers provide significant advantages when it comes to tracking history and storing the numerous details of interaction. By combining these strengths, we seek to provide a reliable and intuitive system for archiving and re-accessing conversations.

Capturing Social Experiences

Written language complements oral discourse as a means to exchange ideas [14]. The predominance of literate culture in recent history has resulted in many changes in the way we remember and think as a society: written records act to accurately detail events and easily disseminate information. One of the most common artifacts of aural meetings is the transcript. They are easy to archive, store, and reread. Computer generated transcripts using speech recognition can capture verbal exchange in a similar form. They, too, are archivable and searchable, but they are prone to errors in recognition.

With the onset of near real-time speech recognition, researchers have explored storing days and weeks of personal transcripts generated with mobile wearable microphones that continuously record. Integrated into daily life, the microphone produced an extensive database that could be stored and searched [17]. With no specific query in mind, a cursory understanding of a transcript relies on personally skimming exact text. To date, methods such as highlighting words based on speech recognition confidence or TFIDF scores have been used to aid in exploring these transcripts [2, 16].

In meeting archival systems, researchers have automated detection of visual and audio cues to summarize and index video [8, 9]. These environments are focused on reexperiencing the raw captured media to review events. Students have benefited from similar automated capture when reviewing lectures with the eClass system [5]. Other systems, such as Google's Audio Indexing[1], have applied speech recognition to automatically create searchable indices for a video recording.

Many applications try to mitigate human involvement in computational tasks such as parsing and object recognition. More recently, popular websites like del.icio.us[2] utilize human knowledge to tag, label, and classify the web. Smaller groups use indices within a single document or video to easily locate useful content[10]. All of these systems rely on user participation. Some, like peekaboom.org, have built games to entertain participants while capturing this information [18]. Our goal with this work is to combine the computed topics with human interpretation to provide a topic labeled image of conversation.

Influence of Social Displays

We propose using a visualization in the form of a community display to present the group status and activity of a meeting. Prior work in this area has demonstrated that such displays externalize the groups' progress, allowing it to focus on core ideas, deliberate more, and produce higher quality output [13].

Work with conversation visualization has produced greater interaction awareness among participants [3, 6]. Subtle feedback, such as showing individual vocal participation, promoted equal contribution during conversation. Though these visualizations focus strictly on the aural interaction, other visualizations have begun exploring content by using speech recognition to detect topics and then suggesting related avenues for discussion [4].

These words suggest people can extract meaningful interaction information from shared displays during conversation by providing augmented group awareness. We aim to provide similar feedback with *Conversation Clusters*.

CLUSTERING CONVERSATION

Conversation Clusters approaches dialog as a stream of ever changing topical threads. Rather than focusing on reproducing an error free word-for-word, transcript, we detect general topics from the words that we can reliably recognize. A dynamic visualization on a shared public tabletop enables participants to see the most recent discussion topics as clusters of words grouped together (Figure 1). Users can also switch to a corresponding timeline visualization to explore the birth and death of topics. With *Conversation Clusters*, we develop a representation of conversation that captures the context of the moment.

Clusters

The first visualization demonstrates a clustering of current and recent topics (Figure 1). Projected directly on the table, the visualization provides an ambient awareness of the computer's topic model which can be manipulated via Mitsubishi's DiamondTouch table.

As the visualization occupies public conversation space, it draws attention from the participants to the up-to-date depiction of salient topics. Though there is potential for distraction, previous work indicates conversation develops naturally as speakers prioritize their attention to conversation and listeners examine the visualization when convenient [3, 6]. We further mitigate accidental distraction by using slow fading transitions and reusing visual elements in the display. If a word is being replaced in a cluster, the visualization will fade from one word to another rather than add one word and subtract another. However, clusters themselves are reused only in cases where the new topic is sufficiently close to the old (we discuss the distance metric in a later section). Emerging topics in conversation that cannot be paired appear as a new cluster. As a group moves into new topic areas,

[1] http://labs.google.com/gaudi
[2] http://del.icio.us

the shared display demonstrates the changing focus of the meeting with multiple new topic clusters.

Words unrelated to the highest rated clusters move freely and unbound in the visualization. These words may be from a previous topic that is no longer being discussed or an emerging topic that may soon appear on the table. With the interactive tabletop, participants can increase the strength of a link between the lesser topics and current clusters by dragging the floating words into clusters. With no interaction, the words simply demonstrate the changing pattern of topics. As speech recognition and topic detection algorithms are not perfect, erroneous words and glaring errors do appear and may be removed by pressing a full hand to the word. In this manner, people present can influence topic formation and topic removal.

Thread History

The thread history visualization shows a historical overview of salient topics in discourse over time. As seen in Figure 2, threads of conversation begin, merge, separate, and end along a timeline from left to right. The computer diagrams history by creating and displaying this model of conversation evolution.

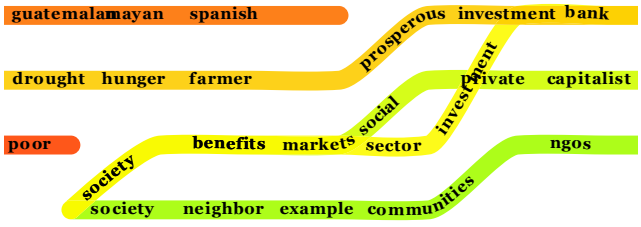

Figure 2. The thread history visualization demonstrates the progression and change of conversational topics.

In this historical view, multiple threads appear in parallel. This does not mean that four or five topics are being separately addressed at once, but demonstrates related topics that are not clustered together. For example, a discussion about car repair might have a thread about cars and another about the financial means of payment. In the context of the conversation they are related so they appear simultaneously. However, the words that generate these threads do not sufficiently overlap to form a single cluster or thread. Additional parallel threads appear as the transcript grows in size. The thread history visualization examines the relation of clusters over time, and the window of potential related clusters grows proportionally. The interface allows for interactive exploration. Longer conversations are viewed at a higher level; participants can zoom in to see a more detailed view. They can also switch between the timeline and cluster modes of the table visualization.

The timeline visualization creates an abstracted view summary to aid recall by filtering words upfront. The full dialog cannot be recreated from the visualization. Threads of conversation use individual salient words to characterize the evolving topic. The selected words allow viewers to infer the general meaning but provide privacy for the individual speakers.

Implementation

Before detecting topics, we must be able to extract words from a conversation. Our initial design calls for a dedicated microphone for each speaker. We use an off the shelf copy of Dragon's Naturally Speaking[3] to provide a general transcript before clustering meaningful word topics.

Both the clusters and timeline visualizations use a sliding window of the conversation transcript to generate topical words. A short window in the cluster visualization ensures only recent words appear. The timeline visualization window varies: a longer window shows more general trends while a smaller window captures shorter lived topics. The initial topic clusters are generated from the first utterances of a group and are prone to fluctuation while the conversation fills the first sliding window.

Related words are clustered by querying an Explicit Semantic Analysis (ESA) server with the windowed transcript [7]. ESA uses a stored version of Wikipedia as a base collection of topics, each represented by an article. Probable topics are identified by comparing the distribution of words in articles to those in the query, and a list of articles are returned with a weighted list of words that support the article (Figure 3). The Wikipedia article titles are less informative for our purpose than supporting words. For now, we ignore the title and use the lists as a proxy for relatedness of words.

Query: How do people interact with others using computers? How can we augment normal experience?

Classifications:

Augmented reality (1299.7)
augment (58.8) interact (26.4) comput (21.9)

Many-worlds interpretation (1039.3)
interact (24.8) augment (22.3) experi (16.0) comput (12.8)

Motion capture (1003.7)
augment (23.1) comput (22.2) normal (13.5) interact (12.7)

Virtual reality (913.2)
comput (24.5) interact (17.6) experi (15.6) augment (15.1)

Game theory (862.0)
interact (22.6) comput (19.7) normal (19.3) experi (16.1)

Figure 3. An example of a ESA query and results.

The ESA results form the basis of the topic detection that feeds our visualization. In our pilot study, the Wikipedia article word distributions used by ESA eliminated erroneous words recorded in the transcript. These words were culled as it was unlikely for the erroneous words to appear in the same article as many of the correctly recorded words.

Even with groups of related words, we must combine and prune our results to avoid overwhelming participants in the shared visualization space. As seen in Figure 3, the query returns a list of Wikipedia articles with overlapping sets of supportive words and an overall score of relevance to the query. To avoid redundancy we combine topics using a weighted sum of supporting words (Figure 4). We limit inclusion in the visualization to results that are in the top

[3]http://www.nuance.com/naturallyspeaking

200 red (25) white (19) blue (3)
150 color (25) pattern (20) red (5)
100 red (10) white (10) green (6)

⇓

200 red (20) white (16) blue (2) green (2)
150 color (25) pattern (20) red (4)

Figure 4. ESA results with significant overlap are combined into similar groups. The weight of the topic weight is the highest weight of the combined topics. Individual words are combined via weighted sum. For 'white,' this calculation is: $(200 \cdot 19 + 100 \cdot 10) \div (200 + 100) = 16$.

weighted results. The topics are combined after calculating a distance between all pairs. The distance is a relative difference between the word weights of the top words; sharing words indicates a much closer match. The pairs are combined until a maximum cutoff distance is exceeded.

As stated earlier, this is a socio-technical system. People and machines work together to influence the clustered visualization. The first, more subtle, human-computer dialog occurs when people edit Wikipedia. ESA then clusters the text based on the word distribution in this socially generated reference. Therefore, a change to Wikipedia affects the clustering each time the ESA back-end is updated. The second human-computer dialog occurs when people interact directly with the table to remove or swap words between clusters. The altered clusters serve to direct the associations made in future clusters.

LIMITATIONS AND FUTURE WORK

Conversation Clusters is not a solution for all archiving situations. Though Wikipedia provides a vast and expansive set of topics, we recognize it does not encompass the whole of all possible topics. Additionally, the system cannot infer basic information that would be obvious to a human listener such as cause and effect. Another point we wish to make is that people do forget, and this act of forgetting serves a purpose. It is not the everyday interactions at the coffee shop that we target. *Conversation Clusters* is aimed at conversations that anticipate a retrospective analysis: a board meeting, a public record, or a series of interviews.

In future work, we hope to include imagery to support individual contribution and conversational context to provide a picture of conversation featuring the interplay of actors and their roles. As seen in previous works [3, 6], participant contribution can serve as meaningful imagery to understand interaction. Incorporating personal interaction patterns, we can hope to demonstrate the meaningful contributions and information persona of each individual. Finally, we plan to study this tool used in meetings at a nearby consulting firm.

CONCLUSION

As society archives an ever increasing amount of material, searching through the vast amount of data and uncovering salient moments becomes increasingly difficult. *Conversation Clusters* illustrates techniques that can be used to identify relevant moments in otherwise unstructured archival by combining human intuition and inference with the computational affordances of computers to create a human-computer dialog.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation, NSF-0643502.

REFERENCES

1. J. Allen. *Natural language understanding*. Benjamin-Cummings Publishing Co. 2nd ed, 1995.
2. S. Basu, S. Gupta, M. Mahajan, P. Nguyen, and J. C. Platt. Scalable summaries of spoken conversations. In *Proc of IUI*, 267–275, 2008. ACM.
3. T. Bergstrom and K. Karahalios. Seeing More: Visualizing Audio Cues. In *Proc of Interact*, 2007.
4. K. Böhm and L. Maicher. *Real-Time Generation of Topic Maps from Speech Streams*, 112–124. Springer Berlin / Heidelberg, 2006.
5. J. A. Brotherton. *Enriching Everyday Experiences through the Automated Capture and Access of Live Experiences: eClass: Building, Observing and Understanding the Impact of Capture and Access in an Educational Domain*. PhD thesis, Georgia Tech, 2001.
6. J. M. DiMicco, A. Pandolfo, and W. Bender. Influencing group participation with a shared display. In *Proc of CSCW*, 614–623, 2004. ACM.
7. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proc Artificial Intelligence*, 2006.
8. L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *Proc of Multimedia*, 489–498, 1999. ACM.
9. A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proc of Continuous archival and retrieval of personal experiences*, 74–85, 2004. ACM.
10. V. Kalnikaité and S. Whittaker. Social summarization: does social feedback improve access to speech data? In *Proc of CSCW*, 9–12, 2008. ACM.
11. J. B. Miller, P. deWinstanley, and P. Carey. Memory for conversation. *Memory*, 4(6), 1994.
12. C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc of CHI*, 493–502, 2006. ACM.
13. J. S. Olson, G. M. Olson, M. Storrøsten, and M. Carter. Groupwork close up: a comparison of the group design process with and without a simple group editor. *ACM Trans. Inf. Syst.*, 11(4):321–348, 1993.
14. W. J. Ong. *Orality and Literacy: The Technologizing of the World*. Routledge, 2002.
15. R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proc of IEEE*, 88(8), 2000.
16. S. Vemuri, P. DeCamp, W. Bender, and C. Schmandt. Improving speech playback using time-compression and speech recognition. In *Proc of CHI*, 295–302, 2004. ACM.
17. S. Vemuri, C. Schmandt, and W. Bender. iRemember: a personal, long-term memory prosthesis. In *Proc on CARPE*, 65–74, 2006. ACM.
18. L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proc of CHI*, 55–64, 2006. ACM.