Tracking Conversational Context for Machine Mediation of Human Discourse

Tony Jebara, Yuri Ivanov, Ali Rahimi, Alex Pentland
MIT Media Lab, 20 Ames St., Cambridge, MA 02139
{ *jebara, yivanov, rahimi, sandy* } *@media.mit.edu*

Abstract

We describe a system that tracks conversational context using speech recognition and topic modeling. Topics are described by computing the frequency of words for each class. We thus reliably detect, in real-time, the currently active topic of a group discussion involving several individuals. One application of this 'situational awareness' is a computer that acts as a mediator of the group meeting, offering feedback and relevant questions to stimulate further conversation. It also provides a temporal analysis of the meeting's evolution. We demonstrate this application and discuss other possible impacts of conversational situation awareness.

The Computer as a Mediator

The recent progress of the Human-Computer-Interface community has produced a variety of intelligent interfaces to assist humans in their interaction with a computer. In the process, various supporting technologies are called upon (tangible interfaces, audio-visual modalities, etc.) to produce a more natural look-and-feel to the computer. However, the emphasis has been primarily on improving the interaction between a single human and a computer.

We propose an alternative paradigm where the computer acts as an external third party. It observes several humans interacting with each other and assists them as an external mediator. Our specific application is a group meeting where individuals are actively involved in a discussion. The computer is no longer the focus of attention and must autonomously track the conversation, provide feedback and augment the meeting. As 'external mediator' the machine must have some situational awareness or understand the conversational context such that it intelligently enhances and assists the overall interaction. A human facilitator has a similar task. He can be assistive without being an expert with background knowledge or detailed understanding of the group meeting.

Another constraint is that a machine should be autonomous such that it doesn't encumber the individuals, restricting their conversation by requiring direct input and manipulation. Therefore, the computer must participate in a non-obtrusive manner, using audio-visual or passive modalities that allow the participants to continue carrying out a conversation without browsing menus, using a GUI or getting slowed down by the 'mediator'. Since it is not expected to directly replace one of the participants (i.e. as an automated bank teller), the computer does not have to carry the conversation itself . It instead stimulates it further and provides relevant information when appropriate contextual cues arise from the conversation between the humans.

Figure 1: Current mediation system setup.

System Overview

The system is portrayed in Fig. 1. It consists of a number of microphones (either head-mounted or clip-on, wireless mikes), a video camera and a large projection screen. The users sit at a meeting table and engage in a conversation. While they speak, the microphones feed a commercial-grade speech recognizer which detects words. The frequency of the past few words is used to compute the general topic area of the conversation and establishes the situational context. In addition, the video camera detects frontal faces using computer vision. When either of the users looks at the screen, the computer gets a reinforcement signal. Feedback from the computer mediator is generated using the projection screen which asks relevant questions and can provide other audio-visual augmentations to the current conversation. To train the topic spotter, we obtain text documents that are representative of each topic class we wish to track (i.e. 'medicine', 'politics', etc.). The system analyzes the statistics of these

text documents to later detect the currently active subject from the words the users generate. For example, if the users are talking about medicine, key words such as 'doctor', 'cancer', etc. will be detected to isolate the topic.

Background

Related work in conversational awareness is found in (Franklin, Bradshaw, & Hammond 2000). Here, a system tracks the topic using a speech recognizer to automatically change slides in a power-point presentation. Techniques for topic-spotting arise in the text classification community which uses typed (vs. speech recognized) words to classify Reuters documents, etc. (McCallum & Nigam 1998). Wearable audio systems track audio context at a coarser signal level in (Clarkson, Sawhney, & Pentland 1998). Finally, wearable keyboard based systems track typed words to pull up relevant reference documents automatically (Rhodes & Starner 1996). Such systems have a coarse understanding of the conversational or textual context and this enables intelligent output that is relevant to the situation.

Technological Issues

As noted, we utilize a computer vision system as well as audio. A small video camera is positioned such that it can detect frontal faces when the users look at the screen. That way, the system can know when its feedback is being elicited and can tell **when** to produce output. The computer vision system searches for skin-colored blobs and detects frontal faces using eigenspace techniques. It is a variant of the system in (Jebara & Pentland 1997), and generates a 1Hz signal indicating how likely a frontal face is present in the image.

Despite many advances in speech recognition the technology is still brittle, preventing its proliferation in the HCI community. Therefore, it is unlikely a contemporary system could accurately respond to natural communication on a word-by-word basis. To circumvent this problem, we look at the frequencies of the past few words (i.e. 200 words) to determine the topic. Therefore, despite poor accuracy in the speech recognizer (as low as 50%), the aggregate performance over a set of 200 words for topic-spotting (vs. word-spotting) has an accuracy in the high 90's. The topic-spotting is used to tell **what** feedback to generate. Furthermore, the mediator system is used in meetings where people converse actively generating *many* words. This is a far better situation for a recognizer than when the computer is interacting with a single user and has to recognize a sentence at a time in a turn-taking situation. Such query-response systems are far too brittle except in constrained applications.

Implementation

Our current system uses live audio input to detect the topic of the conversation. A speech recognizer (IBM's ViaVoice SDK) is run in continuous dictation mode, producing candidate words in real time. The list of recent words is matched against a set of trained topic models to compute their likelihoods. Topic probabilities are continuously compared to each other and the maximum one is used to determine the conversational situation and provide feedback.

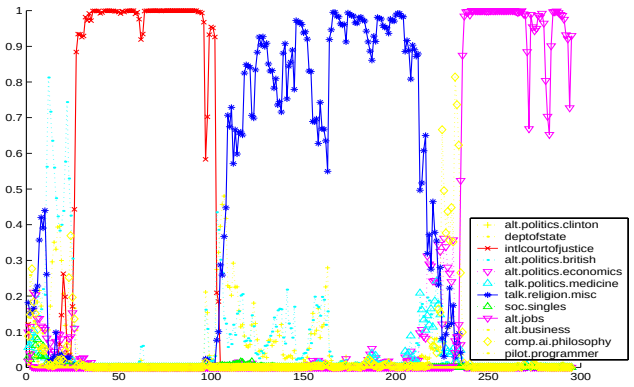

Figure 2: Plot of class probabilities.

Model Training

The model used is a multinomial pdf (bag-of-words) computed from the counts of unique words for each topic:

$$P(word_i|c) = \frac{N_c(word_i)}{\sum_{j=1}^{w} N_c(word_j)} \quad (1)$$

where $P_c(word_j|c)$ is a probability of a particular word coming from class c. In addition, $N_c(word_j)$ is how many times this word was encountered in the training corpus for the class c. The total number of unique words is w. This model is a word frequency model which is guaranteed (in the maximum likelihood sense) to converge to true word probabilities given a large training corpus. We train the system for 12 different conversation topics from the web and newsgroup text documents in Fig. 2.

Topic Classification

After training data is collected and class models are built, the system begins receiving audio input from speakers. A matching algorithm sequentially updates a conversation history ($\mathbf{x}$) which counts the frequency of most recently spoken words and weights them by their recency (which is slowly decaying). The conversation history $\mathbf{x}$ (i.e. a 30,000 dimensional vector of counts of past words), is updated at each step after receiving a new word $word_k$ by decaying $\mathbf{x}$ and adding a count of one for the new word:

$$x_i^t = \alpha x_i^{t-1} + \delta(k,i) \quad (2)$$

where α is the decay parameter, $\delta(k,i)$ equals 1 if the the $word_k$ is the same word as x_i (i.e. $i = k$). Given the conversation history at time t, its class-conditional probability is computed as follows:

$$P(\mathbf{x}|c) = \prod_i P(word_i|c)^{x_i} \quad (3)$$

This probability is converted into the posterior for topic c using Bayes' rule. The prior probabilities $P(c)$ are scalars (one per topic class) estimated by cross-validation:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{\sum_{k=1}^{C} P(\mathbf{x}|k)P(k)} \quad (4)$$

Fig. 2 shows class probabilities for the ongoing conversation. After these probabilities are computed for each class the most likely topic c is selected and the corresponding feedback is given to the users as described below.

Prompt Selection

The prompt is selected using the same technique as is used for the topic classification. For each prompt we precompute a model which keeps the word frequencies within the prompt. After the current topic of the conversation is established we find the prompt within the topic which is the most related to the recent utterances as given by the conversation model $\mathbf{x}$ (Eq. 2). We compute the relevance measure (Eq. 3) with class c being the corresponding prompt model.

Experiments

We trained the system using newsgroup and web data (an average of 150,000 words per topic) and attempted to recover the currently active topic out of the twelve candidates. As depicted, in Fig. 2, the speakers discussed three topics in the following order: 'intlcourtofjustice', 'talk.religion.misc', and 'alt.jobs'. About 100 words per topic were uttered and the system converged to the correct topics. Only the transitions caused some confusion as the speakers migrated from one subject to another (this could be reduced by varying the parameter α which was set to 0.95). If transition errors are counted, the system has an accuracy of 93%. Naturally, in steady state, the system correctly identified all 3 topics.

After the topic is detected, the most appropriate prompt is determined and shown to the users on the large screen display (see Fig. 1). The video camera is used to evaluate how "smoothly" the conversation progresses and if the users are searching for prompts. We use a detection of a full frontal view of a user as a cue that the user is requesting assistance.

Other Applications

We are considering other possible applications of the mediation paradigm. Assistance could provide relevant information (retrieve database documents, query a web search, etc.) or encouraging communication (ask questions, direct conversation flow, prevent off-topic ramblings, etc.). For example, in an INS (immigration) interview, the system could list regulations when they are relevant to the current topic at hand. Or, consider using such a system in a marriage counseling session where its primary role is to encourage further conversation between partners. This is reminiscent of the "Eliza" program without the question-answer prompts. It also does not interrupt the participants when the conversation is progressing nicely.

Conversational Style

In addition to tracking topics, we can also track the discussion's emotional content and overall conversational style. Negative and positive attitudes can be detected based on word choice using a two-class model, similarly to the method of topic classification. Alternatively, formal or casual conversation classes could be other axes that, when tracked, would measure the formality of a situation. This then determines if the meeting can be interrupted (i.e. if pager or cell phone should be allowed to ring).

Conversation Analysis

It is often convenient to revisit the content of a long meeting in a summary form to find out if all the topics of the agenda have been covered. Such a topic-based summary can be produced with our system for review purposes.

Discussion

For more flexibility, we are now developing our own speech recognition engine. Using a recurrent neural network, we compute phoneme probabilities (in a 40 phoneme alphabet) in real-time. This provides a lower level representation of audio that could handle phonetically similar words, proper nouns and foreign languages.

We are also considering other output modalities. For instance, if the conversation indicates users are frustrated (i.e. many harsh words) we could toggle a bluish or greenish lighting to create a more relaxed ambiance.

The system we described required us to manually specify the outputs and design questions it would ask to imitate a human mediator or facilitator. Ultimately, however, a mediation system should be trained from real-world data where the machine studies a human mediator responding to sample meetings and participants. The machine forms a predictive model of the mediator's responses to different key words spoken during the training meeting sessions. The output (i.e. the mediator's responses) could then be automatically associated with their trigger stimuli and these could be later synthesized by the machine (i.e. via audio playback).

We demonstrated a real-time conversational context tracking system. The topic tracking performs well reliably processing natural spoken audio. This situational awareness has many applications and we have shown examples of it as a meeting augmentation tool to prompt speakers with relevant questions. For further results see:

`http://www.media.mit.edu/vismod/demos/conversation.html`

References

Clarkson, B.; Sawhney, N.; and Pentland, A. 1998. Auditory context awareness in wearable computing. In *Workshop on Perceptual User Interfaces*.

Franklin, D.; Bradshaw, S.; and Hammond, K. 2000. Jabberwocky: You don't have to be a rocket scientist to change slides for a hydrogen combustion lecture. In *IUI*.

Jebara, T., and Pentland, A. 1997. Parametrized structure from motion for 3d adaptive feedback tracking of faces. In *Computer Vision and Pattern Recognition*.

McCallum, A., and Nigam, K. 1998. A comparison of event models for naive bayes text classification. In *AAAI / ICML Workshop on Learning for Text Classification.*

Rhodes, B., and Starner, T. 1996. The remembrance agent: A continuously running automated information retrieval system. In *Practical Application of Intelligent Agents and Multi Agent Technology*.