Multi-modal Discourse Investigation with SIDGrid

Principal Investigators: Gina-Anne Levow, Susan Duncan, and David McNeill

1 Introduction

Human communicative discourse is vastly more than simply a sequence of utterances strung together. Rather, it depends on the creation and maintenance of coherence between utterances through a range of devices from lexical chains to discourse markers to referring expressions. Furthermore, spoken and multi-party discourse depend not only on lexical evidence, but also on a range of prosodic cues, such as pitch and loudness, and an ensemble of non-verbal communicative cues such as gesture, gaze, posture, and head nod. Thus, discourse understanding and interpretation rely on a rich, varied array of linguistic, non-verbal and paralinguistic evidence.

Study of discourse ranges from basic research into the phenomena that establish and maintain structure and coherence to computational approaches that automatically recognize and interpret these cues and even synthesize appropriate behavior. As a result, the study of discourse is strongly interdisciplinary, drawing on research in linguistics, psychology, sociology, anthropology, and philosophy as well as computational techniques for signal and language processing and machine learning.

2 Supporting Multi-modal Discourse Study

The current ATI proposal centers on the development of a laboratory and project component for courses in discourse and dialogue. Such a course (without the laboratory component elaborated here) is currently offered in the Computer Science department and will be cross-listed with linguistics and psychology. The course with our proposed elaborated lab component introduces the fundamental theoretical bases of discourse structure and coherence from many perspectives and, through the development of machine learning and computational models, creates systems to recognize and interpret discourse. This course, and other related courses, will be enhanced by the opportunity to perform flexible hands-on descriptive annotations of a standard collection of multiparty conversational discourse, to exploit tools that automate cue and feature extraction, and to perform computational analyses of these phenomena. These exercises support both descriptive analyses and creation of classification algorithms to detect activities of interest such as topic changes, turn boundaries, and discourse cohesion. The materials for these studies include multi-media multimodal data collections. Specifically the planned corpora comprise video and audio recordings of multiparty meetings, such as those recorded by the Speech Group at the National Institute of Standards (NIST) or the International Computer Science Institute (ICSI) at University of California, Berkeley. In addition to the media themselves, manual transcriptions and alignments of transcripts to audio and video are also available. In some cases, additional annotations of phenomena such as dialogue acts and topic segments may also be available. Interactive laboratory sections can be held in the Computer Science Instructional Laboratories.

3 SIDGrid for Multi-modal Study of Discourse

However, these rich data sources of multiple media files, transcripts of multiple speakers, and other annotations pose challenges for novice and even expert researchers to visualize, annotate, and analyze. Students come to this area from a variety of backgrounds including linguistics, psychology, and computer science, and must become familiar with issues in other disciplines. Importantly, in addition to the general concepts and approaches introduced in lectures and readings, students can only gain a true understanding of the challenges of working with this type of data and the phenomena in play by actually manipulating and gaining hands-on experience with real data, annotations, and analysis and classification techniques.

To overcome these challenges, we plan to exploit the infrastructure for annotating, archiving, and analyzing multi-modal, multi-measure time series data under development as part of the NSF-funded project, "Cyberinfrastructure for Collaborative Research in the Social and Behavioral Sciences", conducted jointly by the University of Chicago, Argonne, and University of Illinois at Chicago. This infrastructure, known as the Social Informatics Data Grid or "SIDGrid"(http://sidgrid.ci.uchicago.edu), provides a novel, smooth integration of annotation and analysis as well as an opportunity to employ the computational resources of the TeraGrid(http://www.teragrid.org) for data analysis. The SIDGrid infrastructure comprises three main components: a client-side annotation and analysis interface, a data repository that is browsable and searchable through a web interface, and a web-based portal to the TeraGrid for distributed analysis.

3.1 SIDGrid Infrastructure

The client-side interface builds on an existing open-source annotation tool, Elan[Wittenburg et al.2006], developed by the Max Planck Institute (http://www.mpi.nl/tools/elan.html). Elan supports synchronous playback of multiple video and audio files as well as entry and playback of annotations to these media files. These annotations, such as speech transcriptions or labeled changes of gaze or turn, are time-aligned and displayed in sync with the media in a "music-score"-style interface. They may be browsed and edited interactively, as shown in Figure 1. Under SIDGrid, the tool has been extended to support display of time-aligned time series data such as pitch tracks or motion capture data, interactive execution of user-defined programs on linked media and annotation files, and upload and download to the data repository. The on-line repository allows multi-user distributed access to stored data and annotations through a web-based interface for browsing and simple meta-data search. The repository's internal database representation is designed to facilitate more sophisticated search. Finally, the system enables users to apply user-defined functions to data in the repository using the large-scale distributed computing capabilities of the TeraGrid, a large open science infrastructure. The TeraGrid uses high-speed network connections to link high performance computers and large scale data stores distributed across the United States.[Pennington2002] Software controls the distribution of programs and data to these machines to execute in parallel and also collects the output results. These capabilities enable the user to perform more computationally expensive processing on the complex multimedia data, such as signal processing for speech and video analysis. This new infrastructure was designed explicitly to support data such as that required for studies of discourse and human communication. Use of this technology within our Discourse Analysis course will exploit a synergy between the NSF-funded cyberinfrastructure project and the needs of students studying complex multi-modal behaviors.

Each of the SIDGrid components supports exercises and projects in discourse analysis, from

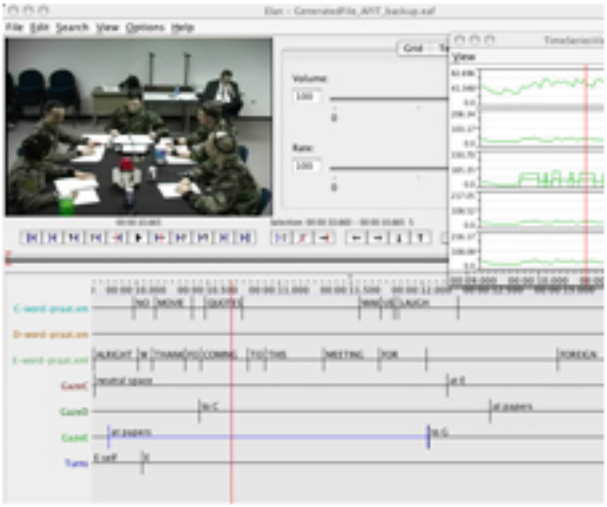

Figure 1: SIDGRID-enhanced annotation and analysis client

theoretical and computational perspectives.

3.1.1 Interactive Data Exploration

Analysis builds on manual and semi-automatic annotation of the data. For example, pitch, loudness, duration, gesture, gaze, and posture all are theorized to participate in floor control to manage turn-taking. To explore how these interact, a combination of measures drawn from signal processing such as pitch track as well as discrete labels of gaze change or gesture initiation are required. The use of the SIDGrid client interface will support exercises in annotation, analysis, and visualization. Much current research on natural discourse and human interaction depends on annotations stored in flat format text or richer XML files or spreadsheets. These formats make visualization of time synchronous phenomena difficult. Even the few tools which support multi-modal or multi-speaker displays have only limited support for concurrent visualization and extraction of measures such as pitch. The SIDGrid client supports all these capabilities and will allow exercises in annotation of prosody to identify events in the ToBI (Tones and Break Indexes) [Silverman et al.1992] phonological framework and also annotation of non-verbal phenomena such as gesture, posture shifts, and gaze. Instruction in annotation and analysis of the latter will be led by researchers from the McNeill lab, with expertise in non-verbal micro-analysis of conversational data.

3.1.2 Distributed Project Support

A core component of these courses is a small group project, often combining team members from different areas of expertise. The SIDGrid repository will greatly facilitate these projects by providing a shared repository for annotations and analysis that can be accessed by groups members and which record the history of incremental changes to support distributed analysis. Furthermore, it will enable the students to escape the limits of relatively restricted home directory space for large multimedia data sets.

3.1.3 Distributed Grid-based Data Analysis

Finally, the use of the TeraGrid will expose students to the large-scale distributed computational resources of the Grid as well as the tools for setting up and running experiments on these resources. Few students who are not active researchers in Grid technology themselves have exposure to these tools, and those who do are concentrated in Physics and Astronomy. This access will both support more computationally demanding analysis, such as signal processing for audio and video, and also give students experience in the use of general Grid-based tools.

4 Courseware Development for SIDGrid Integration

To support these activities to aid students in developing intuitions about discourse and and an understanding of the capabilities and limitations of current techniques, a variety of courseware development is required First, streamlined student-level installation scripts and tutorial documentation will be developed to enhance ease of use of the SIDGrid interface and utilities. Then the data, media resources, and annotations for several standard data sets must be acquired and converted to the SIDGrid format for storage in the repository and access through the annotation client. We plan to develop support for laboratory activities on topics including, but not restricted to the following:

- Annotation and recognition of coreference in multiple modalities,
- Identification of multi-modal correlates of discourse structure, cohesion, and topic change in text, intonation, and non-verbal communication for analysis and classification, and
- Modeling of multi-modal signals of floor control and turn-taking.

Analysis scripts in support of signal processing and data analysis will be developed for each of the desired prosodic measures and multi-measure analysis, employed for the study of the topics above. Analogous scripts and underlying software must also be installed on the TeraGrid to support distributed processing.

Finally, some novel extensions to the SIDGrid search capability will also be performed to provide additional data mining capabilities. Current SIDGrid functionality supports the application of user-defined functions to media, such as audio and video files, to extract timeseries such as pitch tracks. Enhancements to these capabilities include search and data aggregation across both media data and annotations. For example, techniques for topic change detection require computation of text overlap or text similarity across time spans and can be augmented with prosodic evidence. As another example, analysis of turn-taking will be facilitated by extraction of average pitch in words at the beginning and end of hypothesized turns, requiring the use of word and turn annotations as constraints on pitch extraction processing. We also consider aggregation across multiple interactions, integrating additional software packages, such as *Theme*, developed by Magnus Magnusson[Magnusson1996], that support such analysis, and will require additional data format conversion support. For instance, analysis of group interaction dynamics requires detection of small and large-scale patterns among tagged events in multiple co-occurring behavioral streams, as well as temporal, sequential dependencies among multi-modal behaviors, across participating individuals.

This software and the related exercises will be tested by a student for ease of use and clarity of directions. This structure is readily extensible to new analyses and new annotation types as data becomes available. The augmentation of courses in Discourse Analysis with tools and software for

annotation, archiving, and analysis from the SIDGrid infrastructure will enhance the opportunities for students to explore these rich multi-modal phenomena integral to human communication.

References

[Magnusson1996] M. Magnusson. 1996. T-patterns, theme and the observer. In *Proceedings of Measuring Behavior '96, International Workshop on Methods and Techniques in Behavioral Research.*

[Pennington2002] Rob Pennington. 2002. Terascale clusters and the TeraGrid. In *Proceedings for HPC Asia*, pages 407–413. Invited talk.

[Silverman et al.1992] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labelling English prosody. In *Proceedings of ICSLP*, pages 867–870.

[Wittenburg et al.2006] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2006.*