Improving Team's Consistency of Understanding in Meetings: Intelligent Agent Participation and Human Subject Studies

by

Joseph Kim

B.S., University of Illinois at Urbana-Champaign (2012)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Signature Redacted

Author .....
Department of Aeronautics and Astronautics
Signature Redacted , 2015

Certified by .....
Julie A. Shah
Assistant Professor of Aeronautics and Astronautics
Thesis Supervisor

Signature Redacted

Accepted by .....
Paulo C. Lozano
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Improving Team's Consistency of Understanding in Meetings: Intelligent Agent Participation and Human Subject Studies

by

Joseph Kim

Submitted to the Department of Aeronautics and Astronautics
on May 21, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Aeronautics and Astronautics

Abstract

Upon concluding a meeting, participants can occasionally leave with different understandings of what had been discussed. For meetings that result in immediate subsequent action, such as emergency response planning, all participants must share a common understanding of the decisions reached by the team in order to ensure successful execution of their mission. Thus, detecting inconsistencies in understanding among meeting participants is a desired capability for an intelligent system designed to monitor meetings and provide feedback to spur stronger shared understanding within a group.

In this thesis, we present a computational model for the automatic prediction of consistency among team members' understanding of their group's decisions. The model utilizes dialogue features focused on capturing the dynamics of group decision-making. We trained our model using one of the largest publicly available meeting datasets and achieved a prediction accuracy rate of 64.2%, as well as robustness across different meeting phases. To the best of our knowledge, our work is the first to automatically predict levels of shared understanding using natural dialogue.

We then implemented our model in an intelligent system that participated in human team planning meetings about a hypothetical emergency response mission. The system suggested discussion topics that the team would derive the most benefit from reviewing with one another. Through human subject experiments with 30 participants, we evaluated the utility of such a feedback system, and observed a statistically significant mean increase of 17.5% in objective measures of the consistency of the teams' understanding compared with that obtained using a baseline interactive system.

Thesis Supervisor: Julie A. Shah
Title: Assistant Professor of Aeronautics and Astronautics

Acknowledgments

Personal Acknowledgments

First and foremost, I thank Jesus Christ, my Lord and Savior for this thesis. I pray that this work would serve to glorify Him, and that throughout the rest of my graduate years, I can be His good and faithful witness.

I would like to express my sincere gratitude to everyone who had made my last three years at MIT one of the most enriching periods of my life. Not only have I gained knowledge, I have grown significantly as a researcher and as an individual. I would first like to thank my advisor, Professor Julie Shah. Her guidance and support has been tremendously helpful. I am grateful to have an advisor who is not only brilliant, but possesses a genuine understanding and a desire to help me become the best that I can be. Her enthusiastic and encouraging attitude has kept me motivated, and helped me strengthen my weaknesses and accomplish my goals. I am extremely grateful and look forward to her continued guidance over the course of my Ph.D.

Next, I would like to thank all the members of the Interactive Robotics Group. They have made my life as a graduate student very enjoyable. Through all the lab lunches, dinners, game nights, and social events, I have shared many smiles and laughs with them. Also, they are the best peer reviewers and counsellors. I have learned so much and gained wisdom from both the senior and junior members of our group.

I would not be where I am here today without the unconditional love and support from my family. I would like to thank my parents for everything. From them, I have learned the principles of hard work and integrity. Life as immigrants have been difficult for them, but one day, I believe we will all reach our dreams. I would also like to thank my little brother, Joshua, for his love and maturity in taking care of the family matters even at his young age. I would also like to thank my grandfather for always praying for our family.

I would like to thank the fellow brothers and sisters whom I have developed relationships with at the First Korean Church in Cambridge (FKCC). I would like to thank the pastors, the church leaders, and the small group leaders. Thank you all for

your kindness and your serving hearts even when you did not expect much in return. I look forward to giving back to the community as much as I can.

I would like to thank my friend, Przemyslaw (Pem) Lasota for his rapport and trustworthiness. He is also the best roommate and a labmate one could ask for. Throughout the next several years, we will support each other and one day graduate with our Ph.Ds.

Lastly, I would like to thank my girlfriend, Hyehee, for her amazing loving heart. I am grateful to be in a relationship with someone so compassionate. Thank you for being in my life.

Funding

This thesis was supported by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) under grant no. 2388357.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Meetings are an integral component of many collaborative and organized work environments [46]. Each day, over 11 million meetings take place in the United States, and over 2.6 billion occur each year [2]. Meetings are essential and the number of meetings and their duration has been steadily increasing [48], [46]. Managers spend between a quarter and three-quarters of their time in meetings [41], and approximately 97% of workers have reported in a large-scale study [28] that collaboration is essential to do their best work. However, we realize that meetings are often not as efficient as they could be: An estimated $54 million to $3.4 billion is lost annually as a result of inefficient meetings (e.g., getting off-topic, poor preparation, a lack of organization, misunderstandings among participants, etc.) [51]. Consequently, there is a great interest in improving meeting productivity and efficiency.

One common source of inefficiency is inconsistency between team members in their understanding of the outcome of a meeting [51], potentially causing miscommunication and confusion. Figure 1-3 provides a visual illustration of inconsistency between team members' understandings. In highly dynamic crisis domains, where degradation in team performance can lead to high public safety costs, such mishaps can have severe consequences [17]. Therefore, it is imperative in such situations that team members reach a uniform understanding of the decisions made by the group.

Figure 1-1: Pictures of typical meetings in office environments. They can be physical meetings or meetings conducted through an electronic medium such as a video conferencing. Images are from [1] and [12] respectively.

Figure 1-2: Meetings in safety-critical domains. Image on the left hand side shows firefighters using a Web-based situational awareness tool (NICS [18]) to coordinate missions for emergency response. Image on the right hand side shows a crew operating inside an E-2D Advanced Hawkeye, coordinating distributed aerial fleets. Images are from [13] and [31] respectively.

We are interested in developing an intelligent system that would monitor meetings and provide useful feedback to help team members to remain 'on the same page.' The system would suggest a review of the discussion topics with the greatest potential to result in inconsistent understanding among team members, and provide friendly reminders to review those topics before adjourning the meeting (Figure 1-4 illustrates this idea). A system with this capability could serve to reduce misunderstandings and hidden conflicts among meeting participants that could have otherwise gone unnoticed. This has potential to make everyday meetings more efficient.

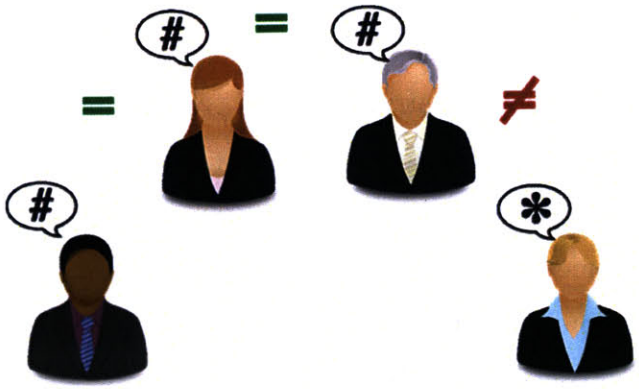

Figure 1-3: A visual illustration of the inconsistency between team members' understandings. The team member on the rightmost side possesses an understanding that is conceptually different from that of the other members.

1.2 Thesis Contribution

In prior literature [49], [58] researchers have proposed qualitative models to explain the process of how teams reach a consistent, or shared, understanding of one another. Although much has been written about the modeling of a shared understanding process, prior work has been purely qualitative i.e., constructing theoretical models inspired by results from observational studies. We build on this prior work by enabling a computational framework such that the level of shared understanding among team members can be quantitatively assessed by a computer.

In this thesis, we present a computational model to predict the consistency among team members' understanding of their group decisions (defined as **consistency of understanding**). Our work enables an automatic framework for assessing a level of shared understanding, a form of shared cognition that has previously only been analyzed qualitatively. To do this, we utilize a set of dialogue features that focuses on capturing the dynamics of group decision-making and incorporate them as features of a machine learning algorithm.

One of the key benefits of our model is generalizability: The model learns by monitoring the dynamics of how teams plan, not what they are planning for — in essence, it does not rely on any domain-specific content. We trained our model using one of the largest publicly available meeting datasets, and achieved a mean prediction

accuracy rate of 64.2%, with robust performance across different meeting phases.

Next, we demonstrate the utility of our computational model when it is implemented for an intelligent agent participating in live meetings. This agent monitors team dialogue over the course of a meeting, and suggests that the participants review discussion topics that the model has predicted will result in inconsistencies. Through human subject experiments involving 30 participants, we evaluated the utility of such a feedback system and observed a statistically significant mean increase of 17.5% in objective measures of consistency of understanding.

Overall, our multi-step study makes the following contributions to the literature: (1) We demonstrate that a computer can automatically assess the consistency of understanding within a team through natural dialogue, with a prediction accuracy rate above random chance. In other words, we show that there is a predictive signal in the monitoring of team planning dynamics through dialogue features proposed from qualitative studies. (2) We contribute to the understanding of how an intelligent agent could participate in human meetings. To our knowledge, no prior studies have explored how shared understanding within a team is affected by receiving a review recommendation from a computer.

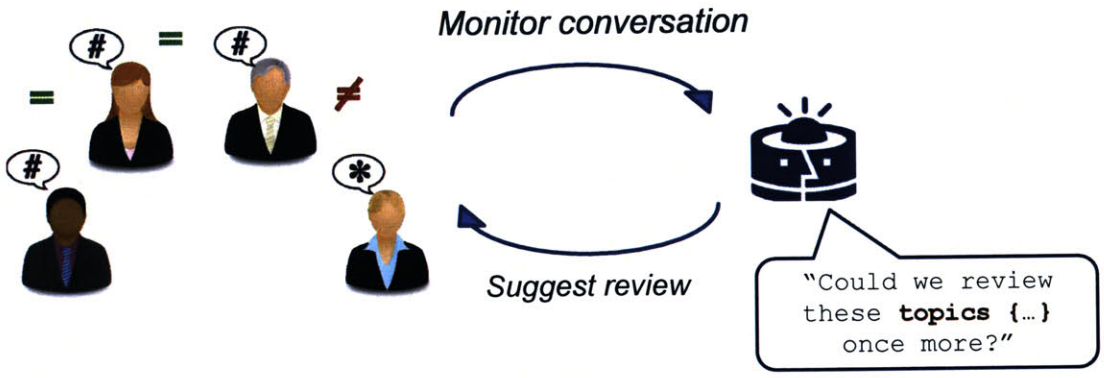

Figure 1-4: An illustration of intelligent system helping teams reach consistent understanding of each other.

Chapter 2

Related Work

2.1 Shared Understanding

Assessing levels of shared understanding through natural dialogue is a challenging task. Human dialogue is complex: Discussions unfold in cycles, agreements are fluid and idea proposals are often communicated and accepted implicitly [19]. Shared understanding represents an alignment of mental states, and therefore presents difficulties for explicitly monitoring its continually evolving process [49].

Despite these challenges, shared understanding has been a topic of multidisciplinary research in the linguistics, cognitive psychology and social science communities. Definitions of shared understanding include "the overlap of understanding and concepts among group members [49]," "the ability to coordinate behaviors toward common goals or objectives [52]" and "having mutual knowledge, mutual beliefs and mutual assumptions (content and structure) on the task [11]." Here, the idea of "sharedness" refers to the commonality of understandings among team members, and not "shared" in the sense of the division of resources. Our definition of "consistency of understanding" is synonymous with previously mentioned definitions, but it provides a clear emphasis on the overlap and alignment of understandings. According to prior work, shared understanding has positive effects on production performance (with regard to both quality and quantity of products) [43], individual satisfaction [38], reduction of iterative loops and re-work [36], innovation [37] and team morale

[15].

The process of how shared understanding is achieved has been investigated previously. Mulder et al. [49] described this process as a three-step transition from an initial phase of conceptual learning (primary exchange, reflection and refinement of facts and concepts), to a feedback phase (confirmations, checks and explanations among group members), and finally to a motivation phase (evaluative expressions of usefulness, certainty and uncertainty). Bossche et al. [58] identified a set of team learning behaviors and explained that collaborative groups express and listen to individual understandings (construction), discuss and clarify them to reach mutual understanding (co-construction) and negotiate an agreement upon a mutually shared perspective (constructive conflict). Eugenio et al. [19] described the process as a three-phase transition between balance, propose and dispose stages, and also highlighted the importance of tracking commitment dynamics across team members.

In short, shared understanding is considered crucial to the quality of group interactions, and much has been written about the essence of shared understanding and the process through which is reached. However, prior work has been purely qualitative, focused on theoretical definitions and modeling motivated by results from observational studies. To the best of our knowledge, the study of monitoring and assessing shared understanding has not yet been generalized to an automatic, predictive framework. In this thesis, we present a computational model to predict the levels of shared understanding, such that a computer would be able to provide quantitative measures.

2.2 NLP: Agreement Detection

Prior work within the natural language processing (NLP) community has explored the related task of automatically detecting "agreements" in meetings [30], [22], [24]. This task involves the detection and classification of agreements as positive or negative through machine learning algorithms incorporating verbal and nonverbal dialogue features.

The detection task is generally performed for each spoken utterance. For example: "Yes, that sounds great" would be classified as a positive agreement, while "I don't like that idea" would be classified as a negative agreement or disagreement. Non-agreement utterances would be classified as neutral. However, this work only captures agreements during single instances, and from a single speaker's perspective; they do not capture the essence of "joint agreement," which is more closely related to the definition of shared understanding.

Speaker	Example Discussion
A	What is our priority?
B	Um, how about we proceed with lower region first?
A	**Okay, I like that idea.** ← {Positive agreement}
	....
C	**I don't think so.** ← {Negative agreement}

Figure 2-1: Example annotation of agreement detection task.

While we believe that momentary agreement is an important feature that may lead to an eventual shared understanding within a group, these two terms are not interchangeable. "Agreement" refers to an accordance with another's opinion at a spoken utterance, while "shared understanding" refers to a state of group consensus resulting from the culmination of an entire discussion. For example, a meeting participant can disagree with another participant during a given moment in a discussion, but may still possess a clear understanding of what the group has decided on upon completion of the meeting. In contrast to the related work in the NLP field, our work focuses on utilizing the full discussion to predict the level of shared understanding within the group.

2.3 Intelligent Agent Participation

Intelligent agents are increasingly being integrated into tasks such as automatic summarization [61], speaker identification [21], plan extraction, detection of meeting ac-

tions [45], modeling of social interactions [23] and audiovisual processing of various cognitive states for analysis [47]. In the case of the latter, researchers are developing models to infer participants' states of concentration, interest, confusion and frustration [20], [35], and have used intelligent agents to predict the outcomes of interviews [50] and the success of negotiations [42]. More recently, we have seen physical intelligent agents (robots) integrated into meetings for example, to serve as moderators in balancing engagement and dominance levels [57] and in predicting levels of interpersonal trust among team members [39]. Our work addresses the novel task of automatically predicting the consistency of understanding during team meetings. This problem is unique, in that it involves prediction of a shared cognitive state.

Figure 2-2: Examples of intelligent agent participation. On the left hand side, a Wakamaru robot is monitoring the human's engagement. The right image shows a Nexi robot interacting with a human and monitoring the levels of interpersonal trust. Images are from [55] and [39] respectively.

Chapter 3

Problem Statement and Approach

The problem statement of our work is to automatically predict the consistency of understanding using a team's natural dialogue, which can be supplied either online or offline. The focus is on learning through textual data; however, we also investigate the potential benefits of incorporating nonverbal features, such as head gestures, into the model.

In our problem, we assume a structural form that meetings are composed of discussions of several topics. These topics can be envisioned as a list of items on a meeting agenda, where **topic discussions** form collections of dialogue relevant to decision-making for individual topics. In our problem, we perform a single prediction task for each topic discussed. We believe this is an important level of granularity for the development of a system that can identify (in)consistent topics over the course of a meeting, rather than outputting a measure for the meeting as a whole.

3.1 System Flowchart

Figure 3-1 depicts the flowchart of our problem statement. A topic of discussion is read as input by the computation model, which then outputs a prediction about the consistency of understanding within the group for that topic. The output is binary i.e., team members can have either a consistent or inconsistent understanding of group decisions. We leave explorations of additional levels (such as a moderate level

of consistency) for future work. Consistency of understanding, especially information on its ground truth labeling, is described in Section 4.2.2. Finally, for topics that the model predicts will result in inconsistency, a system feedback is triggered suggesting that team members review those topics. To avoid potential confusion, we would like to emphasize that the term 'inconsistent' here refers to inconsistencies or misalignment among team members' understandings. We are not using the term to represent the quantitative infeasibility of a plan structure (which is a term frequently utilized in AI).

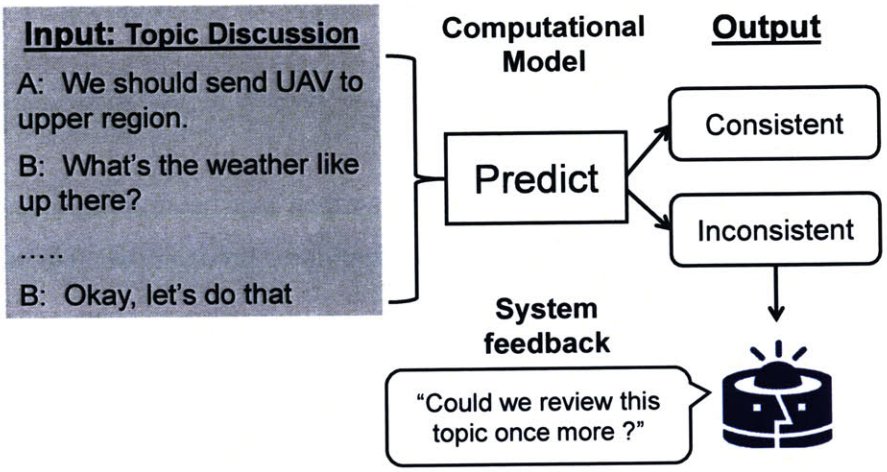

Figure 3-1: Flowchart of the problem statement. The input is a topic discussion, and the output is consistency of understanding. System feedback is triggered for topics predicted to be inconsistent.

3.2 Approach

One of the biggest challenges for our problem statement is the mapping of natural dialogue to a concrete set of features that can capture information about a team's consistency of understanding. In order to accomplish this, we adopt the idea of tracking the conversational dynamics of group decision-making. In essence, we aim to capture the process of how a team plans, which is considered to be an important feature in modeling group consensus [29], commitment [26] and agreement [19]. With regard to the aforementioned cognitive states, we assume that consistency of understanding is

related to shared cognition, and thus utilize the set of features proposed from prior studies for our computational model.

We use a particular set of features defined from Eugenio et al. [19] (referred to here as *Eugenio's features*), which has been shown to monitor the evolving attitude of participants' commitment toward options presented during a meeting.[1] It also describes how joint commitments are achieved by the group. Eugenio's features are types of *dialogue acts* [54], [34], or labels that define the functional role of utterances. Table 3.1 provides a list of dialogue acts, their definitions and example sentences (showing only subset of a full list from [8]). However, in contrast to conventional dialogue acts, Eugenio's features have been shown to facilitate the recognition of implicit and/or passive acceptance of options by team members. These characteristics make Eugenio's features useful for predicting consistency of understanding, as joint commitment toward options would naturally lead to joint understanding of group decisions. We describe how Eugenio's features are generated from conventional dialogue acts in Section 4.2.3.

Table 3.1: Sample Dialogue Act Labels [8]

Label	Definition	Example sentence
Inform	Exchange of information	*UAV is located here.*
Assess	Comment expressing an evaluation	*That is a good plan.*
Suggest	Expression of intention to the actions of another individual, or a group as a whole	*Let's send it over there.*
Offer	Expression of intention relating to own actions	*I can do that task.*
Elicit-Inform	Requesting of an information	*What kind of system is it?*
Elicit-Assess	Attempt to elicit an assessment	*What do you think about this?*
Understanding	Comment on understanding about a discussion point	*Yes I see.*
Elicit-Understanding	Asking for a comment about understanding	*Do you see what I mean?*

We learn a model for consistency of understanding from sequences of Eugenio's features in dialogue. Maintaining sequential information is of particular importance,

[1] 'Options' here refers to proposed ideas and choices to be decided on by the group [19].

because natural turn-taking behavior exists within human dialogue (often called *adjacency pairs*, e.g. proceeding from question to answer, request to acceptance or rejection, etc.). The order in which one dialogue act follows another may provide discriminative information for distinguishing a team's shared understanding: For example, a sequence of "question → question → question" may be a pattern of weaker understanding than a sequence of "question → acceptance → confirm." Our approach is to apply machine learning algorithms to learn discriminative sequences of Eugenio's features. We perform machine learning to derive patterns from real human dialogue, rather than specifying any hardcoded templates. We represent a topic discussion using a sequence of dialogue acts, as follows:

$$\boldsymbol{D} =< DA_1, DA_2, ..., DA_L >, \ where \ DA_i \in \boldsymbol{\Lambda}$$

where $\boldsymbol{D}$ is a topic discussion, DA_i is a dialogue act realized at instance i (which designates a row on a discussion table), subscript L is the length of the topic discussion and $\boldsymbol{\Lambda}$ is a finite set of dialogue acts. For $\boldsymbol{\Lambda}$, our primary feature set incorporates Eugenio's features.

There are several advantages to using dialogue acts to represent a discussion: First, dialogue acts allow for the learning of conversational dynamics without the extraction of keywords or domain-specific content, in turn allowing for generalizability of both qualitative and quantitative models across different topic discussions. The resulting sequence essentially stores information about how teams plan, and does not require the processing of potentially sensitive information. Second, dialogue acts offer a higher level of abstraction than working directly at the word level (a common approach for NLP-related tasks such as topic modeling and document classification). By representing a discussion as a sequence of labels drawn from a finite set, the computational complexity for learning algorithms is significantly reduced.

Also, we investigate the benefits of multimodal fusion by including head gestures as an extended feature set. Head gestures have been used previously to infer a state of agreement, disagreement, concentration, interest or confusion [20]. We test whether

the combination of head gestures with textual features improves the prediction performance of the model.

Chapter 4

Dataset

4.1 AMI Meeting Corpus

The dataset we used to build and train our model comes from the AMI meeting corpus [8]. It is one of the largest publicly available meeting corpora, containing over 100 hours of recordings. In each of these meetings, a team of four people collaborated on a task related to product design. The meetings were divided into four distinct phases of the design process (descriptions are provided in Table 5.1) and were scenario-driven. Each participant served one of four specific roles: project manager, industrial designer, marketing expert or user interface designer. Although the participants were engaged in role-playing, they were guided by personal coaches with regard to how best perform their role and, most importantly, their conversations were collaborative and reflected natural, human-to-human interaction. The length of the meetings ranged from approximately 10 minutes to 45 minutes, which overlaps with the most common lengths of meeting, as discussed by [51].

The AMI meeting corpus is well-suited for our study, because the conversations that occurred during these meetings were tailored toward a group decision-making process. The use of Eugenio's features is also appropriate due to the collaborative environment of the meetings, wherein all decision points were consensual. This makes consistency of understanding an important outcome from these meetings.

Line	Speaker ID	Topic Discussion = "Remote locator"	DAs: Conventional	DAs: Eugenio's	Head gestures
		.....			
1	B	Do we incorporate the idea of trying to locate the remote control again via a beeping noise?	Elicit-Assess	PDO	
2	D	Yeah, think so.	Assess		D: Concord
3	C	Um, I think so, because it's so small	Inform		
4	C	I mean if we only have like two, three buttons it might be essential to have to have that [pause]	Assess		B: Concord
5	B	The ability to locate it again.	Elicit-Inform		
6	C	Yeah.	Inform		B: Concord
7					A: Concord
8	B	That would require a transmitter maybe attached to the TV and a basically small microphone	Inform	UO	
		....			
9	B	If you could look into what we've suggested so far, the feasibility of small transmitter, and ...	Suggest	Proposal	
10	C	Okay. Sure.	Assess	Commit	C: Concord

Figure 4-1: A sample conversation segment taken from the AMI corpus. Here, the participants are discussing a topic related to a remote locator device. Corresponding annotation layers of dialogue acts, Eugenio's features and head gestures are provided in the right-hand columns.

The corpus also contains a rich collection of annotations[1]. In our study, we utilized annotations of topic segmentations, participant summaries, dialogue acts and head gestures. Here, we describe how each annotation layer was used to construct the components necessary to build our computational model.

4.2 Generation of Input Components

4.2.1 From topic segmentations to topic discussions

Topic segmentations partition each meeting according to related topics. They naturally represent our definition of a topic discussion by providing conversation segments that focus on decision-making about a single topic. Some examples of topics from the AMI corpus include: physical appearance, target audience, product customizability, etc.

[1]for a full list of available annotations, we refer readers to [8]

4.2.2 From participant summaries to consistency of understanding

Self-reported participant summaries were used to establish ground truth on consistency of understanding. At the end of each meeting phase, participants were asked to provide written summaries of all the decisions made by the group. We compared the summaries and checked for alignment. If all summaries were aligned, the associated topic discussion was labeled *consistent*; the discussion was identified as *inconsistent* if one or more of the summaries differed in content. Note that this is a "hard" measure of consistency; i.e., if even one individual's summary differed from the others (in a group size of n members), a ground truth of *inconsistent* was applied. For groups of larger sizes, alternate and more conservative methods of labeling consistency can be explored. In the AMI dataset, there were four participants per meeting.

Two annotators performed the comparison of consistency (inter-rater agreement, $\kappa = 0.73$), resulting in ground truth labels for a total of 140 topic discussions. There was an imbalance in the distribution: 93 discussions were identified as *consistent* and 47 discussions were *inconsistent.*

Prior work has utilized an identical approach for comparing participant summaries to form ground truth on shared understanding [3], [4]. Other measurement alternatives include structured interviews and Likert scale questionnaires about perceived shared understanding [60], [49]. However, an individual's perception of the shared understanding within a group may be susceptible to confirmation biases. Therefore, we believe that comparing individual plan summaries provides a more objective measure of shared understanding.

4.2.3 From dialogue acts to Eugenio's features

The AMI dataset provides annotations of conventional dialogue acts (DAs), but not Eugenio's features. However, conventional DAs can be used to construct Eugenio's features given the knowledge of "solution sizes." A solution size is defined as "determinate" when sufficient relevant information has been exchanged between meeting

participants to form options. "Indeterminate" refers to instances wherein further balancing of information is required. Solution sizes are the distinguishing element of Eugenio's features. They represent the state of the discussion on whether or not the participants would need more information before deliberation. Because we did not want to rely on any domain-specific materials, we had to approximate the state from the dialogue acts. We applied the heuristic of marking a portion of a topic discussion as "indeterminate" until the final DA label of "Inform" is displayed, at which point the conversation segment is marked as "determinate." Note that this effectively requires the whole topic discussion to be seen by the system, before solution sizes can be determined.

With DAs and solution sizes, we applied the coding scheme described in [19] to construct Eugenio's features. Table 4.1 provides an overview of Eugenio's features, including their descriptions and coding schemes. Figure 4-1 depicts a sample conversation segment with a full layer of annotations, including Eugenio's features. Note that "action-directives (AD)" correspond with suggestions and all elicit forms of DAs that require actions from partners.

Table 4.1: Eugenio's features, descriptions and coding schemes

Feature	Description	Coding
Partner decidable option (PDO)	Occurs when a speaker offers an option that partners can use during decision-making. Corresponds to options that require further deliberation and balancing of information within the group.	AD, offer + indeterminate
Proposal	Occurs when a speaker offers an option following its full deliberation by the group.	AD, offer + determinate
Commit	Occurs when a speaker indicates commitment to an option after full deliberation.	Offer, assessment (positive) + determinate
Unendorsed option (UO)	Occurs when an option is simply presented during deliberation, without the speaker expecting any corresponding action from other group members.	Open-options + determinate

4.2.4 Head gestures

The AMI corpus provides annotations of head gestures that reflect one's intent rather than simple form. (For example, a nod of the head is further evaluated in order to distinguish between signals of comprehension, emphasis, etc.) We incorporated gestures intended to communicate understanding and comprehension between participants. Table 4.2 highlights the description of head gestures used in our study. Figure 4-1 also depicts the head gestures made during the conversation segment.

Table 4.2: Description of head gestures used in our study

Head gesture	Description
Concord	Signals comprehension, agreement or positive response; often characterized by a head nod.
Discord	Signals comprehension failure, uncertainty or disagreement; often characterized by a head shake or tilt.
Negative	Signals negative response to a yes/no question; usually characterized by a head shake.
Emphasis	Signals effort to accentuate or highlight a particular word or phrase, often characterized by a nod or head bob.

4.2.5 Processed Data

The processed data from the AMI corpus were reduced to 140 topic discussions with labeled consistency of understanding. Each topic discussion is represented as a sequence of DAs. These sequences were of varying length. In the following section, we describe how we utilized the training data to predict consistency of understanding for a test discussion, $\boldsymbol{D_{test}}$.

Training Data:

$\boldsymbol{D_1} =< DA_1, DA_2, ..., DA_L >$	$y_1 = consistent$
$\boldsymbol{D_2} =< DA_1, DA_2, ... >$	$y_2 = consistent$
$\boldsymbol{D_3} =< DA_1, DA_2, ... >$	$y_3 = inconsistent$
...	...

Chapter 5

Computational Model

5.1 HMM Formulation

We designed a computational model to evaluate consistency of understanding using the proposed feature sets. We modeled the problem using hidden Markov models (HMMs) because of their applicability to modeling systems with temporal sequences, as well as for their prior success within the human communication and social interaction domains [44], [39], [59].

An HMM is defined as a 5-tuple $\{S, O, A, B, \pi\}$, where:

- S is the finite set of hidden states, and $m = |S|$ is its cardinality. One potential interpretation of the hidden states is that they serve as representations of different shared understanding processes defined from qualitative literature [49], [58], [19]. For example, they may represent Bossche et al's definitions, wherein the group may be going through a state of construction or co-construction or a constructive conflict during a specific moment of a discussion. A precise, interpretable definition of S is unknown, but only m is required to train and test an HMM. m controls the number of underlying discussion states and serves as a meta-parameter for the prediction model.

- O is the finite set of observations. An observation at each time step is a dialogue

act realized from the speaker's utterance. $|O|$ represents the number of unique observations (i.e., the number of features). The primary O we use consists of Eugenio's features, presented in Table 4.1. However, we also test cases in which O includes conventional DAs, head gestures or combinations of the two, in order to build baseline HMMs to compare performance across different feature sets.

- A is the state transition matrix, with a size of m by m, and describes the probability distribution of transitioning from one discussion state to another. The Markov assumption is generally accepted due to the frequent occurrences of adjacency pairs in human dialogue [14], [6], [5].

- B is the observation probability matrix. It describes the emission probability of an observation (dialogue act) conditioned on a hidden discussion state. With a combination of A and B, the stochastic process of O is fully described.

- π is the initial hidden state distribution.

In order to train HMMs, the distributions of A, B, and π are iteratively learned through an expectation-maximization algorithm known as the Baum-Welch algorithm [16] using the processed training data. Two separate HMMs are learned for prediction — one for consistent class and one for inconsistent class — and their likelihoods are compared to determine the predicted label $\hat{y}$, as described with Equation 5.1. Because Baum-Welch algorithm is a local-maximum search, we ran ten iterations of randomized initial values for each training step and chose the best iteration (based on log-likelihood) for testing. When testing, $P(\boldsymbol{D_{test}} \mid HMM_j)$ represents the "evaluation" step for an HMM and has polynomial complexity of $m^2 * L$ where L is the length of the discussion. In other words, with the trained HMMs, Equation 5.1 can be computed in a quick fashion. For further details and properties of standard HMMs, see [25].

$$\hat{y} = \underset{j \in \{consistent, inconsistent\}}{\text{argmax}} P(\ \boldsymbol{D_{test}} \mid HMM_j) \tag{5.1}$$

Our primary HMM uses Eugenio's features as observations ($\text{HMM}_{\text{Eugenio}}$, $|O| = 4$). A graphical representation is depicted in Figure 5-1. We also built a baseline HMM with conventional DAs ($\text{HMM}_{\text{DAs\_full}}$, $|O| = 11$). In order to balance the number of features and counter the effect of overfitting, a second baseline HMM was built with four conventional DAs[1] (HMM_{DAs}, $|O| = 4$).

In order to incorporate head gestures into our model, we used an early fusion technique of combining both verbal and nonverbal features into a larger feature set. The two modality streams (Eugenio's features and head gestures) were ordered chronologically to form a single stream of observations; i.e., feature-level fusion. Figure 5-2 depicts the resulting $\text{HMM}_{\text{Eugenio+Head}}$, which captures occurrences of both feature sets. The model effectively learns information regarding their transitions. In the future, we intend to investigate alternative fusion techniques, such as decision-level fusion [27], where outputs of single-modality HMMs would be weighted and summed. The baseline for the combined model was an HMM wherein four conventional DAs are added into the set of Eugenio's features ($\text{HMM}_{\text{Eugenio+DAs}}$).

5.2 Prediction Performance

Here, we present the prediction performance of $\text{HMM}_{\text{Eugenio}}$ and $\text{HMM}_{\text{Eugenio+Head}}$. For training and testing, we performed leave-one-out cross validation (LOOCV) in order to maximize the size of the training data per fold. Standard performance measures such as accuracy, recall, precision, F1 score and false positive rate (FPR) were measured. We averaged the results from five different values of m, which we

[1] Four DAs with definitions most relevant to group decision-making were used: assessment, elicit-assessment, comment-about-understanding (CAU) and elicit-CAU.

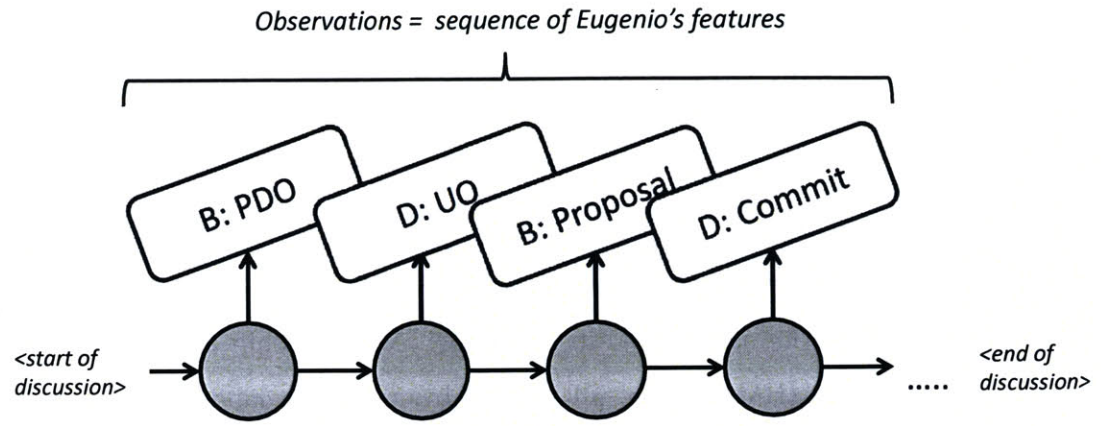

Figure 5-1: A graphical representation of HMM with Eugenio's features as observations (following the order shown in the sample conversation segment in Figure 4-1).

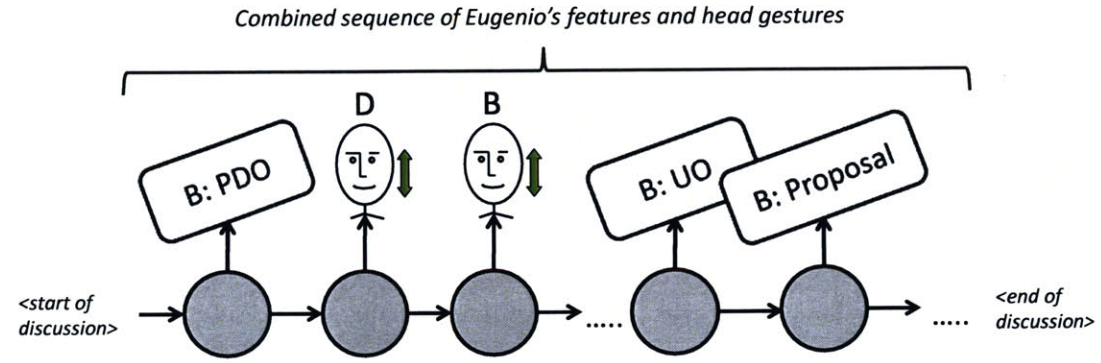

Figure 5-2: A graphical representation of HMM combining both Eugenio's features and head gestures (following the order shown in the sample conversation segment in Figure 4-1).

varied from 1-5.

As shown in Figure 5-3, $\text{HMM}_{\text{Eugenio}}$ resulted in a mean accuracy of 62.1% an increase of 11% compared with HMM_{DAs}. Other measures, such as recall, precision and F1 score, also showed improvement, each with an increase of approximately 10%. There was a 12% reduction to FPR. $\text{HMM}_{\text{DAs\_full}}$ performed very poorly, most likely due to overfitting from a large number of features. Paired t-tests (df = 4) between $\text{HMM}_{\text{Eugenio}}$ and HMM_{DAs} indicated statistically significant differences ($\alpha = 0.05$) across all performance measures. These results demonstrated that using Eugenio's features improves overall prediction performance compared with conventional DAs.

	$\|O\|$	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
HMM_{DAs_full}	*11*	50.7	29.3	23.1	25.8	40.4
HMM_{DAs}	*4*	51.4	36.5	31.0	33.5	41.1
$HMM_{Eugenio}$	*4*	**62.1**	**44.7**	**43.8**	**44.2**	**29.5**
P-value		0.01	0.02	0.001	0.007	<0.001

Figure 5-3: Prediction performance of $HMM_{Eugenio}$ and baselines. The p-values reflect comparisons between $HMM_{Eugenio}$ and HMM_{DAs}.

	$\|O\|$	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
$HMM_{Eugenio}$	4	62.1	44.7	43.8	44.2	29.5
$HMM_{Eugenio+DAs}$	8	45.1	39.1	39.1	39.1	50.0
$HMM_{Eugenio+Head}$	**8**	**64.2**	**55.3**	**47.3**	**51.0**	**31.1**
P-value		0.28	0.02	0.18	0.03	0.49

Figure 5-4: Prediction performance of $HMM_{Eugenio+Head}$ and baselines. The p-values reflect comparisons between $HMM_{Eugenio+Head}$ and $HMM_{Eugenio}$.

When evaluating Figure 5-4, we first noted that $HMM_{Eugenio+DAs}$ performed much more poorly than $HMM_{Eugenio}$. In this case, additional features reduced overall performance. With $HMM_{Eugenio+Head}$, however, there was an increase in mean accuracy, recall, precision and F1 score compared with $HMM_{Eugenio}$. Although $|O|$ was doubled, there did not seem to be a negative overfitting effect. The increases to accuracy and precision were small approximately 2-4% and paired t-tests indicated that only the improvements to recall and F1 score were statistically significant. We observed a small increase to FPR; however, this change was not significant.

5.2.1 Robustness across different meeting phases

We performed four-fold cross validation and compared prediction performance across the four distinct meeting phases in the AMI corpus. As described in Table 5.1, each meeting phase was fundamentally unique with regard to agenda and the topics under discussion. Similar prediction performance across meeting phases would indicate the robustness of our model to phase-specific keywords and topics. Figure 5-5 depicts

this comparison, highlighting the accuracies of $HMM_{Eugenio+Head}$, $HMM_{Eugenio}$ and HMM_{DAs}.

Table 5.1: Four distinct meeting phases in the AMI corpus [8]

Meeting Phase	Discussion
Project kick-off	Getting acquainted with one another and discussing the project goals
Functional design	Setting user requirements, technical functionality and working design
Conceptual design	Determining conceptual specifications for components, properties and materials
Detailed design	Finalizing user interface and evaluating the final product

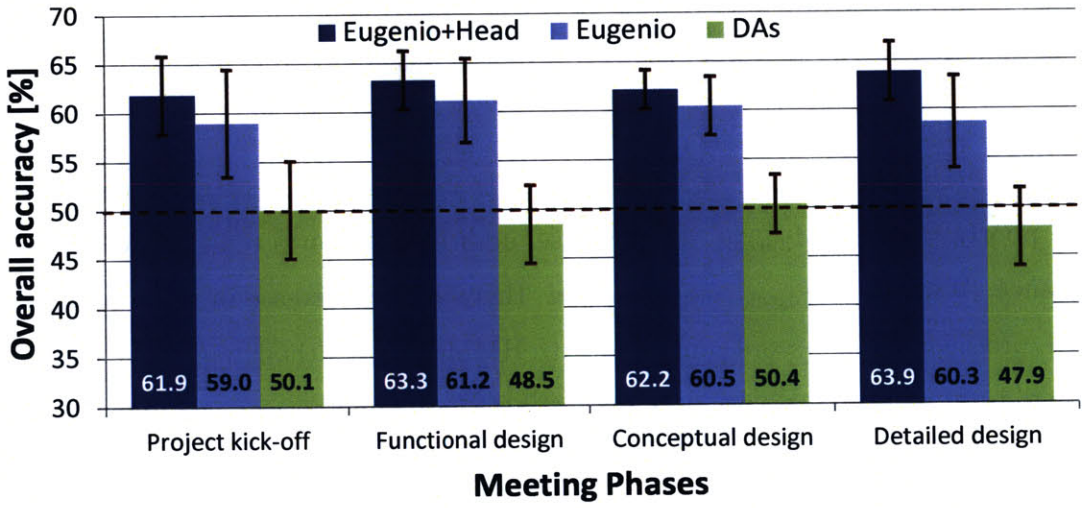

Figure 5-5: Comparison of model accuracies across different meeting phases

The mean accuracies for all three HMMs remained similar across the different meeting phases, though the values were slightly lower than the global numbers presented in Figures 5-3 and 5-4. This was to be expected, as four-fold CV has less available training data per fold than LOOCV. We observed a trend toward increasing accuracy from $HMM_{DAs} \rightarrow HMM_{Eugenio} \rightarrow HMM_{Eugenio+Head}$, which was consistent across all four meeting phases.

5.2.2 Comparison with other learning algorithms

Lastly, we compared the prediction performance of the HMM model to other supervised machine learning algorithms. Specifically, we applied support vector machines (SVM) with radial basis function kernel, logistic regression and a Naïve Bayes classifier with a Gaussian density assumption. The input vectors for these algorithms corresponded to the frequency of Eugenio's features (e.g., a topic discussion can have a total of three "proposals" and two "commitments").

The purpose of our comparison was to investigate the utility of applying generative, dynamic Bayesian models, such as HMMs, against frequency-based approaches. Figure 5-6 shows the comparison on a receiver operating characteristic (ROC) curve. The ROC curve plots recall against the false positive rate at various meta-parameter settings. The diagonal line represents the baseline performance of random chance. We used the area under the curve (AUC) statistic for model comparison. (Head gestures were not incorporated for this section; we focused only on the set of Eugenio's features.)

HMM outperformed the other learning algorithms with an AUC of 0.671, supporting our hypothesis for using HMMs in the context of our problem. $m = 3$ was the best setting for the HMM with regard to maximizing accuracy with reasonable recall and FPR tradeoffs. Naïve Bayes yielded the poorest performance, most likely due to its strong independence assumption between the features.

5.2.3 Discussion

Not only did the HMM trained using Eugenio's features result in prediction performance above random chance, but it also outperformed the HMM trained with conventional DAs. These findings indicate that an informative signal exists within the set of Eugenio's features for predicting consistency of understanding. Essentially, the notion of using DAs to follow how a team generates plans seemed to carry relevant information for distinguishing consistency. However, the choice of the DA set matters, as we found that an HMM trained with conventional DAs resulted in poor

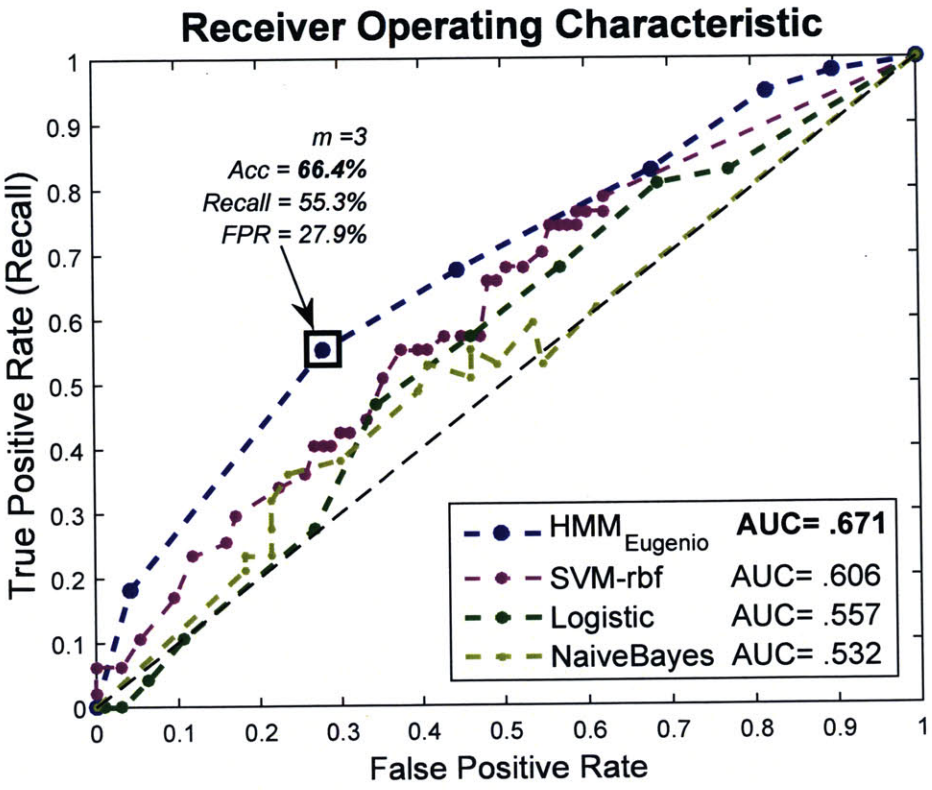

Figure 5-6: ROC curve comparing different prediction algorithms. AUC is reported.

prediction performance. Our results quantitatively verify the utility of Eugenio's features, specifically in the context of capturing information regarding a team's shared understanding.

When head gestures were incorporated into the model, there were statistically significant increases to recall and F1 score, along with non-significant increases to accuracy and precision. Although this combined model yielded positive changes, more statistical evidence is required to conclude improvement. When we tested the performance of an HMM trained only with head gestures, prediction performance was very poor, with accuracy close to 50%. Head gestures alone did not seem to provide an informative signal toward the prediction of consistency of understanding; it was only when they were included with Eugenio's features that signs of a potential benefit emerged. We suspect that this is the product of a strong imbalance within the set of head gestures: 98% of all head gestures in the AMI dataset were characterized by

head nods, with 54% labeled as "concord" and 44% as "emphasis." Head shakes and tilts comprised only 2% of all head gestures. This indicates that participants rarely display head gestures that explicitly convey "discord" or "negative" signals.

We believe that there must be a finer level of granularity particularly within head nods in order to further characterize a person's cognitive intent. However, recovering accurate intentionality from head gestures is a separate and challenging research problem. Also, the utility of features depends upon the chosen learning model. To further investigate the utility of head gestures, alternative computational models or fusion techniques (e.g. coupled-HMMs [7]) can be employed. In the future, we would like to incorporate additional audiovisual modalities, such as vocal intonation, gaze and hand gestures.

With similar accuracies and their consistent ordering through different feature sets ($\text{HMM}_{\text{Eugenio+Head}} > \text{HMM}_{\text{Eugenio}} > \text{HMM}_{\text{DAs}}$), our approach demonstrated robustness across different meeting phases. This was an initial investigation of generalizability, conducted internally within the AMI dataset. We hope to test how our approach generalizes to external meeting datasets, such as the ICSI [33], the VACE [9] and the Wolf [32] corpora, in future study. It is important to focus on meetings that are collaborative and goal-oriented, such that the consistency of understanding is a relevant measure. The biggest challenge to testing other meeting datasets is that they lack sufficient layers of annotations: most do not include self-reported participant summaries, which are necessary to label ground truth on consistency of understanding.

When integrating our computational model for an online system, high recall and low FPR are particularly important. High recall signifies a high hit rate of detecting discussion topics with inconsistency; the system can then provide suggestions to review those topics. Low FPR is also important to reduce the incidence of false alarms within the system. Incorrect predictions and false feedback would be disruptive and could cause human teams to lose trust in the system. The "best" model setting at $m = 3$ (boxed in Figure 5-6) optimizes over these considerations and performs with an accuracy of **66.4**%, recall of 55.3% and FPR of 27.9%. (We later use this setting for model implementation.) A simple predictor labeling the most dominant class would

result in similar accuracy but zero recall, rendering the system useless. A system capturing only 55.3% of true inconsistent topics can still be helpful to human teams as long as the FPR is low, which would cause the system to report inconsistency only when it is highly confident in the result.

Chapter 6

Model Implementation and Evaluation

We implemented and evaluated our computational model using an intelligent agent system that provides review suggestions during meetings. In order to do this, we developed a Web-based collaboration tool and conducted a set of experiments with human subjects.

6.1 Web-based Tool Design

Emergency response teams increasingly use Web-based tools to coordinate missions and share situational awareness among team members. One of the tools currently used by first-responders is the Next Generation Incident Command System (NICS) [18]. This command-and-control system allows a distributed team of responders to efficiently exchange information and coordinate mission planning. It provides a rich set of communication channels, including audio and video conferencing, text chat, a shared map, drawing tools, resource logs and situational information.

We have designed a Web-based tool modeled after this system, with a modification that only allows the team to communicate via text. The tool contains a standard window for text-based chat, a shared map of the environment, a distributed information log and a list of topic discussions. Within the text chat, the software runs

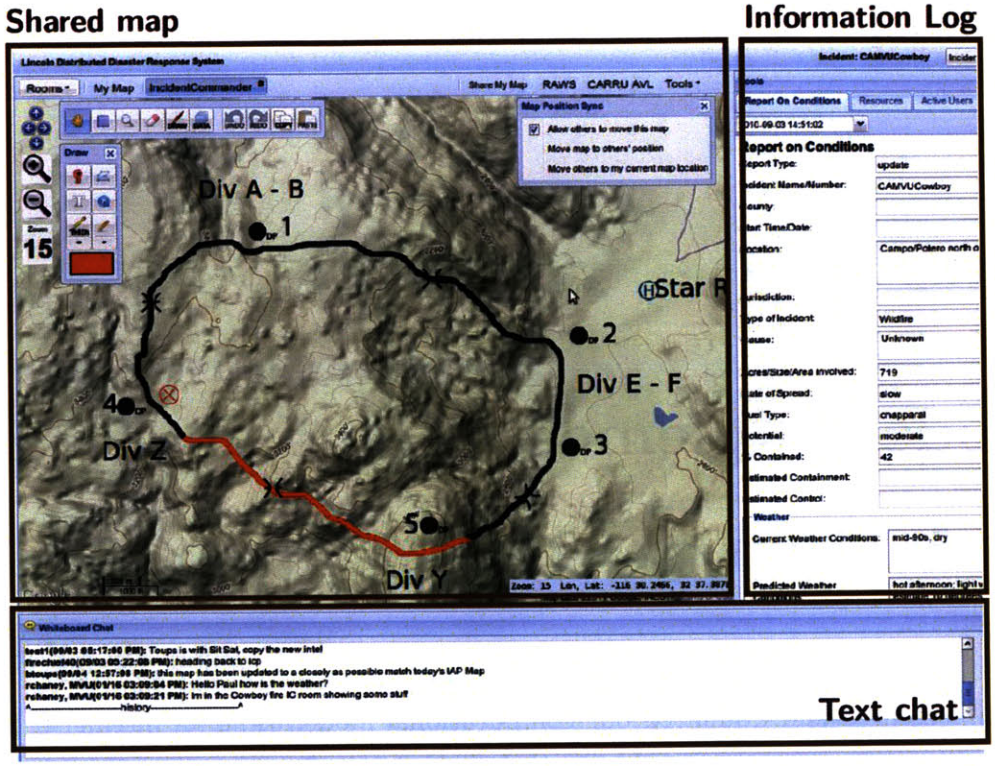

Figure 6-1: A snapshot of the Next Generation Incident Command System (NICS)

an $HMM_{Eugenio}$ trained using the AMI corpus that predicts the consistency of understanding for each topic. Segmentation of utterances is provided by participants' natural turn-taking when using the text chat (a new line of utterance is triggered whenever "Enter" is pressed). Figure 6-2 depicts a snapshot of our collaboration tool. Although our software represents a simplified version of the NICS, it captures the essence of this emerging technology for emergency response coordination.

The tool implements an algorithm that automatically tags dialogue acts [40], assigning the most likely DA label for a given utterance. In implementing the algorithm, we used a bigram classifier with Jelinek-Mercer smoothing [10]. This classifier was trained using the AMI corpus over 11 different DA classes[1], and achieved a classification accuracy rate of 72%.

For the experiment, we preprocessed incoming text in order to reduce noise and

[1]The following AMI DA classes were not considered: 'Stalls,' 'Fragments,' 'Be-Negatives' and 'Other.'

increase DA classification accuracy as much as possible. For example, the preprocessor removed articles, punctuations, verbal fragments and stop words such as {"uh," "um," "hmm," etc.}. We also added additional training utterances pertaining to our scenario, derived from five iterations of pilot study. These steps were taken so that the uncertainty of the intelligent agent system would be primarily attributed to the higher-level $HMM_{Eugenio}$ rather than inaccuracies in the low-level DA classifier. In our post-experimental analysis, the "effective" tagging accuracy was 80%. Using the tagged DAs, we applied the coding scheme in Table 4.1 to generate sequences of Eugenio's features.

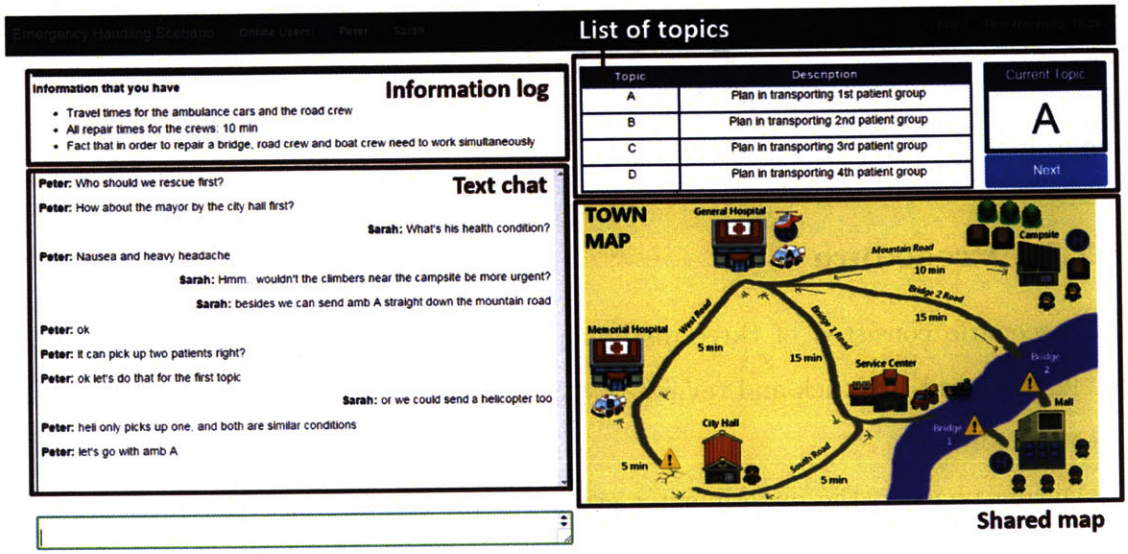

Figure 6-2: A snapshot of our Web-based collaboration tool

6.2 Experiment Design

6.2.1 Task

Fifteen teams of two participants each acted as first-responders in a hypothetical emergency scenario. Their goal was to develop a plan to transport several injured patients to hospitals. There were multiple factors to consider, including variations in the patients' health, travel times, road conditions and transportation capabilities.

Due to the limited number of transports, participants had to prioritize patient delivery and determine ideal travel routes. The overall scenario design was inspired by an existing work on collaborative planning for hypothetical emergency response, the Monroe Corpus [53], and is similar with regard to the process of collaborative problem-solving and encouraging mixed-initiative interaction. However, it should be noted that the Monroe Corpus was an observational study, while our work was an experiment with integration of an intelligent agent: The tool analyzed team chat in real-time and applied a set of experimental treatments during the planning process.

With knowledge and resources distributed among the participants, collaboration was essential for successful completion of the scenario; one participant could not dominate and solve the scenario effectively. The relationship dynamic between the two participants in each team was that of equal collaborators, rather than a supervisor-subordinate relationship.

6.2.2 Procedure

Each scenario consisted of three distinct phases: 1) the main planning session; 2) intelligent agent feedback and review; and 3) individual post-meeting summaries and questionnaires.

During phase 1, the experimenter explained the scenario and described the collaboration tool. Afterwards, participants held their main planning session, communicating with one another through the text chat. Specifically, participants were asked to identify patient groups and set their emergency priority, such that transport plans could be discussed for one group at a time. These partial plans represented distinct topic discussions, where the plan for transporting the first patient group was marked as "Topic A," the plan for the second patient group as "Topic B," and so on. The table of topics depicted in Figure 6-2 illustrates this breakdown. To the right of the table, there was a "Current Topic" indicator that reminded the team which patient group they were currently discussing. Once the team members agreed that they had finished forming a plan for the current patient group, they clicked the 'Next' button to signify that they would move on to discuss a plan for the transport of the next

patient group. This process repeated until the team concluded their discussion about the fourth patient group ("Topic D"). Clicking the 'Next' button naturally provided the topic segmentations. Participants were allotted 20 minutes for the entire main planning session, simulating the time-critical nature of emergency response.

After the team had completed their main planning session, the intelligent agent provided feedback during phase 2 by suggesting two topics out of the four for the team to review. (A detailed explanation of the selection process for review topics is provided in the following section.) The suggestion from the agent was displayed in a pop-up window, as shown in Figure 6-3. Once the team confirmed receipt of the suggestion, they engaged in a 5-minute review session reiterating their plans for the suggested topics.

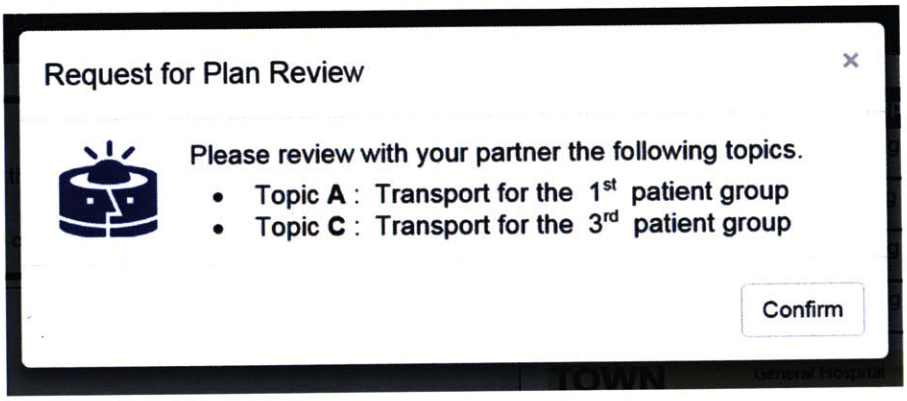

Figure 6-3: Phase 2: The intelligent agent suggests that the team review plans for the selected topics

During phase 3, the participants completed individual post-meeting summaries, writing down detailed plan descriptions for each of their discussion topics. They were permitted as much time as needed to provide the summaries, which were then checked by the annotators to objectively measure consistency of understanding. Participants also responded to post-experiment questionnaires, offering subjective evaluations of their perceived shared understanding and the utility of the review suggestion.

Phases 1 through 3 represented the procedure for a single scenario, and each team completed two scenarios with alternating treatment. (The treatment order was randomized to mitigate learning effect.) Although two scenarios had similar goals for

patient delivery, their detailed environments were different.

The entire experiment took teams approximately 60 minutes to complete. Each participant was compensated $10 for their time.

6.2.3 Experimental Treatment

The topics suggested for review by the intelligent agent varied according to treatment, and the choice of topic represented our treatment levels. The two treatment levels depicted in Table 6.1 were inspired by a related review protocol presented in [56].

In order to explain our treatment levels, we must first need to define a **consistency score**, or a normalization between two HMM likelihoods from Equation 5.1. Mathematically, it represents the posterior probability, $P(\hat{y} = consistent \mid \boldsymbol{D_{test}})$ with a uniform prior assumption. It represents a numerical level of consistency on a scale from 0 to 1, where a score closer to 1 signifies that the discussion is predicted to be highly *consistent* and a score closer to 0 indicates the discussion is highly *inconsistent*. Instead of taking argmax, the normalized score provides more information regarding the "confidence" of consistency. For the sake of brevity, we will refer to this a **predicted c-score**.

Table 6.1: Type of review suggestion by the intelligent agent

Treatment level	Definition
1. Adaptive review	System suggests review of the two topics with the **lowest** predicted c-scores (*weak* topics)
2. Maladaptive review	System suggests review of the two topics with the **highest** predicted c-scores (*strong* topics).

For our treatment, the system always suggested two topics for review. In order to determine which topics to present, the predicted c-scores of the four discussion topics were ranked. In treatment (1) *adaptive review*, the system selected the two topics with the lowest predicted c-scores; we refer to this as reviewing the "weak" topics. Essentially, this treatment represents what is desired for an intelligent system: prompting teams to review the topics with the greatest potential to result in conflicts and misunderstandings. In comparison, the baseline treatment (2) *maladaptive review*

suggested the topics with the highest predicted c-scores, or the topics for which the system already predicts strong consistency within the team. We refer to this as reviewing the "strong" topics.

6.2.4 Dependent Measures

Dependent measures were split into two categories: an objective measure of consistency score and subjective measures self-reported by the participants. The objective measure of consistency (or **objective c-score** to be short) was obtained by comparing the alignment of decision points across the individual post-meeting summaries using a standardized rubric associated with our scenarios. This rubric, depicted in Table 6.2, listed specific decision points and assigned weighted scores for their alignment. An accumulated score of 100% would signify the perfect alignment of all decision points.

Annotation of objective c-scores was completed for each topic discussion. There was a substantial inter-rater agreement between two annotators ($\kappa = 0.70$).

Subjective measures were obtained through participants' rating their perceived utility of the review phase, and whether or not they thought the system suggested the correct topics for review, on five-point Likert scales. These questions are shown in Table 6.3.

Table 6.2: Rubric for Objective Measure of Consistency

Item	Description	Score [%]
☐ Patient	*Same set of patients? Correct health conditions?*	[25]
☐ Transport	*Same transport type?*	[12.5]
	Same letter of the transport vehicle?	[12.5]
☐ Route	*Same start and end locations?*	[12.5]
	Same roads being utilized?	[12.5]
☐ Other details	*Any roads, bridged fixed? Same set of simultaneous events?*	[25]

Table 6.3: Subjective Questionnaires

Measure	Questionnaire Items
Perceived utility	"*The review phase of topics suggested by the system helped my teammate and I reach a stronger understanding over those topics.*"
Perceived recall	"*The system suggested the two topics where there was potential for lack of understanding between my teammate and I.*"

6.2.5 Hypothesis

We formed our hypotheses to test the relationship between the type of review and the measures of a team's consistency.

H1: Adaptive review, or a review focused on topics that had the lowest predicted c-scores, will increase teams' objective c-scores on those topics compared with a baseline without review. Meanwhile, maladaptive review, or a review focused on topics that had the highest predicted c-scores, will not increase objective c-scores for those topics compared with its no review baseline.

H2: There will be an improvement to overall meeting score (the average of all four topic objective c-scores) when participants receive adaptive review compared with maladaptive review.

H3: There will be an improvement to the participants' perceived utility of the review suggestion with adaptive review compared with maladaptive review.

6.2.6 Participants

Fifteen teams of two, for a total of 30 participants (17 males and 13 females), took part in the experiment. Twenty-six of the 30 participants were students from the MIT campus, including undergraduate and graduate students and postdoctoral associates. The remaining four identified themselves as a professional engineer, a software

developer, a scientist and a housekeeper, respectively. The average participant age was 23.8 ($SD = 4.33$) years, ranging from 18 to 38 years. Two-thirds of the participants knew their partners prior to the experiment. On a five-point Likert scale, the participants reported a high degree of familiarity with text-based Web chat ($M = 4.47$, $SD = 0.73$, $Md = 5$, $IQR = 1$).

6.3 Statistical Analysis and Results

Here, we present the details of our statistical analysis of the experimental data and evaluations of the proposed hypotheses.

In order to test *H1*, we performed a set of two paired t-tests to evaluate the utility of an intelligent agent suggesting topics for review following a meeting. The paired t-tests were appropriate for our repeated measures experiment design, wherein each team received both treatments. The t-tests assessed within-subject differences, with "subject" representing a team of two participants. Objective c-scores were measured per topic discussion for each team.

Our experiment was based on the premise that while the act of review would always be helpful for increasing a team's consistency, the significance of this improvement would differ according to the topics reviewed. Our first paired t-test compared the difference in objective c-scores between reviewing and not reviewing the weakest topics (adaptive review), while the second paired t-test compared the difference between reviewing and not reviewing the strong topics (maladaptive review). In satisfying the assumptions of the statistical test, no significant outliers existed in the data, and the assumption of normality was not rejected by the Shapiro-Wilk test ($W = 0.91$, $p = 0.12$).

Figure 6-6 depicts the results of the paired t-tests, with each bar graph indicating the mean values of objective c-scores and standard errors. There was a significant effect on objective c-scores from reviewing weak topics, as indicated on the left plot ($t(14) = 3.29$, $p < 0.01$). The 95% confidence interval of the mean difference was [6.08, 28.92]. The positive direction of the confidence interval confirmed a statistically

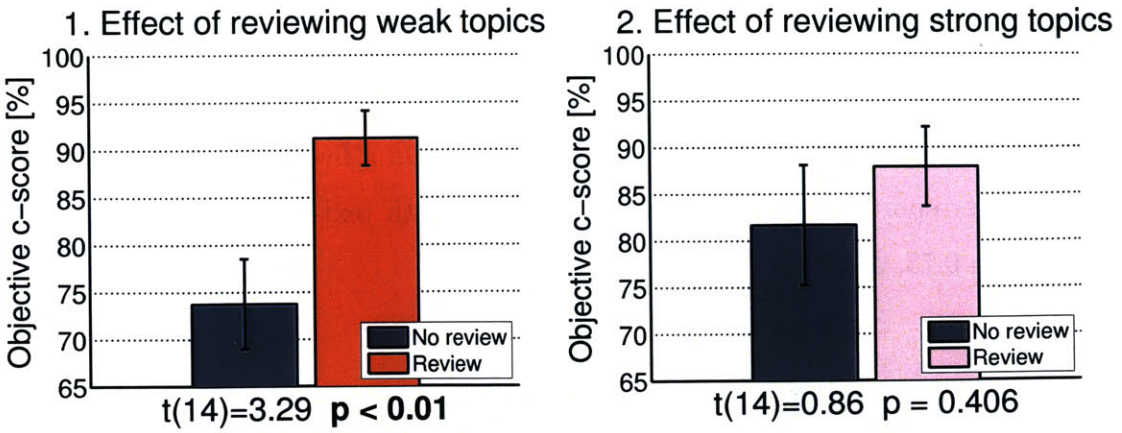

Figure 6-4: Mean values of consistency scores, with error bars indicating standard errors of the mean. The results illustrate that adaptive review had a positive effect on weak topics, increasing the mean of objective c-scores from a no-review baseline of 73.7% to 91.2%. Meanwhile, maladaptive review yielded no statistically significant difference between a review of strong topics and no review.

significant increase, with a mean difference of 17.5%. As illustrated by the right plot, there was no statistically significant difference in objective c-scores when reviewing strong topics ($t(14) = 0.86$, $p = 0.406$). These results provided strong support for both aspects of *H1*.

Figure 6-5 shows a histogram of predicted c-scores, grouped by whether the topics were identified as "weak" or "strong." Mean and standard deviation among the weak topics were 0.471 and 0.058 respectively. Among the strong topics, mean and standard deviation were 0.550 and 0.033 respectively. Most of the distribution lied between the range of 0.4 and 0.6. This plot illustrates the difference in the ranges of the two distributions. Also, it shows that a score of 0.5 can reasonably act as a discriminative threshold.

In order to test *H2*, overall meeting scores were computed using the mean of all discussion topics' objective c-scores. The score represents a numerical level of a team's consistency for the entire meeting, with each topic discussion assigned equal importance. The left plot in Figure 5-4 compares teams receiving adaptive and maladaptive review, and indicates insufficient evidence to support a statistically significant difference between the two ($t(14)=1.20$, $p = 0.25$).

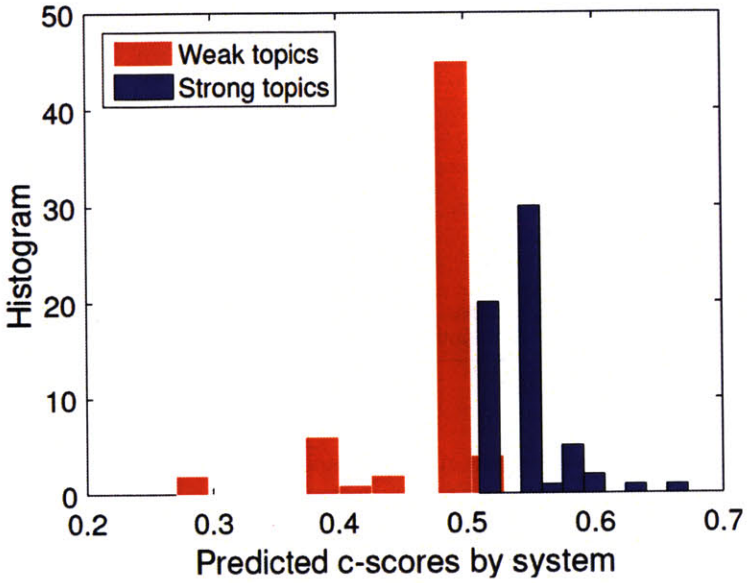

Figure 6-5: Distribution of predicted c-scores.

For subjective measures, we used the Wilcoxon signed-rank test (nonparametric equivalent of paired t-test) to analyze paired differences on a five-point Likert scale. The results showed no significant effect of the type of review on perceived utility ($W = 119$, $p = 0.595$); however, a borderline statistically significant difference was observed for perceived recall ($W = 284.5$, $p = 0.062$).

6.4 Discussion

Reviewing the weak topics suggested by the system (those with low predicted c-scores) resulted in a statistically significant improvement to teams' objective c-scores specifically, a mean improvement of 17.5% over the baseline of not reviewing weak topics. On the other hand, there was no significant difference between reviewing and not reviewing strong topics (those with high predicted c-scores). That a significant improvement occurred only when weak topics were reviewed suggests that the system, on average, chose the "correct" topics for review those with probable inconsistency and a greater potential for the review to improve shared understanding among team members. The experiment demonstrated that the type of review suggested is related

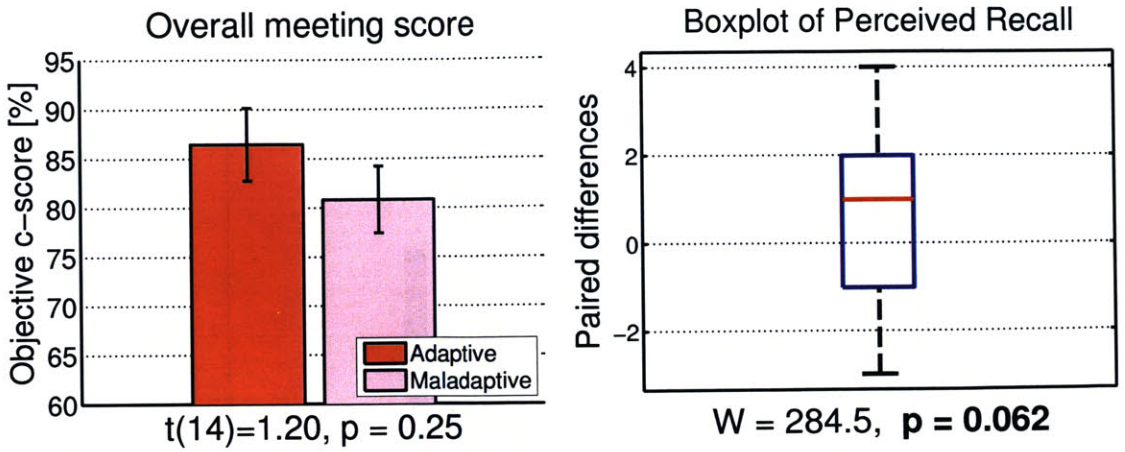

Figure 6-6: A comparison of overall meeting score ($average\{alltopics\}$) is shown in the left plot. The right boxplot depicts the paired difference of medians for perceived recall.

to varying improvements of consistency based on predicted c-scores.

Our results support the notion that simply reviewing all topics is a non-optimal strategy. There is utility behind an intelligent, selective review; one that optimizes over the number of topics discussed during the review session for the most effective improvement to shared understanding. Also, reviewing unnecessary material can potentially be detrimental: It may lead to annoyance among participants, who would be required to re-discuss topics that they already have developed strong opinions about. The frequent occurrence of such false positives can reduce participants' trust in the system and reduce the effectiveness of review.

The difference in overall meeting score was not statistically significant across types of review; therefore, *H2* was not supported. Due to the averaging effect over all four discussion topics, even those with no review, we suspect there may be a loss of sensitivity. These results do not necessarily confound with *H1*, since our original focus was to investigate improvements at the topic level.

We observed no significant effects of the type of review on perceived utility with regard to subjective measures. Participants' perception of the utility of the review phase did not differ significantly across treatments, even though there was a significant objective difference in consistencies. We suspect that the utility of the review

may not be apparent to humans, or it may be that the participants have fundamentally different criteria for judging this utility. Which aspects of review (reiteration, confirmation, clarification, addition of details, changes to plans, etc.) participants consider helpful may vary across different groups of people.

Meanwhile, there was a nearly statistically significant difference in perceived recall: With adaptive review, participants felt more strongly that the system suggested topics that contained the potential for a lack of understanding. In contrast to perceived utility, perceived recall measured participants' direct assessment of the system's topic selection. The borderline significance of this difference in perceived recall was supportive of *H1*. The contrast in the differences of two subjective measures is interesting to note here. It may be possible that even if participants recognize that certain topics have a greater potential to lead to a lack of understanding than others, this does not necessarily mean that they will find a review of those topics to be helpful. For instance, a topic may be difficult to discuss in general, but a participant can still maintain a strong level of confidence in the team's shared understanding of that topic. Another example would be participants who consider the meeting content and planning to be trivial. Such participants would view the review phase as altogether unnecessary, regardless of whether or not certain topics result in greater inconsistency than others. Further statistical evidence would be required to fully support *H3*.

Overall, our computational model learned from the AMI dataset demonstrated utility when implemented in the context of an intelligent review system. Even with 66.4% theoretical prediction accuracy, the model successfully translated to a 17.5% improvement in teams' consistency of understanding, and demonstrated a suitable framework for guiding which topics should be reviewed following initial discussion. The experimental result also provides supporting evidence for the generalizability of the model: While AMI meetings were focused on product design, the learned model transferred and demonstrated utility within the domain of emergency response planning.

Chapter 7

Conclusion and Future Work

In this thesis, we have presented a computational model to predict teams' consistency of understanding in meetings. The model expands upon prior literature by enabling an automatic framework for assessing shared understanding — a form of shared cognition which has previously only been analyzed qualitatively. The model incorporates a set of dialogue acts that focuses on capturing group decision-making dynamics and learns discriminative sequences with a machine learning algorithm. Using the AMI dataset, the model achieved a prediction accuracy rate of 64.2% and demonstrated robustness across different meeting phases. The model's HMM formulation also outperformed other widely used machine learning algorithms.

We then implemented the learned model within an intelligent system that participated in human planning meetings for a hypothetical emergency response mission. Running the computational model, the system suggested the topics that the team would benefit most from reviewing with one another. Through human subject experiments, we evaluated the utility of such a feedback system and observed a statistically significant increase (17.5%) to objective measures of teams' consistency of understanding as compared with a baseline, non-intelligent system.

Overall, we have presented a novel framework for predicting consistency of understanding using only textual data and with no prior knowledge of domain-specific content. We have shown that there exists a predictive signal in the monitoring of team planning dynamics through dialogue features proposed from qualitative studies.

And we have shown that this signal can be leveraged to design an intelligent system that can positively affect a team's shared understanding.

Our problem was motivated with the application for emergency response, but it is applicable to other safety-critical domains and to everyday meetings requiring group decisions. Our model can be potentially integrated to popular web chat software (e.g. Google chat) or other integrated situational awareness-sharing tools like NICS. Our multi-step study combines the strength of human communications research and machine learning with a vision for developing an intelligent system that would help teams to achieve stronger group understanding in meetings.

7.1 Future Work

Our model requires an input stream of Eugenio's features, which were derived from conventional DAs. Therefore, the success of the high-level HMM depends on the low-level DA classifier. We implemented an off-the-shelf algorithm and applied pre-processing to obtain an 80% classification rate. In future work, we would like to investigate the sensitivity of the high-level HMM as it relates to inaccuracies of the low-level DA classifier. We would also like to investigate error propagations resulting from additional input layers, such as a speech recognition tool.

Our current design relies on accurate, manual topic segmentations. For the computational model, the AMI corpus already contained segmented topic boundaries. In our experiment, it was supplied through a signal given by the participants: the clicking of the 'Next' button. In order to design a more independent system, automatic topic segmentation tools must be integrated. This would be especially important when observing physical meetings incorporating live speech. This is a challenging research problem, as participants may switch back and forth spontaneously between or diverge from topics.

In our experimental design, two topics out of four were always suggested by the system. The two topics were ranked with respect to each other and the two with lowest scores were suggested. One could imagine an alternative system where a feedback is

triggered with respect to an absolute threshold (i.e. a topic suggested if it falls below a threshold value. The "best" threshold value can be learned from the data). Consequently, the number of review topics would become a free variable, and it would be possible to have all or none of the topics reviewed. We would like to investigate such a setting in future. We would need to investigate if this generalizes appropriately across teams. Some teams may be better at reaching consistent understanding than others, where the differences of predicted c-scores could be substantial. Aggregating the data and setting a global threshold may not transfer well to all teams.

Instead of just dialogue acts, an advanced system will attempt to uncover more information on the planning details of the conversation. This is a scenario where domain-specific content would have a high utility. The problem then becomes associated with "plan inference" and the resulting feedback from the intelligent agent would change. Instead of designating a topic to be "weak", the agent would scope further and present subcomponents of a topic that cause consistency scores to drop. Such a system would be able to search through different planning predicates (e.g. send(A to B), move(C to D), etc.) and pinpoint where a potential misunderstanding has occurred. We also envision a system in the future that can automatically suggest alternative plans (e.g. *"sending truck to B is not optimal due to heavy traffic, why don't you try route C?"*).

Our design represents a prototype that highlights one potential means for intelligent agent support in meetings. Other avenues for future research might include the design of tools for real-time visualization of consistency of understanding. A numerical score could be visualized and updated dynamically as discussions unfold, providing constant feedback for human teams. Review suggestions could be provided as weak discussion points are discovered online, rather than in a batch format upon completion of the meeting. During physical meetings, the method of feedback from the intelligent agent is also an important variable for investigation (e.g., feedback through speech synthesis or through a screen visualization). It would also be interesting to compare prediction performance between a computer and a human moderator in future work.

Bibliography

[1] http://www.operationhope.org/news/nid/1248. [Online; last accessed 15-May-2015].

[2] Meetings in America: A study of trends, costs and attitudes toward business travel, teleconferencing, and their impact on productivity. In *A network MCI Conferencing White Paper*, volume 3. Greenwich, CT: INFOCOMM, 1998.

[3] Pieter J Beers, Henny PA Boshuizen, Paul A Kirschner, and Wim H Gijselaers. Common ground, complex problems and decision making. *Group Decision and Negotiation*, 15(6):529–556, 2006.

[4] Eva Alice Christiane Bittner and Jan Marco Leimeister. Why shared understanding matters–engineering a collaboration process for shared understanding to improve collaboration effectiveness in heterogeneous teams. In *System Sciences (HICSS)*, pages 106–114. IEEE, 2013.

[5] Kristy Elizabeth Boyer, Robert Phillips, Eun Young Ha, Michael D Wallis, Mladen A Vouk, and James C Lester. Modeling dialogue structure with adjacency pair analysis and hidden markov models. In *Proceedings of Human Language Technologies*, pages 49–52, 2009.

[6] Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. Characterizing the effectiveness of tutorial dialogue with hidden markov models. In *Intelligent Tutoring Systems*, pages 55–64, 2010.

[7] Nuria Oliver Brand, Matthew and Alex Pentland. Coupled hidden markov models for complex action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[8] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, and Melissa Kronenthal. The AMI meeting corpus: a pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39. 2006.

[9] Lei Chen, R Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, et al. Vace multimodal meeting corpus. In *Machine Learning for Multimodal Interaction*, pages 40–51. 2006.

[10] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318, 1996.

[11] Herbert H Clark and Susan E Brennan. Grounding in communication. *Perspectives on Socially Shared Cognition*, 13:127–149, 1991.

[12] Business Meetings Utilizing Video Conferencing. `http://www.tdsengineeringgroup.com/category/video-conferencing/`. [Online; last accessed 15-May-2015].

[13] Business Meetings Utilizing Video Conferencing. `http://www.tdsengineeringgroup.com/category/video-conferencing/`. [Online; last accessed 15-May-2015].

[14] Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. Human-computer dialogue simulation using hidden markov models. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 290–295, 2005.

[15] Peter Darch, Annamaria Carusi, and Marina Jirotka. Shared understanding of end-users' requirements in e-science projects. In *IEEE International Conference on E-Science Workshops*, pages 125–128, 2009.

[16] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[17] Jane F Desforges and Joseph F Waeckerle. Disaster planning and response. *New England Journal of Medicine*, 324(12):815–821, 1991.

[18] Ray Di Ciaccio, Jared Pullen, and Paul Breimyer. Enabling distributed command and control with standards-based geospatial collaboration. In *IEEE International Conference on HST*, pages 512–517, 2011.

[19] Barbara Di Eugenio, Pamela W Jordan, Richmond H Thomason, and Johanna D Moore. The agreement process: an empirical investigation of human-human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53(6):1017–1076, 2000.

[20] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time Vision for Human-Computer Interaction*, pages 181–200. 2005.

[21] Gerald Friedland and Oriol Vinyals. Live speaker identification in conversations. In *Proceedings of the 16th ACM International Conference on Multimedia*, pages 1017–1018, 2008.

[22] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[23] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[24] Sebastian Germesin and Theresa Wilson. Agreement detection in multiparty conversation. In *2009 International Conference on Multimodal Interfaces*, pages 7–14, 2009.

[25] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

[26] Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[27] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pages 3437–3443, 2005.

[28] J. Hall. Americans know how to be productive if managers will let them. *Organizational Dynamics*, 1994.

[29] F Herrera and E Herrera-Viedma. A model of consensus in group decision making under linguistic assessments. *Fuzzy Sets and Systems*, 78(1):73–87, 1996.

[30] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proccedings of HLT-NAACL Conference*, 2003.

[31] S. Hoarn. Distributed battle management: DARPA seeks to manage increasingly complex contested airspace. In *Defense Media Network*, 10 March, 2014.

[32] Hayley Hung and Gokul Chittaranjan. The Idiap Wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the international conference on Multimedia*, 2010.

[33] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, 2003.

[34] Gang Ji and Jeff Bilmes. Dialog act tagging using graphical models. In *Proc. ICASSP*, volume 1, pages 33–36, 2005.

[35] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[36] Maaike Kleinsmann, Jan Buijs, and Rianne Valkenburg. Understanding the complexity of knowledge integration in collaborative new product development teams: A case study. *Journal of Engineering and Technology Management*, 27(1):20–32, 2010.

[37] Maaike Kleinsmann and Rianne Valkenburg. Barriers and enablers for creating shared understanding in co-design projects. *Design Studies*, 29(4):369–386, 2008.

[38] Janice Langan-Fox, Jeromy Anglim, and John R Wilson. Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 14(4):331–352, 2004.

[39] J. J. Lee, W.B. Knox, J.B. Wormwood, C. Breazeal, and D. DeSteno. Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4(893), 2013.

[40] Max M Louwerse and Scott A Crossley. Dialog act classification using n-gram algorithms. In *FLAIRS Conference*, pages 758–763, 2006.

[41] A. Mackenzie and P. Nickerson. *the time trap: the classic book on time management.* Amacom Books, 2009.

[42] William W Maddux, Elizabeth Mullen, and Adam D Galinsky. Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negotiation outcomes. *Journal of Experimental Social Psychology*, 44(2):461–468, 2008.

[43] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2):273–283, 2000.

[44] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *In Proc. IEEE ICASSP, Hong Kong*, 2003.

[45] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. Automatic analysis of multimodal group actions in meetings. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):305–317, 2005.

[46] P.R. Monge, C. McSween, and J. Wyer. A profile of meetings in corporate America: results of the 3M meeting effectiveness study. Annenberg School of Communications, University of Southern California, Los Angeles, CA, 1989.

[47] L Morency. Modeling human communication dynamics. *Signal Processing Magazine, IEEE*, 27(5):112–116, 2010.

[48] R.K. Mosvick and R. Nelson. *We've got to start meeting like this! a guide to successful business meeting management.* Glenview, IL: Scott, Foresman, 1987.

[49] Ingrid Mulder, Janine Swaak, and Joseph Kessels. Assessing group learning and shared understanding in technology-mediated interaction. *Educational Technology & Society*, 5(1):35–47, 2002.

[50] Alex Pentland and Tracy Heibeck. *Honest signals.* MIT Press, Cambridge, MA, 2008.

[51] N.C. Romano and J.F. Nunamaker Jr. Meeting analysis: findings from research and practice. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences.* IEEE, 2001.

[52] Paul R Smart, David Mott, Katia Sycara, Dave Braines, Michael Strub, and Nigel R Shadbolt. Shared understanding within military coalitions: A definition and review of research challenges. *Knowledge Systems for Coalition Operations, Southampton, UK*, 2009.

[53] Amanda J. Stent. The monroe corpus. Technical Report 728 and Technical Note 99-2, University of Rochester, 2000.

[54] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[55] Daniel Szafir and Bilge Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20, 2012.

[56] Daniel Szafir and Bilge Mutlu. Artful: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1001–1010, 2013.

[57] Yasir Tahir, Umer Rasheed, Shoko Dauwels, and Justin Dauwels. Perception of humanoid social mediator in two-person dialogs. In *Proceedings of the 2014 International Conference on Human-robot Interaction*, pages 300–301, 2014.

[58] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional Science*, 39(3):283–301, 2011.

[59] L. Yang, Y. Xu, and C. S. Chen. Human action learning via hidden Markov model. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 27(1):34–44, 1997.

[60] Youngjin Yoo and Prasert Kanawattanachai. Developments of transactive memory systems and collective mind in virtual teams. *The International Journal of Organizational Analysis*, 9(2):187–208, 2001.

[61] Klaus Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.