This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Improving Team's Consistency of Understanding in Meetings

Joseph Kim and Julie A. Shah

***Abstract*—Upon concluding a meeting, participants can occasionally leave with different understandings of what had been discussed. Detecting inconsistencies in understanding is a desired capability for an intelligent system designed to monitor meetings and provide feedback to spur stronger shared understanding. In this paper, 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 the dynamics of group decision-making. We trained a hidden Markov model using the AMI meeting corpus and achieved a prediction accuracy of 64.2%, as well as robustness across different meeting phases. We, then, implemented our model in an intelligent system that participated in human team planning about a hypothetical emergency response mission. The system suggested topics that the team would derive the most benefit from reviewing with one another. Through an experiment with 30 participants, we evaluated the utility of such a feedback system and observed a statistically significant increase of 17.5% in objective measures of the teams' understanding compared with that obtained using a baseline interactive system.**

***Index Terms*—Adaptive review, consistency of understanding, dialogue acts (DAs), hidden Markov models (HMMs), human–computer interaction, intelligent agent participation.**

I. Introduction

MEETINGS are an integral component of many collaborative and organized work environments. However, 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 meeting inefficiencies [1]. One common source of inefficiency is inconsistency between team members in their understanding of the outcome of a meeting [1].

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. A system with this capability could serve to reduce misunderstandings and hidden conflicts that could have gone unnoticed.

Papers [2] and [3] include qualitative models to explain the process of how teams reach a consistent, or shared, understanding of one another. 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.

We present a computational model to predict the consistency among team members' understanding of their group decisions (defined as *consistency of understanding*). We utilize a set of dialogue features that focuses on capturing the dynamics of group decision-making and incorporate them on a machine learning algorithm. Without relying on any domain-specific content, we trained our model using one of the largest publicly available meeting datasets. We demonstrate the utility of our model when it is implemented for an intelligent agent participating in live meetings. Through human subject experiments, we investigate the utility of an intelligent review system.

Overall, our multistep study makes the following contributions: 1) We demonstrate that a computer can automatically assess the consistency of understanding within a team through natural dialogue. 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 provide a review recommendation to improve teams' shared understanding.

A preliminary version of this paper can be found at [4]. This paper expands upon the preliminary version by including the development of a Web-based collaboration tool and an evaluation of the model through human subject experiments.

II. Related Work

A. 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 [5]. Shared understanding represents an alignment of mental states and, therefore, presents difficulties for explicitly monitoring its continually evolving process [2].

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 [2]," "the ability to coordinate behaviors toward common goals or objectives [6]," and "having mutual knowledge, mutual beliefs, and mutual assumptions (content and structure) on the task [7]." Our definition of "consistency of understanding" is synonymous, but provides a clear emphasis on the overlap and alignment of understandings.

Manuscript received June 25, 2015; revised November 23, 2015 and February 24, 2016; accepted March 21, 2016. This paper was recommended by Associate Editor P. Antunes.

The authors are with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: joseph_kim@csail.mit.edu; julie_a_shah@csail.mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2016.2547186

2168-2291 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Shared understanding has positive effects on production performance (with regard to both quality and quantity of products) [8], individual satisfaction [9], reduction of iterative loops and rework [10], and team morale [11].

The process of how shared understanding is achieved has been investigated. Mulder *et al.* [2] 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.* [3] 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 (coconstruction), and negotiate an agreement upon a mutually shared perspective (constructive conflict). Eugenio *et al.* [5] 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.

Prior work has been purely qualitative, focused on formal 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 paper, we present a computational model to predict the levels of shared understanding, using quantitative measures.

B. Natural Language Processing: Agreement Detection and Productivity Analysis

Prior work within the natural language processing (NLP) community has explored the related task of automatically detecting "agreements" in meetings [12], [13]. This task involves the detection and classification of agreements as positive or negative through machine learning algorithms.

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. However, this study 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.

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 with 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.

There has been work on building statistical models to predict productivity [14], [15] and moments when key decisions are made [16]. Various linguistic and structural features of dialogue have been utilized to infer the productivity of a meeting as a whole [14] and to analyze the evolution of productivity levels within a meeting [15]. Although related, we believe productivity and shared understanding are two fundamentally different group interaction variables. Analyzing their correlations and dependences is an interesting problem, which we leave for future work.

C. Intelligent Agent Participation

Intelligent agents are being integrated into tasks such as automatic summarization [17], detection of meeting actions [18], modeling of social interactions [19], and audiovisual processing of various cognitive states [20]. In the case of the latter, researchers are developing models to infer participants' states of concentration, interest, and confusion [21] and have used intelligent agents to predict the outcomes of interviews [22] and the success of negotiations [23]. Smart interfaces have been developed to suggest review topics in online education [24] and to generate feedback in personal tutoring systems [25]. Robots have been integrated into meetings—for example, to serve as moderators in balancing engagement and dominance levels [26], and to influence conflict dynamics during team problem-solving tasks [27], and to predict levels of interpersonal trust [28]. Our work addresses the novel task of predicting the consistency of understanding during team meetings. This problem is unique, in that it involves prediction of a shared cognitive state.

III. Approach

Our problem statement is to automatically predict the consistency of understanding given a team's natural dialogue. The focus is on learning through textual data; however, we also investigate the potential benefits of incorporating nonverbal features, such as head gestures.

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. We perform a prediction task for each topic discussed in a meeting. We believe this is an important level of granularity so that the system can make targeted suggestions on topics that the team would derive the most benefit from reviewing with one another.

Fig. 1 depicts the flowchart of our problem statement. A topic of discussion is read as input by the computational 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. Consistency of understanding, including information on its ground truth labeling, is described in Section IV-B. For topics that the model predicts will result in inconsistency, a system feedback is triggered suggesting that team members review those topics.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

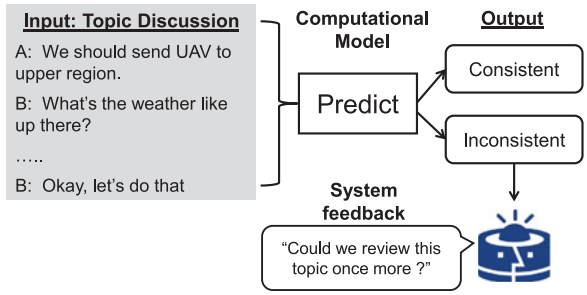

Fig. 1. Flowchart of the problem statement. The input is a topic discussion, and the output is consistency of understanding. System feedback is triggered for topics predicted to be inconsistent.

TABLE I
EUGENIO'S FEATURES

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

One challenge 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. 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], collaborative intentionality [30], and agreement [5]. 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 by Eugenio *et al.* [5] (referred to here as *Eugenio's features*), which has been shown to monitor the evolving attitude of participants' commitment toward options[1] presented during a meeting. Eugenio's features are types of *dialogue acts* (DAs) [31], [32], which are semantic labels that define the functional roles of utterances. A DA expresses the underlying intention of the speaker's utterance. For example, a DA label of "Suggest" expresses the speaker's intention relating to the actions of a group or other team members. "Inform" denotes the speaker's exchange of information relevant to the discussion topic. Additional DA labels with example utterances are shown in Table II. Conventional DA sets [33], [34] are widely used to annotate conversations in many existing meeting corpora [35]–[37].

Eugenio's features are distinguished from conventional DAs in that they are focused on capturing the speaker's commitment toward proposed ideas and choices to be decided upon by the group. These labels are especially suited for annotation of goal-oriented and team decision-making meetings and facilitate the recognition of implicit and/or passive acceptances. For example, the feature "Unendorsed Option" denotes an occurrence in which an option is put forth by the speaker, with no subsequent comment from the other team members. Alternatively, the feature "Commit" denotes the speaker's full commitment toward an option. The full set of Eugenio's features is presented in Table I.

In contrast with existing meeting corpora, we code meetings with Eugenio's features to better capture how joint commitments are achieved by the group. The aforementioned characteristics of Eugenio's features are then useful for predicting consistency of understanding, as joint commitment toward options would naturally lead to joint understanding of group decisions.

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

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

where $\boldsymbol{D}$ is a topic discussion, DA_i is a DA 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 DAs. For $\boldsymbol{\Lambda}$, our primary feature set incorporates Eugenio's features.

There are advantages to using DAs to represent a discussion. First, DAs support 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, DAs 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 reduced.

In addition, we investigate the benefits of multimodal fusion by including head gestures as an extended feature set. Head gestures have been used to infer a state of agreement, disagreement, concentration, interest, or confusion [21]. We test whether the combination of head gestures with textual features improves the prediction performance.

[1]"Options" here refers to proposed ideas and choices to be decided upon by the group [5].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
SAMPLE CONVERSATION SEGMENT FROM THE AMI CORPUS

Line	Speaker	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 .. [pause]	Assess		B: Concord
5	B	The ability to locate 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 basically small microphone ...	Inform	UO	
9	B	If you could look into what we've suggested so far, the feasibility of small transmitter..	Suggest	Proposal	
10	C	Okay. Sure.	Assess	Commit	C: Concord

Eugenio's features were annotated using the coding scheme described in Section IV-C. The remaining annotations were provided as part of the standard AMI dataset.

IV. DATASET

The dataset we used to build and train our model comes from the AMI meeting corpus [36]. In each of these meetings, a team of four people collaborated on a task related to the product design. The corpus consisted of discussions conducted by 35 unique teams. The meetings were divided into four distinct phases of the design process (see Table VI) and were scenario-driven. Each participant served one of four roles: project manager, industrial designer, marketing expert, or user interface designer. Although the participants were given information and documentation about their role, they were free to make decisions as they wished. The conversations that occurred during these meetings were tailored toward a group decision-making process. The use of Eugenio's features is 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.

The AMI corpus contains preexisting annotations of topic segmentations, participant summaries, DAs, and head gestures. We augmented the existing dataset by annotating the conversation with Eugenio's features using the coding scheme described in Section IV-C. Next, we describe the manner in which we used each annotation layer to construct the components of our model.

A. 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, and product customizability.

B. 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 provided written summaries of all decisions made by the group. We compared the summaries and determined whether the participants reported the same decisions for the given topic. If all decision points were consistent, the associated topic discussion was labeled *consistent*; the discussion was identified as *inconsistent* if one or more of the summaries differed in content.[2] Two annotators performed the comparison (interrater agreement, $\kappa = 0.73$), resulting in ground truth labels for a total of 140 topic discussions. Ninety-three 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 [38], [39]. Other measurement alternatives include structured interviews and Likert scale questionnaires about perceived shared understanding [2], [40]. 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.

C. From Dialogue Acts to Eugenio's Features

The AMI dataset provides annotations of conventional DAs, but not Eugenio's features. However, conventional DAs can be used to construct Eugenio's features given the knowledge of "solution sizes" [5]. 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. 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." With DAs and solution sizes, we applied the conversion rules below (defined in [5]) to construct Eugenio's features.

1) *Proposal:* Action-directive (AD)[3], offer + determinate
2) *Partner decidable option (PDO):* AD, offer + indeterminate
3) *Commit:* Offer, assessment (positive) + determinate

[2]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. Alternate methods for "partial" consistency labeling using multiclass classification and regression can be explored for future work.

[3]Action-directives (AD) correspond to suggestions and all elicit forms of DAs that require actions from partners.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
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) *Unendorsed options (UO):* Open-options + determinate

D. 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 and emphasis). We incorporated gestures intended to communicate understanding and comprehension. Table III highlights the description of head gestures we used. Table II also shows the head gesture annotations in the conversation segment.

E. 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. In the following section, we describe how we utilized the training data to predict consistency of understanding for a test discussion, $\boldsymbol{D}_{\text{test}}$.

V. COMPUTATIONAL MODEL

We modeled our problem using hidden Markov models (HMMs) [41] 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 [28], [42]. An HMM is defined as a 5-tuple $\{S, O, A, B, \pi\}$, where:

1) S is the finite set of hidden states, and $m = |S|$ is its cardinality. One interpretation of the hidden states is that they serve as representations of different shared understanding processes [2], [3], [5]. For example, they may represent Bossche *et al.*'s definitions, wherein the group may be going through a state of construction or coconstruction 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 metaparameter for the prediction model.
2) O is the finite set of observations. An observation at each time step is a DA 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 (see Table I). 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.
3) A is the state transition matrix, of size m by m, and describes the probability distribution of transitioning between discussion states. The Markov assumption is generally accepted due to the frequent occurrences of adjacency pairs in dialogue [43].
4) B is the observation probability matrix. It describes the emission probability of an observation (DA) conditioned on a hidden discussion state. With a combination of A and B, the stochastic process of O is fully described.
5) π is the initial hidden state distribution.

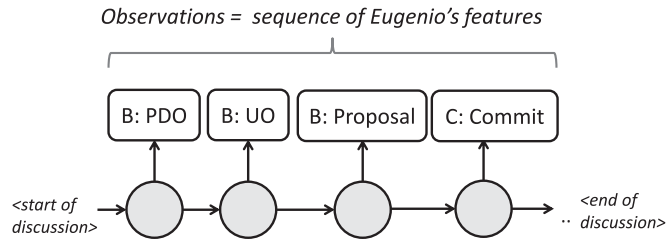

Fig. 2. Representation of HMM with Eugenio's features as observations (order from the sample conversation segment in Table II).

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 [44] 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 $\hat{y}$, described as

$$\hat{y} = \underset{j \in \{\text{consistent, inconsistent}\}}{\text{argmax}} \; P(\boldsymbol{D}_{\mathbf{test}} | \text{HMM}_j). \tag{1}$$

Our primary HMM uses Eugenio's features as observations ($\text{HMM}_{\text{Eugenio}}$, $|O| = 4$). A graphical representation is depicted in Fig. 2. 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[4] (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. Fig. 3 depicts the resulting $\text{HMM}_{\text{Eugenio+Head}}$, which captures occurrences of both feature sets. The model effectively learns information regarding their transitions. The baseline for the combined model was an HMM wherein four conventional DAs are added into the set of Eugenio's features ($\text{HMM}_{\text{Eugenio+DAs}}$).

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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

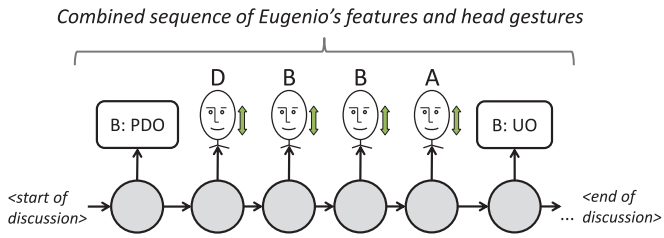

Fig. 3. Representation of HMM combining both Eugenio's features and head gestures (order from the sample conversation segment in Table II).

TABLE IV
PREDICTION PERFORMANCE OF $\text{HMM}_{\text{EUGENIO}}$ AND BASELINES

	$\lvert O \rvert$	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
$\text{HMM}_{\text{DAs_full}}$	11	50.7	29.3	23.1	25.8	40.4
HMM_{DAs}	4	51.4	36.5	31.0	33.5	41.1
$\text{HMM}_{\text{Eugenio}}$	4	**62.1**	**44.7**	**43.8**	**44.2**	**29.5**

TABLE V
PREDICTION PERFORMANCE OF $\text{HMM}_{\text{EUGENIO+HEAD}}$ AND BASELINES

	$\lvert O \rvert$	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
$\text{HMM}_{\text{Eugenio}}$	4	62.1	44.7	43.8	44.2	29.5
$\text{HMM}_{\text{Eugenio+DAs}}$	8	45.1	39.1	39.1	39.1	50.0
$\text{HMM}_{\text{Eugenio+Head}}$	8	64.2	**55.3**	47.3	**51.0**	31.1

TABLE VI
FOUR DISTINCT MEETING PHASES IN THE AMI CORPUS [36]

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

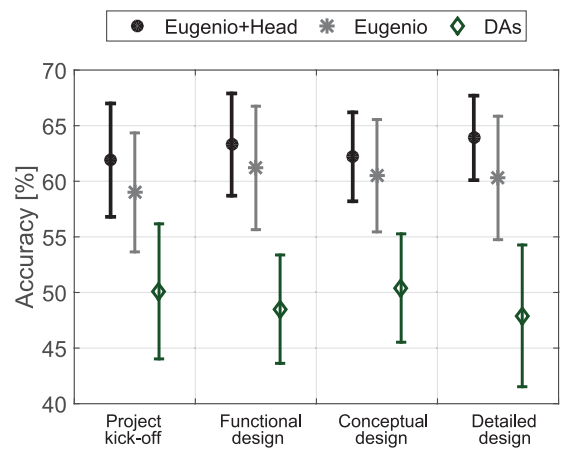

Fig. 4. Model accuracies (with 95% confidence intervals) across different meeting phases.

VI. PREDICTION PERFORMANCE

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 varied from 1–5.

As shown in Table IV, $\text{HMM}_{\text{Eugenio}}$ resulted in a mean accuracy of 62.1%—an increase of 11% compared with HMM_{DAs}. Paired t-tests indicated statistically significant differences across all performance measures between $\text{HMM}_{\text{Eugenio}}$ and each of the two baseline HMMs ($t(4) > 2.78$, $p < 0.02$ for all planned comparisons). This demonstrated that using Eugenio's features improves the overall prediction performance compared with using conventional DAs.

When evaluating Table V, we first noted that $\text{HMM}_{\text{Eugenio+DAs}}$ performed much more poorly than $\text{HMM}_{\text{Eugenio}}$. In this case, additional features reduced overall performance. With $\text{HMM}_{\text{Eugenio+Head}}$, however, there was an increase in mean accuracy, recall, precision, and F1 score compared with $\text{HMM}_{\text{Eugenio}}$. 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 ($p = 0.02$ and $p = 0.03$, respectively).

A. Robustness Across Different Meeting Phases

We performed fourfold cross-validation (CV) and compared prediction performance across the four distinct meeting phases. As described in Table VI, 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. Fig. 4 depicts this comparison, highlighting the accuracies of $\text{HMM}_{\text{Eugenio+Head}}$, $\text{HMM}_{\text{Eugenio}}$, and HMM_{DAs}.

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

B. Comparison with Other Learning Algorithms

Finally, we compared the prediction performance of the HMM model to other supervised machine learning algorithms. Specifically, we applied support vector machines 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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

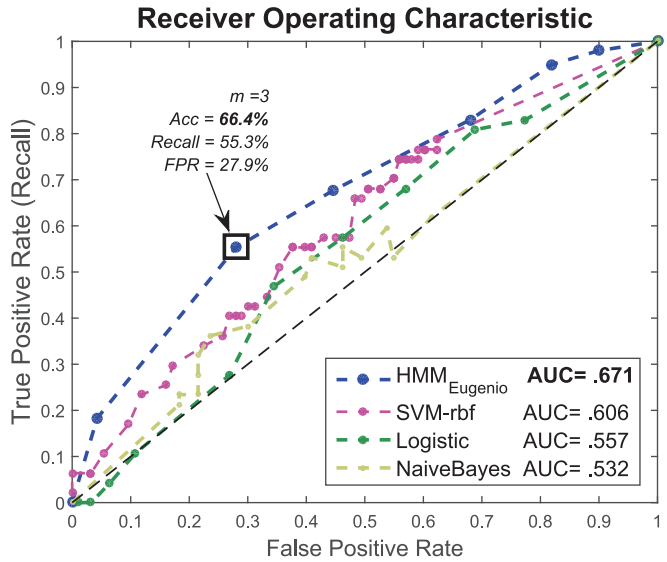

Fig. 5. ROC curve comparing different prediction algorithms. AUC is reported.

Eugenio's features (e.g., a topic discussion can have a total of three "proposals" and two "commits").

The purpose of our comparison was to investigate the utility of applying generative, dynamic Bayesian models, such as HMMs, against frequency-based approaches. Fig. 5 shows the comparison on a receiver operating characteristic (ROC) curve. 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.

Generative models such as an HMM provide the ability to sample from the joint distribution $P(S, O)$ and derive histograms of most frequent observation sequences. We found this capability useful, because the top frequent sequences then allow for interpretability of the model with respect to the trends described in qualitative studies. For example, we can quantitatively verify that consistent meetings often generate the following transition: [PDO → Proposal → Commit], while inconsistent meetings generate [PDO → PDO → UO]. In future work, we would like to test the performances of additional generative and discriminative models for sequence classification, such as hidden conditional random fields [45].

C. 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 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 nonsignificant differences to accuracy and precision. More statistical evidence is required to conclude improvement for the multimodal model. 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. Due to its weak predictive signal, we did not include head gestures as a feature in our experiment (see Section VII).

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. In addition, 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 [46]) 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 [35], the VACE [47], and the Wolf [48] corpora. 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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$m = 3$ (boxed in Fig. 5) 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 the experiment.) 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.

For an online system, we require an automatic DA tagger that assigns the most likely DA label for a given utterance. For our experiment described in Section VII, we used a bigram classifier with Jelinek–Mercer smoothing [49], which was trained using the AMI corpus over 11 different DA classes.[5] This achieved a classification accuracy of 72%.[6] In order to increase the DA tagging accuracy as much as possible, we ran a preprocessor that removed articles, punctuations, verbal fragments, and stop words such as {"uh," "um," "hmm"}. We also added additional training utterances relevant to our experiment scenario, derived from five iterations of pilot study. These steps were taken so that the uncertainty of our computational model would be primarily attributed to the higher level $\text{HMM}_{\text{Eugenio}}$ rather than inaccuracies in the low-level DA classifier. In our postexperimental analysis, the "effective" tagging accuracy was found to be 80%.

VII. Experiment Setup and Protocol

In order to evaluate the utility of our computational model, we implemented it on an intelligent system that provided review suggestions in meetings. It would ask the team to review the inconsistent topics predicted by the system. In order to investigate the change in teams' shared understanding based on the review suggestion, we devised a human subject experiment where participants held their meetings using a Web-based collaboration tool.

A. 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) [52]. 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, resource logs, and situational information.

We have designed a Web-based collaboration tool modeled after this system, with a modification that only allows the team to communicate via text. We leave online integration of audiovisual modalities, such as head gestures, for future work. Our 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. Fig. 6 depicts a snapshot of the tool. Although the software represents a simplified version of the NICS, it captures the essence of the emerging technology for emergency response coordination.

B. Task

Each team of two participants 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 existing work on collaborative planning for hypothetical emergency response: the Monroe corpus [53] and the ELEA corpus [54]. It is similar with regard to the process of collaborative problem-solving and encouraging mixed-initiative interaction. However, it should be noted that aforementioned corpora were observational studies, 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.

C. Procedure

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

During phase 1, participants held their main planning session. The two participants were physically separated from each other and could only communicate through the text chat. Participants were asked to identify patient groups and set their emergency priority, such that transport plans could be discussed for one patient 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 list of topics depicted in Fig. 6 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 min 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. The

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

[6]This is within bounds of current technology, where classification accuracies range from 66% to 75% [50], [51].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

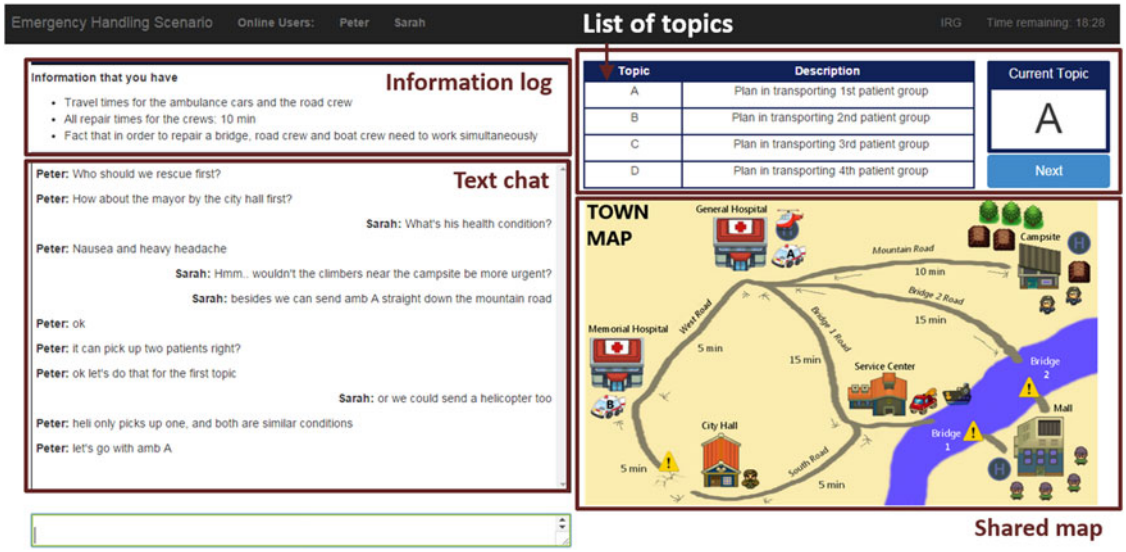

Fig. 6. Snapshot of our Web-based collaboration tool.

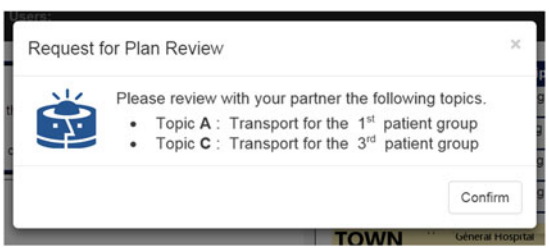

Fig. 7. Phase 2: The intelligent agent suggests that the team review plans for the selected topics.

suggestion from the agent was displayed in a pop-up window, as shown in Fig. 7. Once the team confirmed the suggestion, they engaged in a 5-min review session reiterating their plans for the suggested 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 postexperiment questionnaires, offering subjective evaluations of their perceived shared understanding and the utility of the review suggestion.

Phases 1–3 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 min to complete. Each participant was compensated $10 for their time.

D. Experimental Treatment

The choice of topic suggested for review by the intelligent agent represented our treatment levels. The two treatment levels depicted in Table VII were inspired by a related review protocol presented in [24].

TABLE VII
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).

In order to explain our treatment levels, we must first need to define a *consistency score*, or a normalization between two HMM likelihoods from (1). Mathematically, it represents the posterior probability, $P(\hat{y} = \text{consistent}|\boldsymbol{D}_{\text{test}})$ 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*.

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.

E. Dependent Measures

Dependent measures were split into two categories: an objective measure of consistency score and subjective measures

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VIII
RUBRIC FOR OBJECTIVE MEASURE OF CONSISTENCY

Item	Description	Score [%]
☐ Patient	*Mention the 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 IX
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.*"

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 postmeeting summaries using a standardized rubric associated with our scenarios. This rubric, depicted in Table VIII, 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 interrater 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. These questions are shown in Table IX. The responses were on five-point Likert scales.

F. 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.

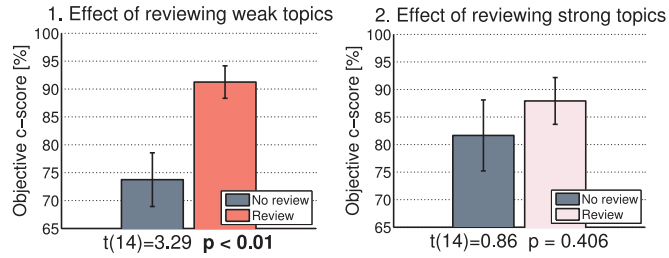

Fig. 8. Mean values of consistency scores, with error bars indicating standard errors of the mean. The results illustrate that adaptive review had a positive effect on weak topics, increasing the mean of objective c-scores from a no-review baseline of 73.7% to 91.2%. Meanwhile, maladaptive review yielded no statistically significant difference between a review of strong topics and no review.

G. 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 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. The participants reported a high degree of familiarity with text-based Web chat.

VIII. STATISTICAL ANALYSIS AND RESULTS

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 ($W = 0.91$, $p = 0.12$).

Fig. 8 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 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*.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

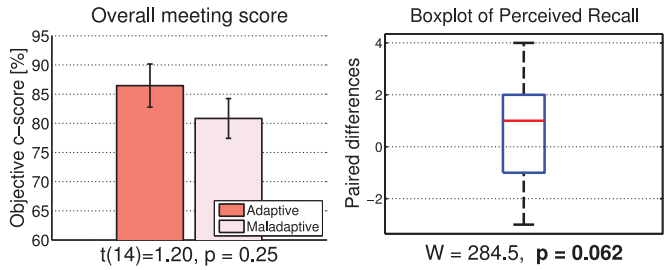

Fig. 9. Comparison of overall meeting score ($average\{alltopics\}$) is shown in the left plot. The right boxplot depicts the paired difference of medians for perceived recall.

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 Fig. 9 compares teams receiving adaptive and maladaptive review and indicates insufficient evidence to support a statistically significant difference between the two ($t(14) = 1.20$, $p = 0.25$).

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 ($W = 284.5$, $p = 0.062$).

IX. 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. The experiment demonstrated that the type of review suggested is related to varying improvements of consistency based on predicted c-scores.

Our results support the notion that simply reviewing all topics is a nonoptimal 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. In addition, reviewing unnecessary material can potentially be detrimental: It may lead to annoyance among participants, who would be required to rediscuss 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 with 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 corpus 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.

A. Limitations and Future Work

In our Web-based tool, segmentation of utterances was provided by participants' natural turn-taking when using the text chat (i.e., a new line of utterance is triggered whenever "Enter" is pressed). If we were to design a speech interface, an automatic segmentation tool would be required.

Our model requires an input stream of Eugenio's features, which were derived from conventional DAs. Therefore, the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

success of the high-level HMM depends on the low-level DA tagging of utterances. We implemented an off-the-shelf algorithm and applied preprocessing 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 tagger.

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, or diverge from, topics.

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).

X. Conclusion

In this paper, 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. The model incorporates a set of DAs that focuses on capturing group decision-making dynamics and learns discriminative sequences with a machine learning algorithm. Using the AMI corpus, the model achieved a prediction accuracy rate of 64.2% and demonstrated robustness across different meeting phases.

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 nonintelligent system.

We have presented a novel framework for predicting consistency of understanding using only textual data and with no prior knowledge of domain-specific content. Our multistep 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.

Acknowledgment

The authors would like to thank Y.-S. Tung for his help in designing the web-based collaboration tool.

References

[1] N. Romano and J. Nunamaker, Jr, "Meeting analysis: Findings from research and practice," in *Proc. 34th Annu. Hawaii Int. Conf. Syst. Sci.*, 2001, p. 13.

[2] I. Mulder, J. Swaak, and J. Kessels, "Assessing group learning and shared understanding in technology-mediated interaction," *Educ. Technol. Soc.*, vol. 5, no. 1, pp. 35–47, 2002.

[3] P. Van den Bossche, W. Gijselaers, M. Segers, G. Woltjer, and P. Kirschner, "Team learning: Building shared mental models," *Instructional Sci.*, vol. 39, no. 3, pp. 283–301, 2011.

[4] J. Kim and J. A. Shah, "Automatic prediction of consistency among team members' understanding of group decisions in meetings," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2014, pp. 3702–3708.

[5] B. Di Eugenio, P. W. Jordan, R. H. Thomason, and J. D. Moore, "The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogs," *Int. J. Human-Comput. Stud.*, vol. 53, no. 6, pp. 1017–1076, 2000.

[6] P. R. Smart, D. Mott, K. Sycara, D. Braines, M. Strub, and N. R. Shadbolt, "Shared understanding within military coalitions: A definition and review of research challenges," presented at the Knowl. Syst. Coalition Oper., Southampton, U.K., 2009.

[7] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives Socially Shared Cognition*, vol. 13, pp. 127–149, 1991.

[8] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, and J. A. Cannon-Bowers, "The influence of shared mental models on team process and performance," *J. Appl. Psychol.*, vol. 85, no. 2, pp. 273–283, 2000.

[9] J. Langan-Fox, J. Anglim, and J. R. Wilson, "Mental models, team mental models, and performance: Process, development, and future directions," *Human Factors Ergonom. Manuf. Service Ind.*, vol. 14, no. 4, pp. 331–352, 2004.

[10] M. Kleinsmann, J. Buijs, and R. Valkenburg, "Understanding the complexity of knowledge integration in collaborative new product development teams: A case study," *J. Eng. Technol. Manage.*, vol. 27, no. 1, pp. 20–32, 2010.

[11] P. Darch, A. Carusi, and M. Jirotka, "Shared understanding of end-users' requirements in e-science projects," in *Proc. IEEE Int. Conf. E-Sci. Workshops*, 2009, pp. 125–128.

[12] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement versus disagreement in meetings: Training with unlabeled data," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguistics Human Lang. Technol.*, 2003, pp. 34–36.

[13] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proc. Int. Conf. Multimodal Interfaces*, 2009, pp. 7–14.

[14] G. Murray, "Learning how productive and unproductive meetings differ," in *Advances in Artificial Intelligence*. New York, NY, USA: Springer, 2014, pp. 191–202.

[15] G. Murray, "Analyzing productivity shifts in meetings," in *Advances in Artificial Intelligence*. New York, NY, USA: Springer, 2015, pp. 141–154.

[16] B. Kim and C. Rudin, "Learning about meetings," *Data Mining Knowl. Discovery*, vol. 28, nos. 5/6, pp. 1134–1157, 2014.

[17] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Comput. Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[18] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[19] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, 2009.

[20] L. Morency, "Modeling human communication dynamics," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 112–116, Sep. 2010.

[21] R. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Proc. Real-Time Vis. Human-Comput. Interaction*, 2005, pp. 181–200.

[22] A. Pentland and T. Heibeck, *Honest Signals*. Cambridge, MA, USA: MIT Press, 2008.

[23] W. W. Maddux, E. Mullen, and A. D. Galinsky, "Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

negotiation outcomes," *J. Exp. Soc. Psychol.*, vol. 44, no. 2, pp. 461–468, 2008.

[24] D. Szafir and B. Mutlu, "ARTFul: Adaptive review technology for flipped learning," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2013, pp. 1001–1010.

[25] D. Fossati, B. Di Eugenio, S. Ohlsson, C. W. Brown, L. Chen, and D. G. Cosejo, "I learn from you, you learn from me: How to make ilist learn from students," in *Proc. 14th Int. Conf. Artif. Intell. Educ.*, 2009, pp. 491–498.

[26] Y. Tahir, U. Rasheed, S. Dauwels, and J. Dauwels, "Perception of humanoid social mediator in two-person dialogs," in *Proc. Int. Conf. Human-robot Interaction*, 2014, pp. 300–301.

[27] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: The case of repairing violations," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2015, pp. 229–236.

[28] J. J. Lee, W. Knox, J. Wormwood, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Frontiers Psychol.*, vol. 4, no. 893, 2013.

[29] F. Herrera and E. Herrera-Viedma, "A model of consensus in group decision making under linguistic assessments," *Fuzzy Sets Syst.*, vol. 78, no. 1, pp. 73–87, 1996.

[30] L. Hunsberger and B. Grosz, "The dynamics of intentions in collaborative intentionality," *Cognition, Joint Action Collective Intentionality, Special Issue, Cognitive Syst. Res.*, vol. 7, nos. 2/3, pp. 259–272, 2006.

[31] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[32] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2005, vol. 1, pp. 33–36.

[33] M. G. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proc. AAAI Fall Symp. Commun. Action Humans Mach.*, Boston, MA, USA, 1997, pp. 28–35.

[34] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. 5th SIGdial Workshop Discourse Dialogue*, 2004, pp. 97–100.

[35] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2003, vol. 1, pp. I-364–I-367.

[36] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, "The AMI meeting corpus: A pre-announcement," in *Proc. Mach. Learning Multimodal Interaction*, 2006, pp. 28–39.

[37] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1992, vol. 1, pp. 517–520.

[38] P. J. Beers, H. P. Boshuizen, P. A. Kirschner, and W. H. Gijselaers, "Common ground, complex problems and decision making," *Group Decision Negotiation*, vol. 15, no. 6, pp. 529–556, 2006.

[39] E. A. C. Bittner and J. M. Leimeister, "Why shared understanding matters—Engineering a collaboration process for shared understanding to improve collaboration effectiveness in heterogeneous teams," in *Proc. 46th Hawaii Int. Conf. Syst. Sci.*, 2013, pp. 106–114.

[40] Y. Yoo and P. Kanawattanachai, "Developments of transactive memory systems and collective mind in virtual teams," *Int. J. Org. Anal.*, vol. 9, no. 2, pp. 187–208, 2001.

[41] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," *Int. J. Pattern Recog. Artif. Intell.*, vol. 15, no. 1, pp. 9–42, 2001.

[42] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Hong Kong, 2003, pp. 748–751.

[43] K. E. Boyer, R. Phillips, E. Y. Ha, M. D. Wallis, M. A. Vouk, and J. C. Lester, "Modeling dialogue structure with adjacency pair analysis and hidden Markov models," in *Proc. Human Lang. Technol.*, 2009, pp. 49–52.

[44] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[45] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[46] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1997, pp. 994–999.

[47] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper *et al.*, "VACE multimodal meeting corpus," in *Proc. 2nd Int. Workshop Mach. Learning Multimodal Interaction*, 2006, pp. 40–51.

[48] H. Hung and G. Chittaranjan, "The Idiap Wolf corpus: Exploring group behaviour in a competitive role-playing game," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 879–882.

[49] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. 34th Annu. Meeting Assoc. Comput. Linguistics*, 1996, pp. 310–318.

[50] R. Serafin and B. Di Eugenio, "FLSA: Extending latent semantic analysis with features for dialogue act classification," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, p. 692.

[51] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1303–1314, Sep. 2008.

[52] R. Di Ciaccio, J. Pullen, and P. Breimyer, "Enabling distributed command and control with standards-based geospatial collaboration," in *Proc. IEEE Int. Conf. Technol. Homeland Security*, 2011, pp. 512–517.

[53] A. J. Stent, "The Monroe corpus," Univ. Rochester, Rochester, NY, USA, Tech. Rep. 728/Tech. Note 99-2, 2000.

[54] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez, "An audio visual corpus for emergent leader analysis," in *Proc. Workshop Multimodal Corpora Mach. Learning: Taking Stock Road Mapping Future*, 2011.

Joseph Kim received the B.S. degree in aerospace engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2012, and the S.M. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015, where he is currently working toward the Ph.D degree with the Department of Aeronautics and Astronautics.

His research interests include computational modeling for human–robot interaction and interactive machine learning.

Julie A. Shah received the S.B. and S.M. degrees in aeronautics and astronautics and the Ph.D. degree in autonomous systems from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2004, 2006, and 2010, respectively.

She is currently an Associate Professor with the Department of Aeronautics and Astronautics, MIT, and leads the Interactive Robotics Group of the Computer Science and Artificial Intelligence Laboratory. She has developed innovative methods for enabling fluid human–robot teamwork in time-critical safety-critical domains, ranging from manufacturing to surgery to space exploration. Her group draws on expertise in artificial intelligence, human factors, and systems engineering to develop interactive robots that emulate the qualities of effective human team members to improve the efficiency of human–robot teamwork.

Dr. Shah was recognized with an NSF CAREER Award for her work on "Human-aware Autonomy for Team-oriented Environments," and by the MIT Technology Review TR35 list as one of the world's top innovators under the age of 35.