Common Ground Tracking in Multimodal Dialogue

Ibrahim Khebour[1], Kenneth Lai[2], Mariah Bradford[1], Yifan Zhu[2], Richard Brutti[2],

Christopher Tam[2], Jingxuan Tu[2], Benjamin Ibarra[1], Nathaniel Blanchard[1], Nikhil Krishnaswamy[1], and

James Pustejovsky[2]

[1]Colorado State University, Fort Collins, CO, USA

[2]Brandeis University, Waltham, MA, USA

{ibrahim.khebour, nkrishna}@colostate.edu,jamesp@brandeis.edu

Abstract

Within Dialogue Modeling research in AI and NLP, considerable attention has been spent on "dialogue state tracking" (DST), which is the ability to update the representations of the speaker's needs at each turn in the dialogue by taking into account the past dialogue moves and history. Less studied but just as important to dialogue modeling, however, is "common ground tracking" (CGT), which identifies the *shared* belief space held by all of the participants in a task-oriented dialogue: the task-relevant propositions all participants accept as true. In this paper we present a method for automatically identifying the current set of shared beliefs and "questions under discussion" (QUDs) of a group with a shared goal. We annotate a dataset of multimodal interactions in a shared physical space with speech transcriptions, prosodic features, gestures, actions, and facets of collaboration, and operationalize these features for use in a deep neural model to predict moves toward construction of common ground. Model outputs cascade into a set of formal closure rules derived from situated evidence and belief axioms and update operations. We empirically assess the contribution of each feature type toward successful construction of common ground relative to ground truth, establishing a benchmark in this novel, challenging task.

Keywords: common ground, multimodality, belief updating

1. Introduction

In the context of increasingly sophisticated interactions involving natural language dialogues with an AI, there is considerable attention being spent on "Dialogue State Tracking" (DST), which is the ability to update the representations of the speaker's (user's) needs at each turn in the dialogue, by taking into account the past dialogue moves and history. In this paper, we address the related but less-studied problem of *Common Grounding Tracking* (CGT), which identifies the shared belief space held by all of the participants in a task-oriented dialogue. We describe the procedure for training CGT models to both identify the current set of beliefs, as well as determine the level of evidence for each, to condition where the dialogue will go (the "questions under discussion", or QUDs). The goal is to provide a more informative snapshot of the dialogue situation, after each action in the task, to develop a policy incorporating shared beliefs in addition to past dialogue history.

A major challenge facing the development of computational models for multimodal interactions involves tracking the intentions, goals, and attitudes of the participants (Cassell et al., 2000; Foster, 2007; Kopp and Wachsmuth, 2010; Marshall and Hornecker, 2013; Schaffer and Reithinger, 2019; Wahlster, 2006). For task-oriented dialogues, just as important is the problem of identifying and tracking the *common ground* between participants (Clark and Brennan, 1991; Traum, 1994; Asher, 1998; Dillenbourg and Traum, 2006).

In this work, we specifically: (a) identify both communicative expressions (speech, gesture) and jointly perceived actions in a multi-party dialogue, in order to convert them into propositional content; and (b) add them to a dynamic data structure we call the *Common Ground Structure (CGS)*. This consists of three parts: **FBank**, a set of facts that are assumed to be known by the group; an **EBank**, a set of evidences available to the group; and **QBank**, the "questions under discussion", a set of topics remaining to be discussed in order to solve the task.

In total, this work encompasses three novel contributions:

- A challenging new task: multimodal common ground tracking, with a formal model of common ground in a shared, situated task;

- A novel incorporation of the formal model into an automated pipeline that tracks the evolution of group common ground over time;

- An augmentation of the Weights Task Dataset (Khebour et al., 2023) with gesture, action, and common ground annotations, to enable the operationalization of our formal model.

Our code may be accessed at https://github.com/csu-signal/Common-Ground-detection

2. Related Work

The present work draws on several diverse areas of research, from modeling common ground in HCI
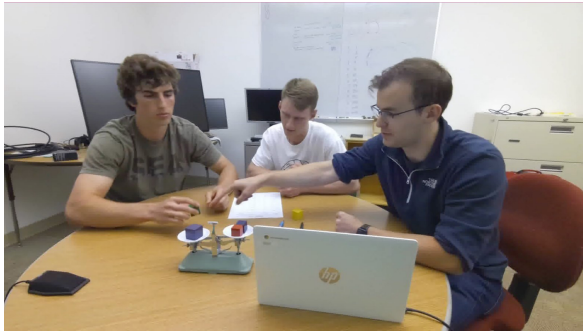
Figure 1: Sample still from the Weights Task Dataset showing communication with multiple modalities. The accompanying utterance at this time is "Put the twenty on there; take off a ten".

and HHI, and Dialogue State Tracking, to the role of gesture in multimodal interactions.

When engaged in dialogue, our shared understanding of both utterance meaning (content) and the speaker's meaning in a specific context (intent), involves the ability to link these two in the act of situationally grounding meaning to the local context, what is typically referred to as "establishing the common ground" between speakers (Grice, 1975; Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998; Traum and Larsson, 2003). The concept of common ground refers to the set of shared beliefs among participants in a Human-Human interaction (HHI) (Markowska et al.; Traum, 1994; Hadley et al., 2022), as well as HCI (Krishnaswamy and Pustejovsky, 2019; Ohmer et al., 2022) and HRI interactions (Kruijff et al., 2010; Fischer, 2011; Scheutz et al., 2011). Del Tredici et al. (2022) have recently employed the notion of common ground operationally to identify and select relevant information for conversational QA system design. Stewart et al. (2021) and Bradford et al. (2023) both study human-human collaboration through the lens of an AI agent.

Dialogue state tracking (DST) aims to estimate the current dialogue state or belief state of the users during the conversation (Budzianowski et al., 2018; Liao et al., 2021; Jacqmin et al., 2022). Current DST models can be categorized into three types: fixed ontology (Henderson et al., 2014; Mrkšić et al., 2017; Chen et al., 2020), open vocabulary (Gao et al., 2019; Hosseini-Asl et al., 2022; Wu et al., 2019) and hybrid methods (Goel et al., 2019; Zhang et al., 2020a; Heck et al., 2020). Recently, pretrained language models have been widely used to model slot relations, while Graph attention networks (GATs) have been used to model the hierarchical structure of DST, enabling the incorporation of semantic compositionality, cross-domain knowledge sharing and coreference (Lin et al., 2021; Li et al., 2021; Cheng et al., 2020).

Understanding the role of nonverbal behavior in multimodal communication has long been a research interest in HCI, but has recently taken on new interest within CL and the broader AI community. Gestures offer an array of unique dimensions in communication, ranging from denoting situational references to indicating specific spatial locations or even conveying manner and orientation (Rohrer et al., 2020; Efthimiou and Kouroupetroglou, 2011; Kong et al., 2015; Kendon, 1997, 2004; Mcneill, 2005). Gesture AMR (GAMR) (Brutti et al., 2022) considers gestures that convey the same propositional content and intentionality as speech acts. Gesture may have meaning on its own, or it may enhance the meaning provided by the verbal modality (Goldin-Meadow, 2003; Krishnaswamy and Pustejovsky, 2020). Also critical to multimodal dialogue is human action, which in addition to communicating deictic and bridging information can also make lasting changes to the world, affecting the common ground (Tam et al., 2023). Much work has been done to facilitate action identification from video (Sigurdsson et al., 2016) (Gu et al., 2018) (Li et al., 2020) as well as to annotate specific semantic roles (Sadhu et al., 2021).

Di Maro et al. (2021) implement dynamic belief sets as graphs, which we do not do explicitly. However, such an approach is theoretically and computationally compatible with ours, as the result (postcondition) of a public announcement or observed action can act as the preconditions for promoting QUDs to evidence, or evidenced propositions to strong beliefs, leading to a natural interpretation of common ground tracking as a graph.

3. Dataset

The Weights Task (Khebour et al., 2023) is a collaborative problem-solving task in which groups of three work together to deduce the weights of differently-colored blocks by making comparisons of block weights using a balance scale. In this activity, the group has a balance scale and five blocks of various colors, sizes, and weights. They are told the weight of one block and must identify the weights of the remaining blocks and, eventually, the algebraic relation between them, which is an instance of the Fibonacci Sequence (Sigler, 2002; Bonacci, 1202). Due to the co-situated nature of the task and its inclusion of physical objects and reasoning about their properties, this task involves communication in multiple modalities, such as language, gesture, and action (see Fig. 1), meaning that knowledge is shared using multiple communicative channels. The Weights Task Dataset (WTD) comes with automatic and human transcriptions of the speech, as well as gesture annotated using Gesture-AMR (GAMR) (Brutti et al., 2022), and collaborative problem solving

(CPS) indicators according to the framework of Sun et al. (2020). All groups successfully deduce the correct block weights, giving a consistent end state against which to assess our models.

3.1. Example Dialogue

Figure 2: Example dialogue. Participant 3 (right) says "looks like they're fairly equal" after placing the red and blue blocks on different sides of the scale. We refer back to this example elsewhere in the paper.

In the WTD, participants are canonically indexed from 1–3, left to right. In Fig. 2, Participant 3 makes a statement that the red and blue blocks are "fairly equal", which is interepreted as an assertion of belief that $red = blue$. Participant 1 give a qualified assent to this through the utterance "yeah, I suppose," meaning that at this point, $red = blue$ and other necessarily entailed propositions can be considered part of the common ground (for example, if we had established that $red < yellow$, then $blue < yellow$ also becomes part of the common ground).

4. Common Ground in Dialogue

Here we assume the context of a multi-participant, task-oriented conversation, involving communication by multiple content-generating modalities (language, gesture) and mutually interpretable nonverbal behaviors (e.g., actions) (Kruijff et al., 2010; Pustejovsky and Krishnaswamy, 2021). To this end, we need a data structure representing the common ground in such a context, that can be dynamically updated throughout the dialogue. We adopt a version of a Dialogue Game Board (DGB), as developed in Ginzburg (2012).

Because of the evolving and dynamic nature of co-interactive dialogue and situated actions, following van Benthem et al. (2014) and Pacuit (2017), we adopt an *evidence-based model* of belief, where our commitments to propositions describing situations or facts are not binary, but are graded, where they can weaken or strengthen de-

pending on available *evidence* for them as the dialogue progresses.

First, however, we define the minimal structure of a task-oriented interaction as a sequence, D, of dialogue steps, where each move in the dialogue takes it into another situation or state. Let $P = \{p_1, p_2, p_3\}$, be the participants in our dialogue. From any situation s_k, we define a D move, m_i, as $m_i = (p_j, C_j, s_{k+1})$: participant p_j performs a communicative act C_j, bringing the multimodal dialogue into situation s_{k+1}. The D can be defined as the sequence of these moves: $D = m_1, \ldots, m_n$.

Here our interest is in tracking the situation content resulting from each move: the set of propositions that captures the current state of the world, the current progress towards a goal, or the status of a task. In addition, it captures the current questions under discussion and beliefs in the dialogue.

Given these considerations, we identify three components for tracking common ground in dialogue: a minimal static model of degrees of belief; a data structure distinguishing the elements of the agents' common ground that are being tracked; and a dynamic procedure which updates this structure, when new information and evidence is available to the agents. We consider each of these in turn below.

4.1. Evidence-based Belief

Pacuit (2017) provides a model for evidence-based belief, where agents obtain evidence in favor of a proposition, φ, and can , to eventually believe φ. We adopt a simplified model of the evidence-based Dynamic Epistemic Logic (EB-DEL) as developed in van Benthem et al. (2014) and Pacuit (2017). We define a model as a tuple, $\mathcal{M} = (W, E, V)$, where

(1) a. W is a non-empty *set of worlds*;
   b. $E \subseteq W \times \wp(W)$ is an *evidence relation*;
   c. $V: At \to \wp(W)$, is a *valuation function*.

Let $E(w)$ denote the set $\{X \mid wEX\}$, the worlds accessible to w through the evidencing relation, E. The evidence-based epistemic language, $\mathcal{L}$, will be the set of formulas generated by the grammar below:

(2) $p \mid \neg\varphi \mid \varphi \wedge \psi \mid [E]\varphi \mid [B]\varphi \mid [A]\varphi$

We distinguish the situation where an agent has "evidence in favor of" a proposition φ, as $[E]\varphi$. Because an agent can have evidence for propositions that convey contradictory information, she can consider both $[E]\varphi$ and $[E]\neg\varphi$. This corresponds to an agent having multiple neighborhoods, X, that are each evidenced in their unique way by w. However, consider the set of non-contradictory worlds as a unique subset of X, one which has what van Benthem and Pacuit (2011) refer to as

the *finite intersection property (fip)*. This property allows us to identify a neighborhood of accessible worlds with non-contradictory propositional content. When this occurs, we say an agent has *belief* in a proposition, $[B]\varphi$. Finally, the universal modality is considered "knowledge" of a proposition, $[A]\varphi$.

4.2. Common Ground Structure

Capturing situational state information in a task-oriented dialogue is critical for reflecting current common ground as well as predicting future dialogue moves (Traum and Larsson, 2003; Schlangen and Skantze, 2011; Zhang et al., 2020b; Jacqmin et al., 2022). For our present purpose, we adopt the notion of a *Dialogue Game Board* (Ginzburg, 1996, 2012), modified to reflect the varying degrees of evidence associated with propositions under discussion. A Common Ground Structure, cgs, is a triple, (QB, EB, FB), consisting of:

(3) a. Questions Under Discussion (QBᴀɴᴋ): set of topics or unknowns that need to be answered to solve the task;
b. Evidence (EBᴀɴᴋ): set of propositions for which there is some evidence they are true;
c. Facts (FBᴀɴᴋ): set of propositions believed as true by all participants.

The task begins with a set of unknowns referred to as the "Questions under Discussion" (QUDs). For this implementation, we create a finite model, including a finite model of questions. For all objects in the domain relating to the task, questions are generated for each relation implicated in the task for that object. For example, in the Weights Task, the goal is to identify the weights of five distinct blocks, and then the algebraic relation between them, i.e., the Fibonacci sequence. The weight of a block ranges between 10 and 50 grams, in 10 gram intervals. Hence, for each block in B, where $B = \{red, blue, yellow, green, purple\}$, we have five possible values, expressed as yes/no questions. Hence, initialization of the QBank results in the following set:

(4) QBank = $\{Eq(r, 10)?, \dots, Eq(r, 50)?, \dots, Eq(p, 10)?, \dots Eq(p, 50)?\}$

At the outset of our dialogue, we set both EBank and FBank to nil, since no task-relevant propositions have been established as commonly evidenced or believed. In the next section, we address the task of determining how information is updated in the dialogue, thereby changing the common ground.

4.3. Updating the Common Ground

Given the epistemic logic presented above, we introduce the mechanisms that update the information state within a dialogue. Following Plaza (1989) and subsequent developments of Public Announcement Logic (Baltag et al., 2016), we introduce a new operator to the model, the announcement operator, !. Public announcements are statements that are made to all agents, and after the announcement, all agents know that the statement has been announced and that it is true.

If $[!\varphi]$ represents the act of announcing φ, then $[!\varphi]\psi$ means "after φ is announced, then ψ is believed to be the case."

In order to distinguish evidence for φ from belief in φ, we relativize the impact of a statement to the context within which it is uttered. Let us interpret $[!\varphi]\psi$ as follows.

(5) a. *Update with Evidence*:
$[!\varphi][E]\psi$: Given the announcement of φ, there is evidence for ψ;

b. *Update with Belief*:
$[E]\varphi \rightarrow [!\varphi][B]\psi$: Belief in φ is conditionalized on φ's announcement in the prior context of evidence for φ.

Semantically, an update represents the state of affairs after an announcement. This entails transforming the current model by removing all states where the announced formula is false. With evidence distinguished from belief/knowledge, we also update the evidence function, where $[!\varphi]$:

(6) a. Updates the worlds: $W' = W \cap \varphi$
b. Updates the Evidence function: $E'(w) = E(w) \cap \varphi$
c. $(M, w) \models \varphi$ implies $(M|_\varphi, w) \models [E]\psi$

This update actually changes the underlying evidence sets themselves. The announcement is taken as a piece of direct evidence. Hence, to capture that the announcement of φ becomes evidence and not just belief, the evidence sets for each agent get restricted (or updated) to reflect the worlds where φ is true. Subsequently, the belief function will then naturally adjust based on the new evidence sets.

Operationally, after (5a) is run, the model is relativized to evidencing neighborhoods, where φ is true. This corresponds to moving a proposition from QBank to the EBank. Then, if the same proposition is "announced" again, as with an *ACCEPT* move, then (5b) promotes that proposition from the EBank to the FBank.

In Section 6.4, we illustrate how these updates are applied when running over the output of the move classifier, in order to determine the content of the current common ground.

5. Annotation

We augmented the existing WTD annotations with dual annotation of GAMR, and participant actions using VoxML (Pustejovsky and Krishnaswamy, 2016). Finally, we also tracked the group's collective surfacing of evidence and acceptance of task-relevant facts by supplying another layer of "common ground annotations" (CGA):

Annotation in the dialogue involves identifying categories relating to the cognitive state of participants to actions and knowledge, concerning the task. This includes the following categories: (a) *OBSERVATION*: participant P_i has perceived an action, a; (b) *INFERENCE*: deduction from φ; (c) *STATEMENT*: announcement of evidence φ; (d) *QUESTION*: introducing role interrogative relating to φ; (e) *ANSWER*; supplying filler to question about φ; (f) *ACCEPT*: agree with evidence φ; (g) *DOUBT*: disagree with evidence for φ.

In the example from Sec. 3.1, Participant 3's utterance would be considered a *STATEMENT* of the proposition $red = blue$ while Participant 1's utterance would be an *ACCEPT* of that proposition. Participant 3 subsequently says "that's 20, these two [referring to the red and blue blocks] are 10" (*STATEMENT* of proposition $red = 10 \wedge blue = 10$), to which Participant 1 says "wait, let's see" signaling a *DOUBT* in $red = 10 \wedge blue = 10$. Therefore at this stage of the dialogue, $red = blue$ can be considered an agreed-upon fact (element of FBank), but none of the participants have accepted that $red = 10 \wedge blue = 10$, so that proposition is still only an element of EBank. This example illustrates the subtleties captured through the annotation.

GAMR, action, and common ground annotations were all dually-annotated. GAMR annotations achieved a SMATCH-F1 score of 0.75. Action annotation achieved an F1 score of 0.67 and Cohen's κ of 0.59. CGA achieved F1 of 0.54 and Cohen's κ of 0.50. Each was adjudicated by an expert to produce the gold standard.

6. Experiments: Modeling Common Ground Tracking

Our experimental pipeline consists of 3 primary components: a *move classifier*, which predicts which cognitive state is being expressed in an utterance (Sec. 5); a *propositional extractor*, which may either consult a dictionary of expressed propositions that is collected from the annotated data with all modalities considered, or may automatically extract the propositional content of an utterance through vector-similarity methods; and a set of *closure rules* that unify the cognitive state and propositional content and update the status of QBank, EBank, and FBank.

Our primary metric for the entire pipeline is Sørensen-Dice coefficient (DSC) (Sørensen, 1948; Dice, 1945). DSC indicates how much the set of propositions extracted by the model matches the set of propositions in the ground truth. It also normalizes for the size of the samples being compared, as the cardinality of the different banks may fluctuate widely as the task proceeds. DSC can also be evaluated as a group proceeds through the task, or averaged over a single group. Since the groups have different numbers of utterances, and hence moves, when aggregating across groups to calculate DSC over time, we pad the length of the shorter groups out to the maximum length by copying the final state of the banks, assuming a "steady state" in the common ground once the task is finished.

6.1. Preprocessing

We first mapped the annotated data to the "oracle" (manually-segmented) utterances in the WTD (Terpstra et al., 2023). If more than one annotation for a given modality was present in the same utterance, we used the one that had the biggest overlap with the oracle utterance.

We encoded the manually transcribed utterances in the WTD into embedding vectors using BERT (`base-uncased`) (Devlin et al., 2019), and extracted the 768D `[CLS]` token embedding from the final encoder layer. Following Bradford et al. (2023), we processed the audio into 88D prosodic features using openSMILE (Eyben et al., 2010). The CPS indicators for an utterance were transformed into their corresponding high-level *facets* according to the Sun et al. (2020) framework, and encoded as 3D one-hot vectors.

GAMR representations were featurized as k-hot encodings of size 81. The first 4 components describe the gesture type (icon, gesture-unit, deixis and emblem) along with a fifth component that represents *and* in cases where 2 types are annotated. Following this is one hot encoding of the GAMR `ARG0` (gesturer), then a k-hot encoding vector of components of the GAMR `ARG1` (gesture content, such as object of deixis). Given the vocabulary of items in the data, this comprises 68 dimensions. Finally, GAMR `ARG2` (gesture recipient) was represented by a one-hot encoding of size 5 (group/researcher/1 of 3 participants). As more than one participant can have a GAMR annotated for them over the same utterance, we allow for 1 GAMR feature vector per participant, resulting in a total GAMR feature size of 243.

Action annotations comprise scale actions, which were vectorized as a one-hot representation of the scale status (left/balanced/right), and participant actions. Participant actions comprise a 2D representation of "lift" vs. "put", a one hot encoding for the block being acted upon, a 2D one hot encoding of "in" and/or "on", and a 2nd one-hot object

representation of block, scale, or table (the destination of the action, where applicable). A participant's action vector is of size 25 ($\times$ 3 for 3 participants), resulting in a total action feature size of 78, including the scale actions.

To facilitate propositional extraction, we decontextualized each utterance from the dialogue context, inspired by the *dense paraphrasing* method (Tu et al., 2022, 2023) that rewrites a textual expression to reduce ambiguity and make explicit the underlying semantics. We filtered the utterances containing at least one pronoun from a predefined set, and had annotators identify the blocks denoted by the pronouns, if any, based on the aligned actions and video frames. Utterances were dually annotated (Cohen's κ = 0.88) and adjudicated by an expert. Utterances were paraphrased by replacing the pronouns with the adjudicated annotation of the block colors (e.g., **they [red block and blue block]** *are probably equal*).

The annotated dataset presents a number of challenges related to sparsity, imbalance, and cross-group diversity. The individual feature channels either capture a single communicative modality or cross-cut two or more (viz. prosodic and CPS features). For input to the model, we concatenate the features of each utterance to the w previous utterances. We remove utterances with no CGA, unless they fall within the w-utterance context window of an utterance with CGA. Of the 1,822 utterances in the dataset, only 271 have any common ground annotation, and these annotations are heavily biased toward the *STATEMENT* class (195, vs. 61 *ACCEPT*s and 15 *DOUBT*s).

6.2. Move Classifier

The move classifier is a multimodal LSTM-based model, intended to capture contextual information that conditions the sequence of cognitive states in a dialogue. Each utterance, including a prior context of $w = 3$ previous utterances, was processed through two linear layers (256 and 512 units) followed by ReLU activation and an LSTM block of 512 units. The final hidden states of the LSTM block for each modality of interest were concatenated and passed through a 512-unit linear layer, $tanh$, another 512-unit linear layer, and SiLU before the classification layer. Fig. 3 shows this architecture.

We optimized for the detection of *STATEMENT*, *ACCEPT*, and *DOUBT*. To alleviate imbalance during training, we augmented the data with SMOTE (Chawla et al., 2002). We trained using Kaiming initilization with a uniform distribution (He et al., 2015). All layers except the classification layer are trained using a triplet loss with a margin of 1 (Balntas et al., 2016) for 200 epochs and a learning rate of 10^{-4}. Subsequently the entire model was trained using cross-entropy loss and a learn-
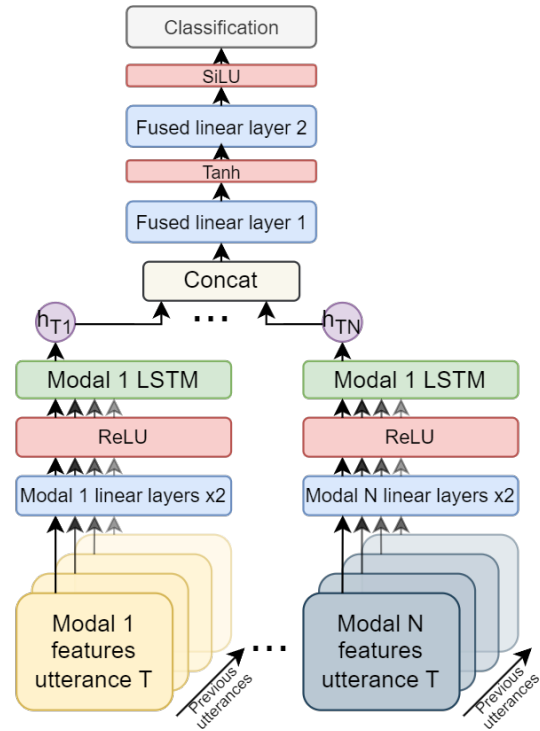

Figure 3: Move classifier architecture.

ing rate of 10^{-3} for 100 epochs, and for 200 further epochs with a learning rate of 10^{-4}. Hyperparameters were fixed using a search with one group held out as validation and one group as test, after which each group was held out in turn while an instance of the model is trained on the other 9 groups, for evaluation on the held-out test group. We ended up with ten trained instances of the archicture, one for each group.

6.3. Propositional Extractor

In additional to the cognitive state expressed by the utterance, we also needed to retrieve the task-relevant propositional content expressed relative to the QUDs. We used two methods for this: 1) **CGA** (*Common Ground Annotation*): We automatically mapped the statement IDs to the propositions expressed as captured in the common ground annotation. Upon move prediction, we consulted this mapping to retrieve the propositional content to be associated with the move in the common ground update. Because annotators had access to the video channel and all other modalities when annotating the propositions expressed, this method is a multimodally-informed method of propositional extraction. 2) **DP** (*Dense Paraphrase*): We encoded the dense paraphrase of the input utterance through BERT (`base-uncased`). Stop words were filtered out before encoding. The stop words came from a standard list, augmented

with words that occured in the transcriptions in 5 or fewer bigrams *and* are not number words, color words, or words describing equality or inequality. BERT vectors were computed by summing over the last 4 encoder layers and taking the average of the `[CLS]` token vector and all individual token vectors in the utterance. Upon move prediction, we chose the proposition whose similarly-encoded BERT vector had the highest cosine similarity to the utterance embedding. Only text and a language model were used in this method, making it unimodal. However, it is important to note that the annotators of the dense paraphrased utterance still had access to the video channel when determining which objects were the denotata of demonstrative pronouns, meaning that is some distant signal from the multimodal data reflected here.

These two methods provide an additional axis of comparison when computing the common ground and allow us to measure the performance gain provided a targeted annotation that directly takes into account all modalities vs. a method using only text and a language model.

6.4. Closure Rules

In order to determine the contents of the CG banks after each utterance, we developed a set of closure rules consistent with the epistemic model presented in Section 4. These rules describe how utterances (specifically, *STATEMENT*s and *ACCEPT*s) affect what is known about the weights of the blocks, and whether certain possibilities are more or less likely than others. When the task begins, the set of possibilities for each block is initialized to $\{10, 20, 30, 40, 50\}$, with no evidence for or against any of those possibilities (i.e., the *evidence_for* and *evidence_against* sets are empty).

Given a *STATEMENT* or *ACCEPT*, we first parsed the propositional content of the utterance into one or more atomic propositions, $p \in At$, where an atomic proposition consists of a block name, a relation ($=, <, >,$ or $\neq$), and a right-hand side. The right-hand side can either be a weight $\in \{10, 20, 30, 40, 50\}$, a block name, or a set of block names connected by $+$. Atomic propositions generally update knowledge about the block on the left-hand side of the relation; the exception is if there is a single block on the right-hand side, and less is known about the right-hand-side block (as measured by the relative sizes of the possibility and evidence sets), in which case that block's possibilities are updated instead.

Then for each atomic proposition, we updated the knowledge associated with the relevant block, according to the move type:

STATEMENTs add evidence for or against certain weights. Statements of propositions "block = weight" directly add that weight to the *evidence_for*

set; e.g., $[!Eq(b, 10)][E]Eq(b, 10)$. For other statements, we compute the set of weights inconsistent with that statement, and add them to the *evidence_against* set.

ACCEPTs remove weights (specifically, those inconsistent with the proposition) from the set of possibilities (and both evidence sets).

Then, from the contents of the possibility and evidence sets for each block, we generated the contents of the CG banks:

If there was only one possibility for the weight of a block, "block = weight" was added to **FBank**; e.g., $[E]Eq(b, 10) \rightarrow [!Eq(b, 10)][B]Eq(b, 10)$.

Otherwise, for those weights in the block's *evidence_for* set, we added "block = weight" to **EBank**. Similarly, for those weights in the block's *evidence_against* set, we added "block $\neq$ weight" to EBank. Inequalities for which evidence existed also were added to EBank.

For the remaining weights (not in either evidence set) in the set of possibilities for the block, we added "block = weight?" to **QBank**.

Considering the example in Sec. 3.1, since the group already knows that $red = 10$ at that point, once $red = blue$ is accepted as a fact, the closure rules also elevate $blue = 10$ to the same epistemic status as $red = 10$. All other possibilities for blue block's weight are also removed from both evidence sets.

The ground truth contents of the CG banks were computed by running the closure rules directly over the annotated data.

7. Results

Averaged across all groups, the move classification model achieves a weighted F1 of 0.61. Most misclassifications are confusions of *STATEMENT*s and *ACCEPT*s, which affect the level of evidence assigned to extracted propositions but not the propositions themselves.

Table 1 shows average DSC per group, for each bank, with propositional extraction using the Common Ground Annotation (CGA) method (Sec. 6.3). We also assess the union of the fact bank and evidence bank, to assess how different modalities contribute to the extraction of propositional content and elevation to either status. We compare the performance using all modalities to using language features only, in the form of BERT vectors.

We find that in most cases, our common ground tracker has trouble not with retrieving the right propositions with the multimodal CGA method, but with assigning the right level of evidence. This is seen in the values for the union of FBank and EBank, which remain high across all groups, even when the Sørensen-Dice coefficients of the individual FBank or EBank are comparatively lower. This also tracks the misclassifications made by

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
					All modalities					
QBank	0.777	0.663	0.811	0.841	0.575	0.868	0.845	0.834	0.987	0.551
EBank	0.250	0.574	0.709	0.926	0.391	0.734	0.793	0.063	0.985	0.250
FBank	0.425	0.480	0.418	0.348	0.318	0.315	0.637	0.574	0.000	0.794
F ∪ E	1.000	0.864	0.939	0.866	0.875	0.880	1.000	0.600	0.996	0.903
					Language only					
QBank	0.767	0.911	0.829	0.817	0.514	0.868	0.972	0.834	0.987	0.392
EBank	0.344	0.713	0.712	0.812	0.335	0.691	0.904	0.049	0.985	0.262
FBank	0.000	0.528	0.501	0.045	0.165	0.372	0.825	0.526	0.000	0.000
F ∪ E	1.000	0.922	0.925	0.832	0.959	0.799	0.967	0.585	0.996	0.827

Table 1: Average DSC per group over all CG banks, comparing multimodal features and language only features. Propositions are extracted using the CGA method.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
All	1.000	0.864	0.939	0.866	0.875	0.880	1.000	0.600	0.996	0.903
BERT	1.000	0.922	0.925	0.832	0.959	0.799	0.967	0.585	0.996	0.827
openSMILE	1.000	0.922	0.900	0.832	0.880	0.839	1.000	0.947	0.996	0.827
CPS	1.000	0.922	0.900	0.832	0.880	0.815	0.000	0.947	0.996	0.827
Action	1.000	0.922	0.900	0.832	0.880	0.873	0.571	0.947	0.996	0.827
GAMR	1.000	0.922	0.900	0.832	0.880	0.731	0.658	0.947	0.996	0.827

Table 2: Average DSC per group over FBank ∪ EBank, comparing multimodal features and each individual modality. Propositions are extracted using the CGA method.
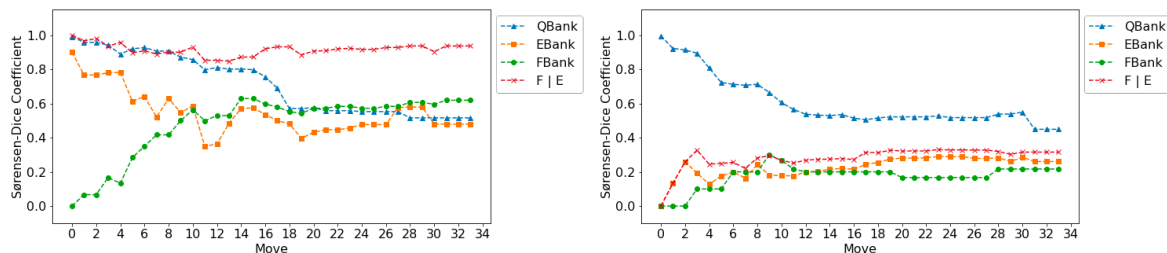

Figure 4: DSC for each bank aggregated across groups, plotted vs. utterance, using all modalities in the move classifier. [L]: propositional extraction performed using the multimodal CGA method. [R]: propositional extraction performed using the language-only Dense Paraphrase (DP) method.

the move classifier, as an *ACCEPT* will elevate a proposition to a fact, while a *STATEMENT* will keep it in evidence without removing the corresponding QUD from QBank.

Incorprating multiple modalities into the move classifier model usually helps assign propositions to the correct level of common ground and maintain greater overlap in the retrieved QUDs relative to ground truth. However, there is great variety across groups. For example, Group 2 performs better using only language, while in Group 9 non-linguistic features do not change the result. These differences can be attributed to how different groups use different modes of communication to complete the task (see Sec. 8).

Table 2 shows average DSC per group over FBank ∪ EBank, comparing each individual modality vs. all modalities.

We see here that often, each individual modality performs similarly or identically, but in 4 out of 10 groups, using all multimodal features results in the highest performance. However, in other groups,

multimodal features may make no difference, or some other individual feature type is the strongest predictor of performance. Compare Group 1 and Group 9, where all modalities perform identically, with Groups 6 and 7, where they all perform differently but multimodal features perform the highest. This supports the previous observation: different groups may adopt radically different modal combinations to communicate equivalent information.

Fig. 4 shows the progression of DSC over time, aggregated across all groups, using all modalities in the move classifier, but comparing and contrasting the multimodal Common Ground Annotation (CGA) and language model-based Dense Paraphrase (DP) methods for proposition extraction. Including multimodal information improves the retrieval of the correct propositions independent of the level of evidence or factuality assigned to them—as shown by the consistently high DSC of FBank ∪ EBank in the left plot.

3594

8. Discussion

Some specific examples show how the language model-based DP method struggles to extract propositions from complex utterances. Table 3 shows how vector comparison over only linguistic information tends to struggle with propositions involving multiple objects. Certain groups, like Group 1, tended to speak full propositions aloud, while others, like Group 10, mixed modalities ("ten and ten" is accompanied by gesture and action, which are accounted for directly by Common Ground Annotations but not Dense Paraphrases).

Group	Utterance (DP)	Proposition (Correct?)
1	red block's ten so then	red = 10 (✓)
1	yeah ok so now we know that blue block is also ten	blue = 10 (✓)
5	so red block, blue block are both ten in theory ten ten twenty	red = 20 and green = 40 and purple = 10 (✗)
5	so the green we think is twenty ok so let's see we can use our hands as well	green = 20 (✓)
10	i guess green block is like twenty and red block, blue block is like ten and ten	red = 50 and green = 20 and purple = 10 (✗)

Table 3: Utterances and propositions retrieved using DP method.

This reflects Table 2 where, even using the multimodal CGA extraction method, Group 1 achieves perfect overlap of FBank ∪ EBank with ground truth using just language, while the model has to combine modalities to reach its best performance for Group 10. That Group 1 performance over FBank ∪ EBank is also perfect using each individual modality alone suggests that their utterances are strongly aligned with their non-verbal behavior. Meanwhile, Group 6 stands out as a particular case where each individual modality is contributing something distinct.

Misclassifications of *STATEMENT*s as *ACCEPT*s, or vice versa, may elevate the utterances of certain participants to fact status, or leave elements in the Questions Under Discussion when they have already been resolved. This could also lead to a certain participant having more apparent influence over the dialogue. One participant's beliefs may update the common ground of the group, and leave other participants' beliefs unconsidered. Table 4 shows some examples from Group 10, and demonstrates how affirmative language like "yeah" may be indicative of *ACCEPT*s elsewhere in the training data, while "okay" or restatements of propositional content are typically indicative of *STATEMENT*s even when in context they indicate acceptance of a previously-stated proposition.

Timestamp	Utterance	Label	Prediction
117.46-118.87	yeah they're together.	*STATEMENT*	*ACCEPT*
217.89-219.78	thirty one thirty two so thirty	*ACCEPT*	*STATEMENT*
218.23-219.00	so okay	*ACCEPT*	*STATEMENT*

Table 4: Sample of utterances from Group 10 misclassified by move classifier.

9. Conclusion and Future Work

In this paper we have presented a challenging novel task: *multimodal common ground tracking*, and a novel benchmark over the challenging Weights Task Dataset. We presented a formal model of common ground over a shared task and augmented the WTD with additional gesture, action, and common ground annotations. We performed a set of experiments to evaluate the contributions of different modalities toward modeling the cognitive states of the group, extracting the propositions expressed, and building common ground structures as the group proceeds through the task. Our model will be particularly useful for AI systems deployed in environments such as classrooms, where they can track the collective knowledge of a group and facilitate productive collaborations.

Certain modalities may be more prone to misclassifications based on the speaker. For instance, future work could examine how prosidic features could be used to detect power dynamics that may bias the construction of common ground toward certain people or assertions. Giving the common ground model additional separate banks for each speaker would allow an agent to facilitate knowledge sharing and collaboration if it seems like a subgroup has arrived at a belief not shared by the whole group. In a task-based environment, the agent could use the model of common ground to make task-relevant inferences itself, such as the algebraic relationship between the block weights here, allowing it to learn from watching and interacting with the group. Finally, because there is a one-to-many mapping between propositions and potential ways to phrase or express them in utterances, the dense paraphrase method for propositional extraction could benefit from a cross-encoder approach, as used in coreference research.

Limitations

Although our work addresses a novel and challenging problem, scaling the pipeline to other use cases confronts some (surmountable) limitations. For a given task, the relevant propositions that may populate the common ground need to be determined and enumerated. The number of propositions scales naturally to increased cardinality of items, attributes, and relations within a similar domain (e.g., by computing the Cartesian product of items, attributes, and binary relations, and subsequently the powerset of atomic propositions to account for conjunctions like $red = 10 \land blue = 10$). Therefore the complexity of proposition construction is subject to the complexity of the task and number of task items. Enumerating the closure rules is straightforward once the propositions are

determined. The move classifier itself should require no changes unless there is a change in input modalities. Imbalance within the data categories presents a further challenge that needs to be addressed. In this paper we used data augmentation approaches like SMOTE, but precise handling would need to be determined on a task-specific basis.

Acknowledgements

Bibliographical References

Nicholas Asher. 1998. Common ground, corrections and coordination. *Journal of Semantics*.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3.

Alexandru Baltag, Lawrence S Moss, and Sławomir Solecki. 2016. *The logic of public announcements, common knowledge, and private suspicions*. Springer.

Leonardo Bonacci. 1202. Liber Abaci. Unpublished lost manuscript.

Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2023. Automatic detection of collaborative states in small groups using multimodal features. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education*.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with elan. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT press.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *AAAI Conference on Artificial Intelligence*.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D, editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association.

Marco Del Tredici, Xiaoyu Shen, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2022. From rewriting to remembering: Common ground for conversational qa models. *arXiv preprint arXiv:2204.03930*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. 2021. Cutting melted butter? common ground inconsistencies management in dialogue systems using graph databases. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):157–190.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Pierre Dillenbourg and David Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151.

Eleni Efthimiou and Georgios Kouroupetroglou. 2011. *Gestures in Embodied Communication and Human–Computer Interaction*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Kerstin Fischer. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine*, 32(4):31–38.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. In *International Conference on Universal Access in Human-Computer Interaction*, pages 828–837. Springer.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.

Jonathan Ginzburg. 1996. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory. Blackwell, Oxford*, pages 359–423.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking. In *Proc. Interspeech 2019*, pages 1458–1462.

Susan Goldin-Meadow. 2003. *Hearing Gesture: How Our Hands Help Us Think*, volume 14.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions.

Lauren V. Hadley, Graham Naylor, and Antonia F. de C. Hamilton. 2022. A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1(1):42–54. Number: 1 Publisher: Nature Publishing Group.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2022. A simple language model for task-oriented dialogue.

Léo Jacqmin, Lina M Rojas-Barahona, and Benoit Favre. 2022. ” do you follow me?”: A survey of recent approaches in dialogue state tracking. *arXiv preprint arXiv:2207.14627*.

Adam Kendon. 1997. Gesture. *Annual Review of Anthropology*, 26(1):109–128.

Adam Kendon. 2004. Gesture: Visible action as utterance.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. 2023. The Weights Task Dataset: A Multimodal Dataset of Collaboration in a Situated Task.

Anthony Pak-Hin Kong, Sam Po Law, Connie Kwan, Christy Lai, and Vivian Lam. 2015. A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage). *Journal of Nonverbal Behavior*, 39.

Stefan Kopp and Ipke Wachsmuth. 2010. *Gesture in embodied communication and human-computer interaction*, volume 5934. Springer.

Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51.

Nikhil Krishnaswamy and James Pustejovsky. 2020. A formal analysis of multimodal referring strategies under common ground. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5919–5927.

Geert-Jan M Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, Hendrik Zender, Ivana Kruijff-Korbayová, and Nick Hawes. 2010. Situated dialogue processing for human-robot interaction. In *Cognitive systems*, pages 311–364. Springer.

Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The ava-kinetics localized human actions video dataset.

Xinmeng Li, Qian Li, Wansen Wu, and Quanjun Yin. 2021. Generation and extraction combined dialogue state tracking with hierarchical ontology integration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2249, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.

Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021. Knowledge-aware graph-enhanced GPT-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Magdalena Markowska, Adil Soubki, Gary Mar, Seyed Abolghasem Mirroshandel, Owen Rambow, and Anita Wasilewska. Formal representation of common ground in dialogue.

Paul Marshall and Eva Hornecker. 2013. Theories of embodiment in hci. *The SAGE handbook of digital technology research*, 1:144–158.

David Mcneill. 2005. *Gesture and Thought*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Xenia Ohmer, Marko Duda, and Elia Bruni. 2022. Emergence of hierarchical reference systems in multi-agent communication. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5689–5706.

Eric Pacuit. 2017. *Neighborhood semantics for modal logic*. Springer.

Jan Plaza. 1989. Logics of public communications. In *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. *Proceedings of LREC*.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327.

Patrick Rohrer, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve-Gibert, Ada Ren, Stefanie Shattuck-Hufnagel, and Pilar Prieto. 2020. The multimodal multidimensional (m3d) labelling scheme for the annotation of audiovisual corpora.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding.

Stefan Schaffer and Norbert Reithinger. 2019. Conversation is multimodal: thus conversational user interfaces should be as well. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3.

Matthias Scheutz, Rehj Cantrell, and Paul Schermerhorn. 2011. Toward humanlike task-based dialogue processing for human robot interaction. *Ai Magazine*, 32(4):77–84.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

LE Sigler. 2002. *Fibonacci's Liber Abaci. Leonardo Pisano's Book of Calculation*. New York: Springer.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5-6):701–721.

Angela E. B. Stewart, Zachary Keirn, and Sidney K. D'Mello. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 31(4):713–751.

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

Christopher Tam, Richard Brutti, Kenneth Lai, and James Pustejovsky. 2023. Annotating situated actions in dialogue. In *Proceedings of the 4th International Workshop on Designing Meaning Representation*.

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? In *Workshop on Interoperable Semantic Annotation (ISA-19)*, page 75.

David Traum. 1994. A computational theory of grounding in natural language conversation. *PhD thesis, University of Rochester*.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. *Current and new directions in discourse and dialogue*, pages 325–353.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS)*, Nancy, France. Association for Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, and James Pustejovsky. 2022. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1521–1533, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Johan van Benthem, David Fernández-Duque, and Eric Pacuit. 2014. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133.

Johan van Benthem and Eric Pacuit. 2011. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92.

Wolfgang Wahlster. 2006. Dialogue systems go multimodal: The smartkom experience. In *SmartKom: foundations of multimodal dialogue systems*, pages 3–27. Springer.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020a. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020b. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

A. Group-wise Move Classifier Performance

Table 5 shows the performance of the 10 classifiers with each being trained on 9 different groups, and evaluated on the remaining 10th.

B. Annotation Procedures and IAA

ELAN (Brugman and Russel, 2004) was the tool used for most annotation, supplemented by collating annotations in spreadsheets. As we can see in Fig. 5, this tool allows annotators to visualize the data at any point in the videos, and also see all other annotated modalities. The data is then featurized and used as input for the move classifier.

Tables 6–8 show inter-annotator agreement (IAA) metrics for the Common Ground, action, and GAMR annotation per group. Because gestures are individualized, GAMR annotations are also broken down by participant. Means are also provided.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
Accuracy	0.625	0.720	0.750	0.500	0.735	0.690	0.429	0.556	0.833	0.438
Micro F1	0.625	0.720	0.750	0.500	0.735	0.690	0.429	0.556	0.833	0.438
Macro F1	0.564	0.360	0.621	0.235	0.512	0.375	0.271	0.389	0.652	0.203
Weighted F1	0.605	0.680	0.768	0.412	0.706	0.737	0.453	0.509	0.791	0.342
Micro Precision	0.625	0.720	0.750	0.500	0.735	0.690	0.429	0.556	0.833	0.438
Macro Precision	0.583	0.369	0.611	0.200	0.525	0.382	0.288	0.355	0.912	0.167
Weighted Precision	0.604	0.654	0.796	0.350	0.687	0.837	0.526	0.473	0.863	0.281
Micro Recall	0.625	0.720	0.750	0.500	0.735	0.690	0.429	0.556	0.833	0.438
Macro Recall	0.567	0.365	0.650	0.286	0.516	0.462	0.298	0.433	0.625	0.259
Weighted Recall	0.625	0.720	0.750	0.500	0.735	0.690	0.429	0.556	0.833	0.438
AUROC	0.500	0.539	0.500	0.563	0.500	0.501	0.476	0.669	0.500	0.531

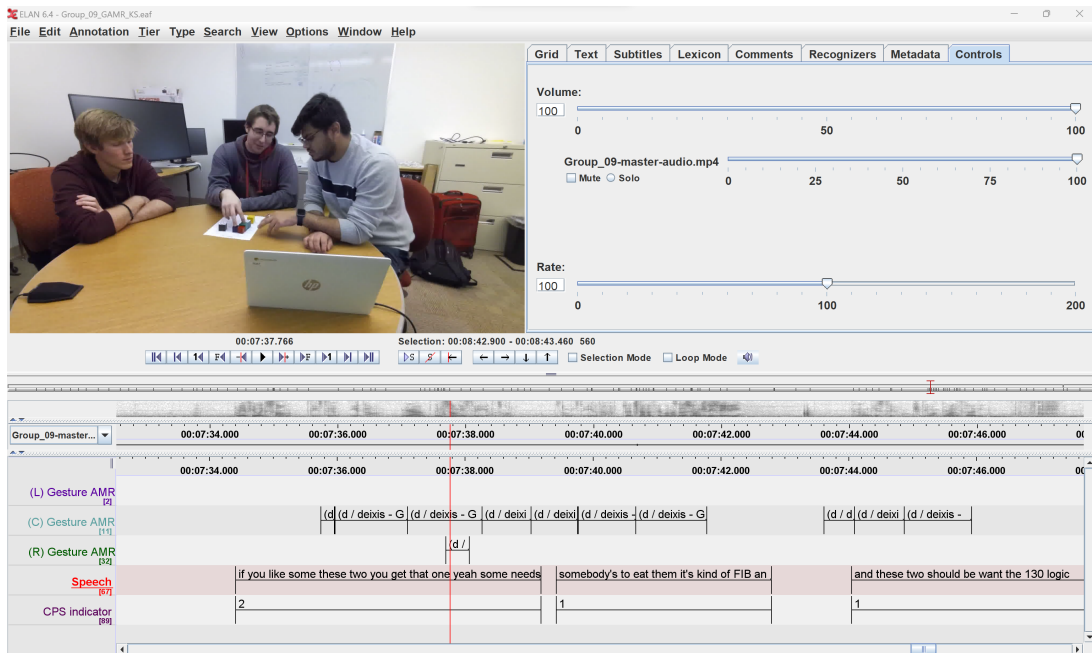Table 5: Group-wise performance of the move classifier using hold-one-group-out evaulation method.

Figure 5: Still of annotation procedure using ELAN.

Group	F1 score	Cohen's κ
1	0.520	0.359
2	0.454	0.295
3	0.492	0.356
4	0.411	0.267
5	0.471	0.503
6	0.639	0.603
7	0.678	0.572
8	0.522	0.712
9	0.575	0.564
10	0.645	0.772
mean	0.541	0.500

Table 6: IAA on Common Ground Annotations.

Group	F1 score	Cohen's κ
1	0.557	0.464
2	0.651	0.666
3	0.750	0.688
4	0.719	0.654
5	0.804	0.689
6	0.737	0.798
7	0.761	0.660
8	0.583	0.466
9	0.519	0.432
10	0.629	0.458
mean	0.671	0.597

Table 7: IAA on action annotations.

Group	Participant	F1	Precision	Recall
1	1	0.921	0.953	0.890
1	2	0.943	0.917	0.971
1	3	0.899	0.912	0.886
1	μ	0.921	0.927	0.915
2	1	0.846	0.798	0.902
2	2	0.947	0.938	0.957
2	3	0.895	0.850	0.944
2	μ	0.896	0.862	0.934
3	1	0.686	0.720	0.656
3	2	0.809	0.796	0.824
3	3	0.793	0.775	0.811
3	μ	0.763	0.763	0.763
4	1	0.791	0.837	0.750
4	2	0.658	0.807	0.556
4	3	0.817	0.779	0.859
4	μ	0.755	0.808	0.722
5	1	0.824	0.836	0.813
5	2	0.693	0.642	0.754
5	3	0.835	0.853	0.817
5	μ	0.784	0.777	0.795
6	1	0.697	0.704	0.691
6	2	0.628	0.667	0.594
6	3	0.480	0.462	0.500
6	μ	0.602	0.611	0.595
7	1	0.865	0.874	0.857
7	2	0.736	0.724	0.748
7	3	0.667	0.662	0.671
7	μ	0.756	0.753	0.759
8	1	0.782	0.725	0.850
8	2	0.745	0.710	0.784
8	3	0.907	0.925	0.891
8	μ	0.812	0.787	0.841
9	1	0.846	0.846	0.846
9	2	0.600	0.525	0.700
9	3	0.800	0.818	0.783
9	μ	0.749	0.730	0.776
10	1	0.386	0.810	0.254
10	2	0.487	0.631	0.396
10	3	0.584	0.444	0.851
10	μ	0.749	0.730	0.776
μ	1	0.765	0.754	0.806
μ	2	0.738	0.712	0.752
μ	3	0.768	0.789	0.761
μ	μ	0.752	0.752	0.773

Table 8: IAA on Gesture-AMR (GAMR) annotation.