

Common Ground Tracking in Multimodal Dialogue

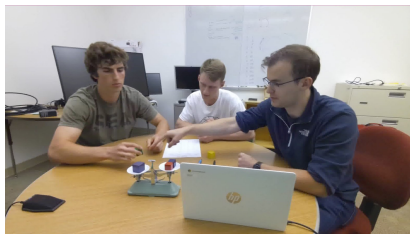
Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky



LREC-COLING 2024
Turin, Italy

Introduction

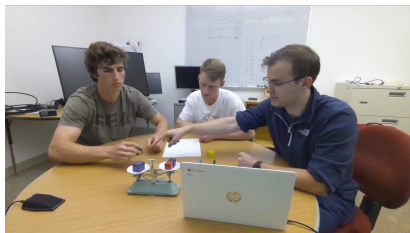
- Major challenge for computational models of multimodal interactions: tracking intentions, goals, and attitudes of the participants
- *Common Ground Tracking*: identifies the shared belief space of all participants in a task-oriented dialogue



Introduction

Our work:

- A challenging new task: multimodal common ground tracking
- Incorporation of a formal model into automated tracking of group common ground over time



Common Ground in Dialogue

- We adapt Ginzburg's **Dialogue Game Board**
- Our innovation: Common Ground Structure, *cgs*, consists of:
 - 1 Questions Under Discussion (Q_{BANK}): set of topics or unknowns that need to be answered to solve the task;
 - 2 Evidence (E_{BANK}): set of propositions for which there is some evidence they are true
 - 3 Facts (F_{BANK}): set of propositions believed as true by all participants

Evidence-based Belief

- Let $P = \{p_1, p_2, p_3\}$ be participants
- From situation s_k , let move $m_i = (p_j, C_j, s_{k+1})$: p_j 's communicative act C_j moves the dialogue to situation s_{k+1}
- Simplified model of evidence-based Dynamic Epistemic Logic (EB-DEL)
- $\mathcal{M} = (W, E, V)$, where
 - W is a non-empty set of worlds
 - $E \subseteq W \times \wp(W)$ is an evidence relation
 - $V: At \rightarrow \wp(W)$, is a valuation function
- Let $E(w)$ denote world accessible to w through evidencing relation E
 - World w may contain evidence for ϕ , $[E]\phi$

Common Ground Updating

- $[!\phi][E]\psi$: Given the announcement of ϕ , there is evidence for ψ
 - Moves a proposition ψ from QBANK to EBANK
- $[E]\psi \rightarrow [!\phi][B]\psi$: Given the announcement of ϕ , there is *belief* for ψ conditional upon prior evidence for ψ
 - Moves a proposition ψ from EBANK to FBANK

Dataset

- The Weights Task: 10 triads deduce the weights of blocks using a balance scale
- Multimodal communication: language, gesture (e.g., pointing), and action (e.g., picking up, placing)
- Knowledge is shared with multiple communicative channels



Figure: Participant 3 [R] says “looks like they’re fairly equal” after placing the red and blue blocks on different sides of the scale.

Annotation

- Weights Task Dataset (WTD) contains transcribed speech, prosodic features, collaborative problem solving indicators, Gesture-AMR (GAMR)
- Augmented data with dual annotation of GAMR, action annotations with VoxML, and “common ground annotations” (CGA)
- GAMR, action, and CGA were dually-annotated, adjudicated by an expert
 - GAMR: SMATCH-F1 .75
 - Action: F1 .67
 - CGA: F1 .54

Common Ground Annotation

- CGA identifies the cognitive state of participants concerning the task:
 - *OBSERVATION*: participant P_i has perceived an action a
 - *INFERENCE*: deduction from ϕ
 - **STATEMENT**: announcement of evidence for ϕ
 - *QUESTION*: interrogative relating to ϕ
 - *ANSWER*: answering question about ϕ
 - **ACCEPT**: agree with ϕ
 - **DOUBT**: disagree with ϕ
- This work focuses on the detection of *STATEMENTS* ($n = 195$), *ACCEPTS* ($n = 61$), and *DOUBTS* ($n = 15$).

Common Ground Annotation



- P3: “looks like they’re fairly equal” \rightarrow $STATEMENT(red = blue)$
- P1: “yeah” \rightarrow $ACCEPT(red = blue)$
- P3: “that’s 20, these two [red, blue] are 10” \rightarrow $STATEMENT(red = 10 \wedge blue = 10)$
- P1: “wait, let’s see” \rightarrow $DOUBT(red = 10 \wedge blue = 10)$

Result: $red = blue$ is a *fact* (F_{BANK}), $red = 10 \wedge blue = 10$ is only *evidenced* (E_{BANK})

Experiments

3 components:

- Move classifier, predicts cognitive state expressed in an utterance
- Propositional extractor, extracts what is being expressed (consults annotation or performs inference)
- Closure rules, unify the cognitive state and propositional content and update QBANK, EBANK, and FBANK

Primary metric: Sørensen-Dice coefficient (DSC)

Move Classifier

- Multimodal LSTM-based classifier
- Input utterances with context window of 3
- Data augmentation with SMOTE
- Training with triplet and cross-entropy losses
- Hyperparameters tuned with one group held out as validation and one group as test
- Tuned model trained with 9 groups, evaluated on 1 in turn

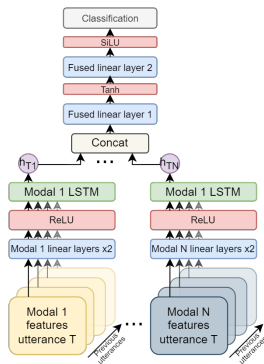


Figure: Move classifier architecture.

Propositional Extractor

2 methods:

- 1 Common Ground Annotation (CGA): automatically mapped utterance IDs to propositions captured in the common ground annotation
 - Implicitly multimodal method as annotators had access to all modal channels
- 2 Dense Paraphrase (DP): Paraphrased utterances by replacing demonstrative pronouns with explicit referents (e.g., “they [red block and blue block] are probably equal”)
 - Utterances were encoded through BERT and compared to embeddings of candidate propositions using cosine similarity

Results

	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 U 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 U E	1.000	0.922	0.925	0.832	0.959	0.799	0.967	0.585	0.996	0.827

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

Results

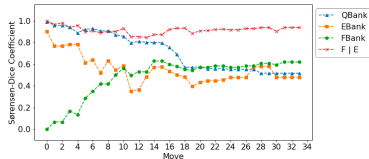


Figure: Aggregate DSC across groups, using all modalities and multimodal CGA propositional extraction.

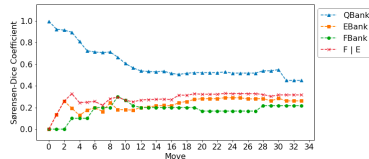


Figure: Aggregate DSC across groups, using all modalities and language-only DP propositional extraction.

- Incorporating multiple modalities helps assign propositions to the correct level of evidence
- Wide variation across groups

Discussion

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: Utterances and propositions retrieved using DP method.

- Vector comparison over language embeddings tends to struggle with propositions involving multiple objects
- Certain groups tended to speak full propositions aloud, while others mixed modalities

Conclusion

- We 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
- We evaluated the contributions of different modalities toward modeling cognitive states, extracting propositions expressed, and building common ground structures as the group proceeds through the task

Future Work

- Our model will be particularly useful for AI systems deployed in environments such as classrooms
 - Track the collective knowledge of a group and facilitate productive collaborations
- Examine how prosodic features can detect power dynamics
- Agent could use the model of common ground to make task-relevant inferences
- Cross-encoder for propositional extraction (to appear at EDM 2024!)

