MultiCAT: Multimodal Communication Annotations for Teams

Adarsh Pyarelal[1], John Culnan[1], Ayesha Qamar[2], Meghavarshini Krishnaswamy[1], Yuwei Wang[1], Cheonkam Jeong[1], Chen Chen[1], Md Messal Monem Miah[2], Shahriar Hormozi[1], Jonathan Tong[2], Ruihong Huang[2]

[1] The University of Arizona [2] Texas A&M University

Computational Language Understanding

Machine Learning for Artificial Intelligence

NAACL 2025

Download and use our dataset!

Web: https://multicat.lab.pyarelal.xyz

Email: adarsh@arizona.edu

Long-term goal

Can we build AI teammates that intervene on team communication to improve team performance?

Contributions

(1) Novel multiparty spoken task-based dialog **dataset** with annotations for related paralinguistic and conversational classification tasks.

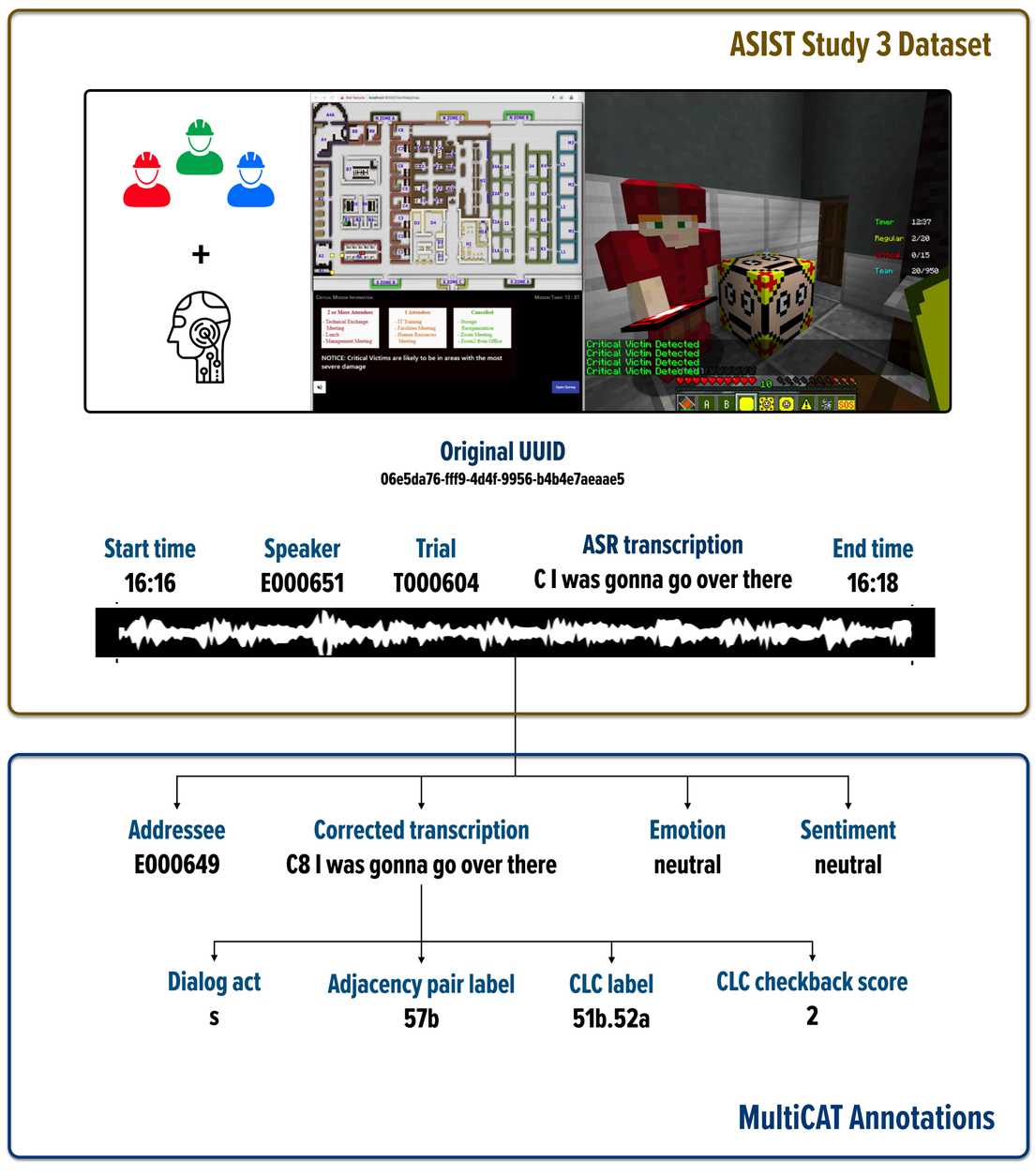
(2) Baseline results for labeling dialog acts, adjacency pairs, sentiment, emotion, and closed-loop communication (CLC) events.

(3) Exploratory analyses relating our annotations to each other and to team outcomes, with results suggesting that our tasks are interrelated, and that **our annotations may be better predictors of team performance than participants' self-reported task proficiency and expertise.**

Dataset

Background and Construction

MultiCAT builds upon the **ASIST Study 3 dataset**, which contains spoken dialog in teams of 3 humans conducting virtual search-and-rescue missions in Minecraft.



Selected statistics

Quantity	Total
Trials	49
Teams	25
Speakers	73
Utterances	11024
Word types	2607
Word tokens	108475

(a) Totals

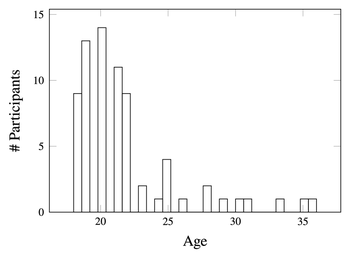
	Mean	Min	Max	SD
Utts./spkr	151	42	287	54
Utts./trial	225	91	348	65
Utts./spkr/trial	79	19	156	28
Word types/utt.	9	1	74	8
Word tokens/utt.	10	1	118	11

(b) Mean, minimum, maximum, and SD.

Annotation	# Trials	# Utts
Emotion	46	7731
Sentiment	46	7731
CLC	36	6544
Gold transcript	45	4666
Dialog act	45	10342
APs	45	6846
Entrainment	8	2896

(c) Number of trials and annotated utterances for our annotation types.

Selected demographic details



Sex	Count
Male	56
Female	12
Nonbinary	1
Prefer not to respond	2

Ethnicity	Count
White/Caucasian	41
Asian	13
Hispanic or Latino	8
Middle Eastern	1
White/Caucasian, Hispanic or Latino	5
Hispanic or Latino, Other	2
White/Caucasian, Asian	2
White/Caucasian, Asian, Pacific Islander	1
White/Caucasian, Hispanic or Latino, Asian	1

Baselines

Model	Fine-grained Macro F_1 (SEM)	Fine-grained Accuracy (%)	Coarse-grained Macro F_1 (SEM)	Coarse-grained Accuracy
He et al.'s (2021)	30.75	63.24	42.15	93.92
LLaMA-3	34.76 (0.48)	66.47 (0.15)	44.55 (0.90)	94.66 (0.07)

Table 3: Macro F_1 and accuracy for DA classification on fine-grained and coarse-grained classes.

		Models		
Emotion	Support	Strat.	Multi.	LLaMA-3
Anger	18	5.4	0.0	3.9 (0.03)
Disgust	25	0.0	9.3	15.8 (0.0)
Fear	70	3.2	16.2	27.2 (0.02)
Joy	154	4.2	20.1	19.6 (0.01)
Neutral	1799	77.5	76.5	87.7 (0.0)
Sadness	145	5.6	30.5	36.7 (0.01)
Surprise	80	3.7	29.2	31.6 (0.02)
All	2291	14.2	26.0	31.8 (0.92)

Table 6: Results for emotion prediction.

		Models		
Sentiment	Support	Strat.	Multi.	LLaMA-3
Negative	370	15.0	43.5	52.1 (0.0)
Neutral	1310	51.5	62.7	68.8 (0.0)
Positive	611	28.0	49.8	54.0 (0.0)
All	2291	31.5	52.0	58.4 (0.24)

Table 5: Results for sentiment prediction.

Stage	Accuracy	F_1
Call-out detection	.77	.79
Check-back detection	.76	.43
Complete CLC event detection	.51	.45

CLC Detection

Exploratory analyses

Relation between tasks

	DA	AP	CLC	Sentiment	Emotion
DA	–	N/A	✓	✓	N/A
AP		–	✓	✓	✓
CLC			–	✓	✓
Sentiment				–	✓
Emotion					–

Table 7: Combinations of label types for which the p-values computed using a χ^2 test of independence is less than 0.000125 (indicated by ✓). The matrix is symmetric, hence we omit the entries below the diagonal. We enter 'N/A' in the cells corresponding to the DA/AP and DA/Emotion combinations, as they do not satisfy the rules-of-thumb for χ^2 tests of independence discussed by Kroonenberg and Verbeek (2018).

Team performance prediction

	Mission 1	Mission 2	Combined
# of trials	17	16	33
Proficiency	130 (26)	104 (19)	118 (17)
AP	126 (17)	100 (13)	118 (12)
CLC	125 (13)	99 (11)	116 (9)
DA	125 (11)	99 (9)	117 (8)
Sent	125 (10)	100 (8)	116 (7)
Emo	123 (9)	98 (7)	115 (6)
Multicat	123 (8)	97 (7)	115 (6)
All	122 (8)	97 (6)	115 (5)

Table 8: MAE (with SEM in parentheses) over all folds for our score prediction models.

Notable Firsts

(1) First publicly available dataset for CLC detection.

(2) First dataset for remote multi-party vocal entrainment detection.

Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under awards W911NF- 20-1-0002 and W911NF-24-2-0034.

DARPA

DEVCOM Army Research Laboratory