CAMAL: A Novel Dataset for Multi-label Conversational Argument Move Analysis

Viet Dac Lai, Duy Pham
Jonathan Steinberg, Jamie Mikeska, Thien Huu Nguyen
vietl@uoregon.edu

LREC-COLING 2024

Introduction

- Pre-Service Teacher (PST) is an individual enrolled in a **teacher preparation program** who must earn a degree/certificate to work as a teacher.
- **Simulated classrooms** are often used as practice spaces to **support teacher learning**.
- Scoring of PST are done **manually** by *human raters/coaches/teacher educators.*
- 1-1 scoring/feedback is **extremely expensive**.
- An **automatic analysis system** of teacher-student discussion can provide **much more (frequent) feedbacks** at a much lower cost.
- However, there is no dataset for studying **Conversational Argument Move AnaLysis**.

Takeaways

- Study the **Conversational Argument Move AnaLysis** (**CAMAL**) in educational assessment.
- Created the first dataset for **CAMAL**:
  - **Two science tasks**
  - Size: **44 hours**
  - Environment: **Simulated classrooms**
- Proposed **Speaker Identification Graph** to improve model's performance
- Applications:
  - Pre-Service Teacher (PST) assessment
  - PST Discussion Analysis
  - PST Automatic Scoring

Data Collection

- Discussion tasks:
  - Mystery Powder (Chemistry): discussion how to identify an unknown powder
  - Ordering Fractions (Math): discussion how to order a list of fractions
- Simulated classroom environment:
  - A subject PST
  - A expert acts as 4 higher elementary students.
- All the audios are transcribed into texts
- Total:
  - **165 sessions**
  - **44 hours**
  - **16 minutes/session**

Taxonomy

Claim:

- ELC: Eliciting A Claim
- STC: Stating A Claim

Data:

- ELD: Eliciting Data
- PVD: Providing Data

Reasoning & Justification:

- ELR: Eliciting Reasoning & Justification
- PVR: Providing Reasoning & Justification

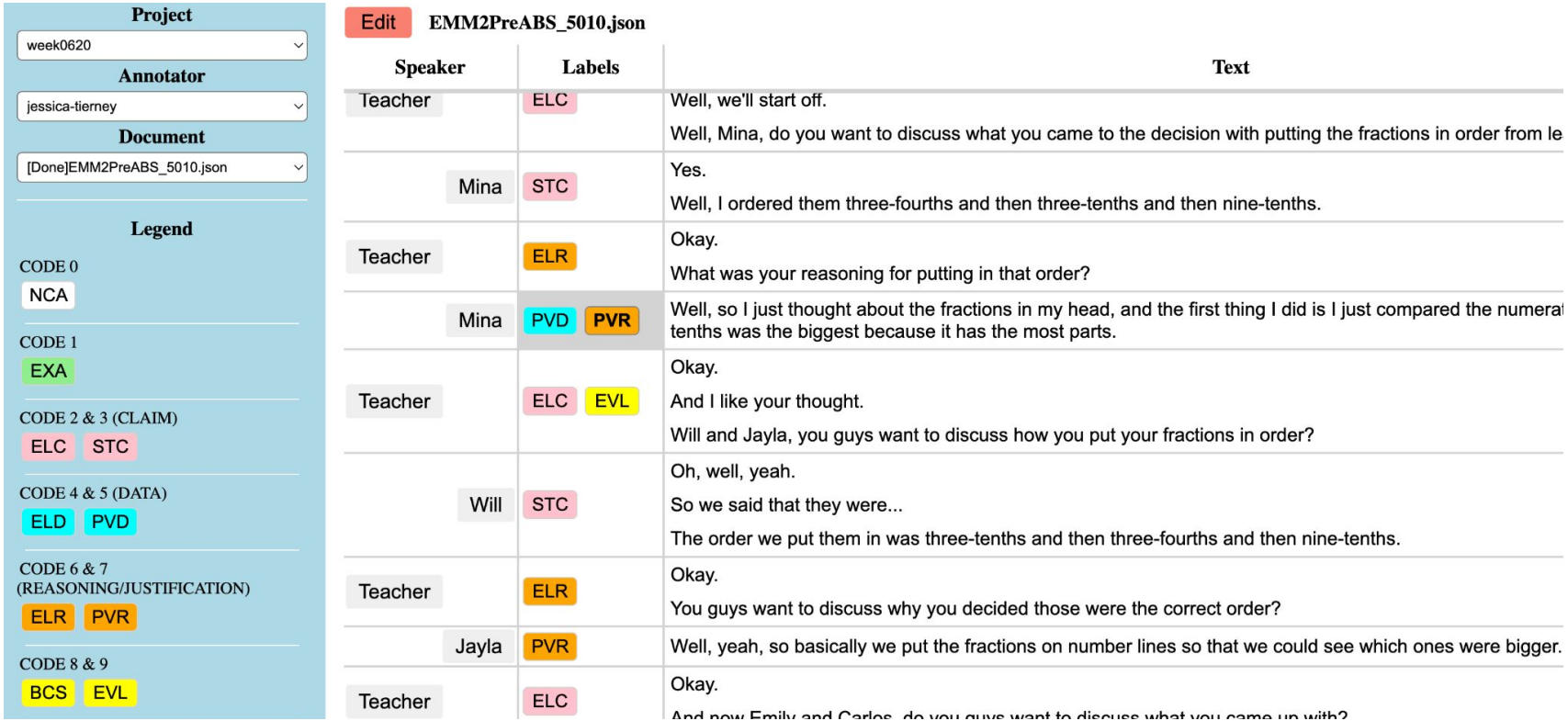
Misc:

- EXA: Explicating Argumentation
- BCS: Building Consensus
- EVL: Evaluating

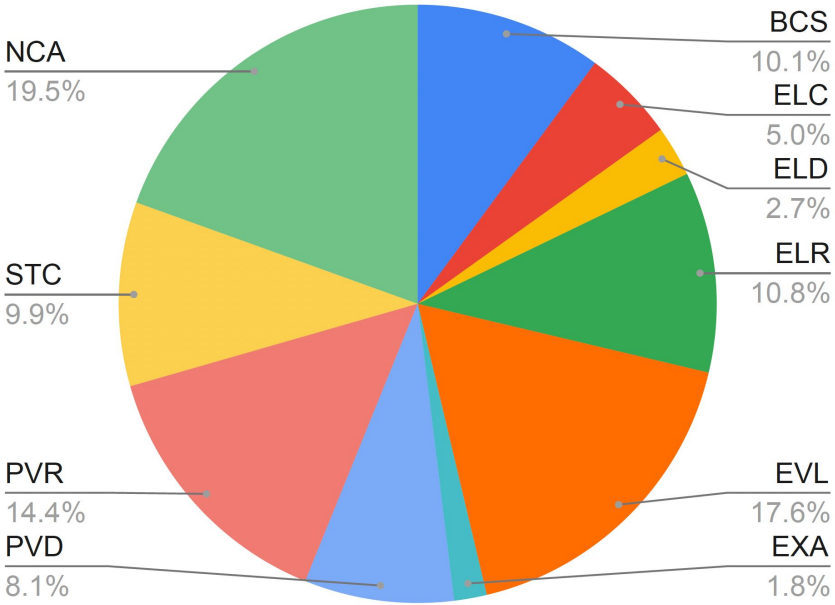
Null:

- NCA: No Code Applied

Annotation Tool



Data Distribution



Multi-Label Challenge

"Yeah, but they all end up with the same number on the top and bottom to make it one.

So it just depends on how many parts we're separating.

The denominator means that's how many parts you're separating one into, and the numerator is how many parts are filled up out of that denominator.

So using these two fractions, you see 3 5 is quite obviously smaller than 4 5 . Correct?"

PVD: Providing data

PVR: Provide Reasoning & Justification

BCS: Build Consensus

Figure: An example of multiple moves in a single utterances

Cross-Utterance Context Challenge

***PST**:*
“Okay. So, since we’re not looking at weight as an important property,
what would be another way that we can measure
to test out the mystery powder that isn’t one of these properties?”

***Jayla**:*
“That’s not one of these properties?”

***PST**:*
“Yes. Turn to talk to your partner.”

Figure: An example of implicit cross-utterance reference.

Dataset Comparison

CAMAL features:

- Multi-label scheme
- Conversation-level annotation

Dataset	Multi-label	Level	#Labels	#Samples
HWU64 (Liu et al., 2021)	No	Sentence	64	25,716
Clinic150 (Larson et al., 2019)	No	Sentence	150	23,700
Banking(Casanueva et al., 2020)	No	Sentence	77	13,083
CAMAL	Yes	Conversation	9	18,460

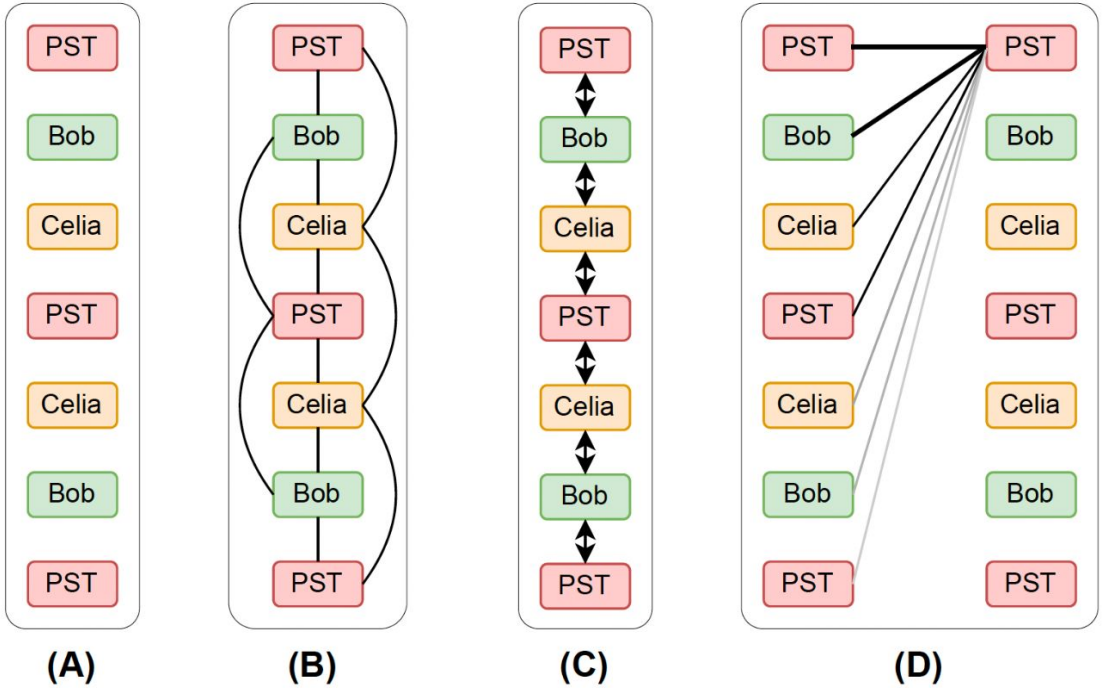
Models

(A) Simple MLP

(B) CNN

(C) BiLSTM

(D) Transformer



Speaker ID Graph

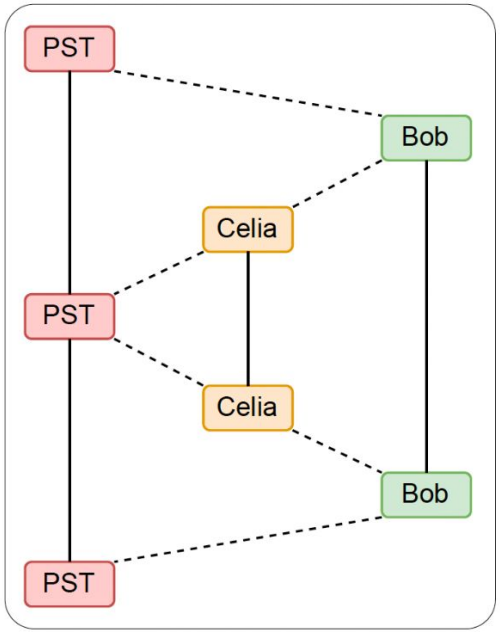
Discussion involves:

- Multiple speakers
- Arbitrary turns

> Encoding without knowing the speakers is suboptimal

(E) Speaker ID Graph's relations:

- Chronological relations (dash)
- Speaker ID relations (solid)



(E)

Experimental Results

Model	Dev				Test			
	P	R	F	↑	P	R	F	↑
MLP	48.2	**69.0**	55.2		59.0	48.4	50.7	
BiLSTM	59.0	67.5	59.4		63.4	60.1	61.0	
+ Graph	63.6	65.4	64.3	+4.9	65.6	63.1	63.8	+2.8
Transformer	**67.6**	61.0	62.9		68.0	58.5	61.5	
+ Graph	67.0	62.3	63.8	+0.9	67.9	59.1	62.8	+1.3
CNN	66.6	60.2	63.0		**68.8**	58.6	62.0	
+ Graph	63.1	65.1	63.8	+0.8	65.1	63.0	63.6	+1.6
CNN+LSTM	65.5	63.3	64.0		68.0	61.8	64.2	
+ Graph	61.4	68.4	**64.5**	+1.5	64.8	**65.8**	**65.1**	+0.9
Human performance					83.1	74.9	78.2	

CAMAL: A Novel Dataset for Multi-label Conversational Argument Move Analysis

Thank you!

Question → vietl@uoregon.edu