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

Viet Dac Lai[1], **Duy Pham**[2,3], **Jonathan Steinberg**[2]
Jamie Mikeska[3], **Thien Huu Nguyen**[1]
[1]Department of Computer Science, University of Oregon
[2]Educational Testing Services
[3]University of Massachusetts Amherst
{vietl, thien}@cs.uoregon.edu
{duypham,jsteinberg,jmikeska}@ets.org

Abstract

Understanding the discussion moves that teachers and students use to engage in classroom discussions is important to support pre-service teacher learning and teacher educators. This work introduces a novel conversational multi-label corpus of teaching transcripts collected from a simulated classroom environment for Conversational Argument Move AnaLysis (CAMAL). The dataset offers various argumentation moves used by pre-service teachers and students in mathematics and science classroom discussions. The dataset includes 165 transcripts from these discussions that pre-service elementary teachers facilitated in a simulated classroom environment of five student avatars. The discussion transcripts were annotated by education assessment experts for nine argumentation moves (aka. intents) used by the pre-service teachers and students during the discussions. In this paper, we describe the dataset, our annotation framework, and the models we employed to detect argumentation moves. Our experiments with state-of-the-art models demonstrate the complexity of the CAMAL task presented in the dataset. The result reveals that models that combined CNN and LSTM structures with speaker ID graphs improved the F1-score of our baseline models to detect speakers' intents by a large margin. Given the complexity of the CAMAL task, it creates research opportunities for future studies. We share the dataset, the source code, and the annotation framework publicly at http://github.com/uonlp/camal-dataset.

Keywords: educational assessment, teacher, student, simulated discussion, speaker graph identification

1. Introduction

In the last decade, simulated classrooms are being increasingly used in teacher education to support both pre-service teachers (PSTs) and in-service teachers in developing their ability to engage in core teaching practices. These practices include but are not limited to learning how to facilitate classroom discussions, elicit student thinking, manage student behavior, work with diverse student populations, and communicate with families (Badiee and Kaufman, 2014; Chazan and Herbst, 2012; Mikeska and Howell, 2020; Mikeska et al., 2021). Prior research suggests that opportunities to practice teaching and receive feedback and reflect on the experience can positively influence teachers' learning. Moreover, targeted, timely, and actionable feedback specific to the teaching task and teachers' performances have been shown to improve their instructional practice (Kazemi and Cunard, 2016; Benedict-Chambers and Aram, 2017; Kleinknecht and Groschner, 2016; Snead and Freiberg, 2019).

Simulated classrooms are often used as practice spaces to support teacher learning. However, the current approach of relying on human raters, coaches, or teacher educators to evaluate live or recorded teaching performances and provide feedback presents challenges. The process of manually scoring and generating feedback for simulated teaching is time-consuming and not scalable. Therefore, developing an automatic analysis system is crucial to accelerate the provision of scores and feedback for the participating PSTs, which can support their learning. To enable the development and evaluation of automated analysis and automated scoring of teacher-student discussions, benchmark datasets play a crucial role. Unfortunately, most of the currently available datasets for intent detection are not suitable for developing systems to detect PSTs and students' argument moves due to three reasons. First, the current datasets are collected for analyzing customer needs in online customer services but not for teaching practices (Liu et al., 2021). Second, the texts in these datasets are very short (e.g., one to a few sentences) (Larson et al., 2019; Casanueva et al., 2020). As such, applying these datasets and models devised to extract features from them in longer classroom sessions with context dependency of a multi-speaker conversation in teacher-student discussion might hinder its application (Erduran et al., 2004). Third, these datasets only associate an utterance with a single label, while in a complex con-

LREC-COLING 2024, pages 2673–2682
20-25 May, 2024.

versation like a teacher-student discussion, a single utterance might exhibit multiple intents(Larson et al., 2019; Casanueva et al., 2020). As a result, the existing corpora are not a good fit for detecting teacher-student argument moves in simulated classrooms.

To address this issue, in this paper, we present a novel dataset for Conversational Argument Move AnaLysis, called **CAMAL**. The argument moves in the CAMAL dataset were human annotated by trained annotators with expertise in mathematics and science teaching, research on teacher learning, and/or assessment development (Mikeska et al., 2023). The dataset comprises 165 full-length teacher-student transcripts transcribed from audio or video recordings of classroom instruction in which PSTs facilitated discussions with five student avatars in a simulated environment environment on a science (Mystery Powder science discussion) or mathematics (Ordering Fractions mathematics discussion) topic. Each discussion that a PST facilitated could last up to 20 minutes, although PSTs could end the discussion any time before that time. In total, across the 165 discussions (XX using the Mystery Powder science discussion task, XX using the Ordering Fractions discussion task), there is approximately 44 hours of discussion recorded. The dataset features nine argumentation moves that have been suggested in the empirical literature as key components of high-quality, argumentation-focused discussions in these disciplines. Each teacher or student utterance can be associated with multiple argumentation moves. Finally, we propose a graph-based neural network model, which features a speaker ID graph, to significantly improve the performance of all the baseline models.

The contribution of this paper includes:

- We formulated the convolutional argument move analysis in educational assessment as an NLP task.
- We created the first dataset for convolutional argument move analysis to standardize the study of the topic.
- We proposed graph convolutional neural network with speaker identification graph to improve the performance of the baselines model for the convolutional argument move analysis.

2. Data Creation

2.1. Data Collection

In our study, we collected video and audio files of elementary PSTs who facilitated an argumentation-focused discussion with five upper elementary student avatars in a simulated classroom. These videos feature two performance tasks: the Mystery Powder science task (Mikeska and Howell, 2020) and the Ordering Fractions mathematics task (Mikeska et al., 2021). During each discussion, a trained human called an interactor plays the role of the five student avatars and uses specialized technology to speak and behave like upper elementary students. The PST stays in front of a screen in which the student avatars appear and the PST interacts with them in real-time through the screen. The interactor was trained by content and simulation experts to be responsive to the PST's facilitation during the discussion. Certain student ideas (e.g., a student's initial claim about the ordering of the factions) remain consistent at the beginning of each discussion, but can be changed based on what the other students and teachers say and do during the discussion. We transcribed each discussion video recorded into textual transcripts using a transcribing company (rev.com) for use in this study.

In the Mystery Powder science task, the PST moderates a discussion among the student avatars to support them in identifying an unknown powder among six known powders (e.g., baking soda, baking powder, salt, sugar, flour) using data about the powders' properties to come to a consensus about which properties were most helpful to reveal the mystery powder's identity. The student avatars start the discussion with different claims and justifications about the unknown powder and the properties that are most useful to identify it. In the Ordering Fractions mathematics task, the PST facilitates a discussion among the student avatars to build consensus around ordering three different fractions and evaluate the strategies' generalizability for use in ordering any fraction sets. Similar to the science task, the student avatars start the discussion with different strategies to order the three fractions and reasoning about whether the approaches they used could be applied to all sets of fractions.

2.2. Taxonomy

Our team designed an annotation framework to characterize how the PSTs prompted students to engage in mathematical and scientific argumentation and how the students engaged in argumentation during these discussions. Toward this end, we designed an intent taxonomy that harvests knowledge from three primary sources: (1) prior empirical and practitioner literature in education, (2) our own observations and experiences in our other research projects, and (3) expertise of our assessment developers and research scientists in assessment and linguistics.

The development of the annotation framework occurred collaboratively and through several iter-

Argument Move	Acronym	Description
Explicating Argumentation	EXA	Communicating the key features or characteristics of high-quality argumentation discussions (needs to be explicit)
Eliciting A Claim	ELC	Asking or encouraging others to share their claims related to the discussion's targeted student learning goal (without data or reasoning to support the claims)
Stating A Claim	STC	Sharing a claim related to the discussion's targeted student learning goal (without providing any data or reasoning).
Eliciting Data	ELD	Asking or encouraging others to provide data to support or refute a claim without reasoning to support the claim).
Providing Data	PVD	Sharing data to support or refute a claim (without providing reasoning).
Eliciting Reasoning & Justification	ELR	Asking or encouraging others to share their reasoning to support or refute a claim (without data to support the claim).
Providing Reasoning & Justification	PVR	Sharing reasoning to support or refute a claim (without providing data)
Building Consensus	BCS	A focus on consensus and/or providing opportunities for building consensus among participants.
Evaluating	EVL	Providing an evaluation of an argument, or part of an argument (claim, data, and/or reasoning), or an evaluation of whether an argument is strong or weak.
No Code Applied	NCA	A special label used when none of the nine codes above is applicable to the utterance.

Table 1: A brief description of argumentation moves.

ations among team members. In the first step, we relied upon our knowledge of the empirical research literature and our team's previous experience observing these discussions to brainstorm a comprehensive list of 23 possible moves for inclusion in the framework. Seven of these initial moves (e.g., eliciting data; providing reasoning/justification) were selected to capture the argumentation moves that teachers used to ask students to provide data and reasoning and the moves that students took to state a claim, explain reasoning, and justify their own or others' statements (Erduran et al., 2004; Oyler, 2019). Eight of these initial moves (e.g., raising or responding to counter arguments/challenges/rebuttals; evaluating) were proposed to characterize the argumentation moves that PSTs used to encourage students to debate with each other and the moves that the students took to examine, analyze, and convince one another about specific claims. The other eight initial moves reflected conversational moves that a student might make during a discussion, such as distilling another person's point or recommending another person to come back to or continue to discuss something that has not been debated in depth (Oyler, 2019).

Second, our research team used this set of 23 initial moves to annotate PST and student utterances in two mathematics and and two science discussion transcripts. In our first week, we conducted a group annotation of one transcript in each content domain. In the second week, we did individual annotation with group reconciliation of another mathematics transcript and science transcript. Learning from the group and individual annotation with reconciliation, we significantly refined the annotation taxonomy by combining similar moves and removing moves that did not directly capture the PST skills of facilitating students' argumentation. The finalized annotation taxonomy consisted of nine moves that characterized the important argumentation moves that the PSTs made to facilitate the students' engagement in argumentation and the moves that the students made to engage in argumentation during these discussions. Table 1 presents the list of the labels and their brief descriptions.

2.3. Annotator Training

Before the annotation of the full set of transcripts, we used a small number of transcripts to train all our annotators. During the training, we began by discussing each of the moves, then their descriptions and examples of the utterances that should be associated with that move. Then, each team member independently annotated one science and one mathematics transcript. In the next step, we conducted a group reconciliation for each transcript. After that, our annotators worked individually to annotate one more mathematics and one more science transcript. To conclude the training, we asked our annotators to reconcile their labels with their

Argument Move	Kappa score	Raw match
EXA	0.44	96.78
ELC	0.51	93.59
STC	0.56	89.74
ELD	0.56	96.83
PVD	0.65	93.29
ELR	0.59	90.64
PVR	0.62	89.81
BCS	0.61	91.59
EVL	0.57	84.80
All	0.60	91.90

Table 2: Initial agreement score before reconciliation.

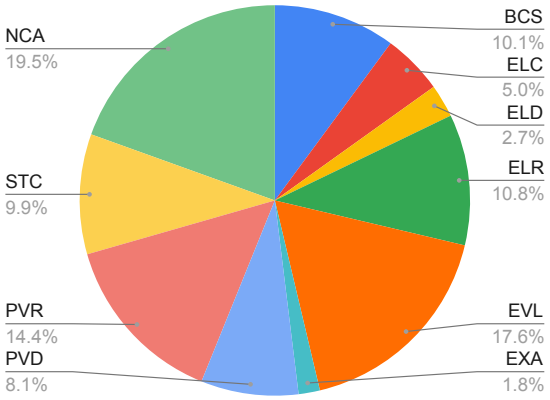

Figure 1: Distribution of the argument moves in the CAMAL corpus.

partners. Once the training on one mathematics and one science transcripts was complete, our five annotators conducted the annotation work on a weekly basis for 20 weeks to finish the annotation of 165 transcripts. Each week, each rater annotated two to three transcripts individually, as well as reconciled one or two transcripts from their previous week's annotations with one of the other annotators. To monitor the annotation quality, we held bi-weekly meetings for the whole team to share and discuss annotation challenges and make minor revisions or add more examples to the annotation taxonomy to ensure that we shared a mutual understanding of how to consistently assign the moves to each utterance within the transcripts.

To ensure the quality of the annotation process, we monitored several statistics to evaluate the annotation quality. The statistics included percent exact agreement, Cohen's kappa, and the intraclass coefficient. These statistics were estimated at the transcript and per-label level for 27 doubled-annotated transcripts. Nine of these transcripts received Kappa score value below 0.50. We asked our annotators to reconcile their annotations for these transcripts to increase annotation agreement. After reconciliation, the agreement statistics were not adjusted because we presumed the final annotation results would show a perfect agreement between raters. Table 2 shows the average Kappa score statistics for individual moves after the initial annotation was done and before reconciliation. Our overall evaluation shows a Cohen's Kappa score of 0.60, which indicates a moderate to substantial agreement between the annotators.

Figure 1 shows the overall distribution of labels across the PST and student avatar utterances within the 165 discussion transcripts. Table 3 compares the CAMAL corpus with existing benchmark corpora in intent detection. To the best of our knowledge, our CAMAL corpus presents the first multi-label conversation-level intent detection dataset. While it offers fewer intent types, its sample per intent is much higher due to the comparable total number of annotated intents.

2.4. Challenges

During our human annotation process, we aimed to ensure that annotators had a shared and consistent understanding of when and how to apply the moves to individual utterances within the transcripts, including when none of the moves should be applied. Given the complexity of the interaction within our transcripts, five annotation challenges were noticed and we went through a few iterations to mitigate these challenges.

2.4.1. Taxonomy design

The first challenge for us was to finalize our annotation framework to capture the argumentation moves used by PSTs and the student avatars without being too complicated to make the annotation work not feasible or impractical. From the literature on argumentation discourse, we initially proposed 23 moves to annotate the transcripts. These labels reflect argumentation moves regarding prompting students to provide data and reasoning, debate arguments, or general moves such as paraphrasing someone's argument or redirecting a person to return to a discussion point. After working with our experts to apply these moves to two sample transcripts, we realized that 23 moves were too many which made it hard for the annotators to decide which moves to use for which utterances. For example, the difference between paraphrasing and distilling moves was too subtle for our annotators to distinguish between them. For that reason, we decided to drop these two moves out of our final annotation framework. We use the label NCA which stands for No Codes Applied for utterances that do not reflect any aspects of the construct being measured and nine moves that capture the quality

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

Table 3: Statistics of the datasets.

of argumentation facilitation of the PST. Table 1 presents these argument moves in detail.

2.4.2. Argument Move Detection

Our next challenge was to decide which utterances did not reflect the construct being measured which is the nature of the argumentation moves used by the PSTs and student avatars, so were assigned the NCA label in table 1. Take the following utterances as an example:

PST: *"Great, okay. So let's move into, what do you guys think is the next property we should, we tested?"*

...

Student:*"Well, we decided as a group that color wasn't important, so reaction with vinegar."*

One of our annotators thought the student's utterance does not reflect any aspect of the construct being measured because it is just a restatement of a point that had previously been discussed, thus should be assigned the label NCA. However, some other annotators thought we should assign the label STC (State a Claim) to this utterance because the student made that statement to answer a question from the PST earlier in the discussion. The claim, in this case, was that color was not an important feature that can help the group identify the mystery powder and reaction to vinegar was actually helpful to solve the problem of figuring out the powder.

2.4.3. Cross-Utterance Context

The third challenge was how we took into account the content surrounding an utterance while we were assigning moves to it. For example, we assigned the label Eliciting Reasoning and Justification (ELR) for the utterance below because the PST tried to ask the student avatars to provide reasoning for their claim of which properties of the substance would be helpful for them to identify the powder.

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."*

An annotator was unsure whether we should assign the label ELR again for the last utterance because the PST answered Jayla and told the students to discuss it with their partners We decided not to assign the label of ELR to this utterance because the elicitation of reasoning really happened in the first utterance but not in the last one.

2.4.4. Multi-label Utterances

The fourth challenge was to annotate utterances that can receive multiple labels. These utterances usually had multiple sentences and touched on a few different aspects of the discussion. Below is one example of such an utterance:

Student: *"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 $\frac{3}{5}$ is quite obviously smaller than $\frac{4}{5}$. Correct?"*

One of our annotators assigned Building Consensus (BCS) for this utterance. Whereas another one chose multiple moves, which included (Eliciting a Claim) ELC, Stating a Claim (STC), Eliciting Reasoning & Justification (ELR), and Providing Reasoning & Justification (PVR) for it. We then had a third annotator who came in and worked with the first two annotators to reconcile the label. After discussing the text and anchoring the discussion in the annotation framework, we decided to assign three moves of PVD, PVR, and BCS to the utterance. Below is the reasoning of the third annotator: *"No ELC (refer to specified claims in the document); no STC for the same reason, even if you believe this is relevant we don't annotate sub-claims as STC; Yes to PVD and PVR (stating known facts about what the numerator and denominator are and the observation that their equality means the interval is still one, reasoning supporting the statement that 3/5 < 4/5). Agree with BCS, probing for points of consensus."*

2.4.5. Ambiguous Argument Moves

The last challenge was to draw the fine line for a few pairs of moves such as "Stating a Claim" versus "Providing Data" or "Eliciting a Claim" and

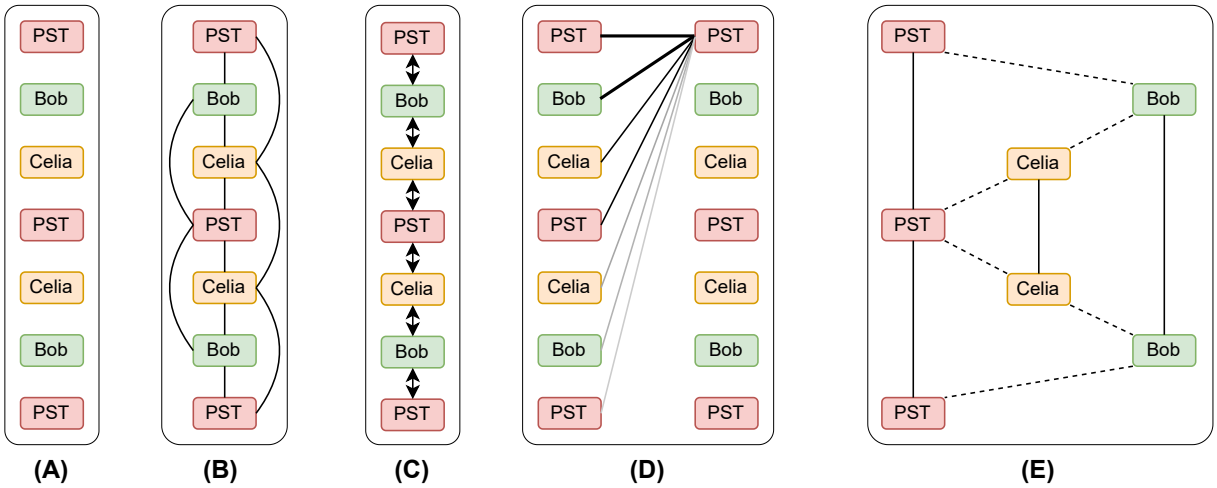

Figure 2: Architectures of the examined models. (A) Baseline model with isolated utterances. (B) CNN model with kernel=5. (C) Bidirectional LSTM. (D) Self-attention transformer. (E) Speaker ID Graph with chronological edges (dash lines) and speaker edges (solid lines).

"Eliciting Data". For example, one of our annotators applied the "Stating a Claim" for this utterance *"Well, we can eliminate baking powder because it looks cloudy."* Meanwhile, another annotator argued we should use "Providing Data" instead of the "Stating a Claim" label because the second part of the sentence was to provide the information that the baking powder looks cloudy.

In summary, given the complexity of the interaction among the PSTs and student avatars in these discussions in our transcripts, we decided to use fewer moves in our final annotation taxonomy to focus only on the key features related to the quality of the argumentation facilitation of the PST. The five annotation challenges we described here reflected the difficulty our annotators faced when they annotated the transcripts. These challenges were more present during the first few weeks of the annotation. The more we went into the annotation, the more we know what moves we should assign to each utterance and our inter-rater agreement was improved over time through reconciliation and bi-weekly meetings to discuss the annotation work.

3. Models

3.1. Baseline

In this work, we examine the text classification format in previous intent detection work (Larson et al., 2019). In particular, given an utterance of L_U sentences $(S_1, S_2, \cdots, S_{L_U})$, an input sequence $[CLS], S_1, [SEP], S_2, \cdots, [SEP], S_{L_U}, [SEP]$ is fed to a large pre-trained language model (PLM). Then, the representation h^{CLS} of the $[CLS]$ token is obtained as the representation for the whole utterance. Then, nine separate binary classifiers are employed to predict the Positive/Negative label for each argument moves.

3.2. Document-level Sequence Labeling

The distinction of the CAMAL task presented in this paper compared to other previous intent detection tasks is that the label of an utterance depends on the content of not only the current utterance but also the preceding utterance. Hence, modeling each utterance independently is suboptimal. As such, to address this issue, the surrounding context must be considered to predict the labels of an utterance. To do that, we employed CNN, LSTM, and Transformer to encode the context of the conversation.

In particular, given a sequence of utterances $(U_1, U_2, \cdots, U_{L_U})$, a sentence-level encoder embeds every utterance similar to the baseline model, producing a sequence of utterance-level hidden states $(h_1^{CLS}, h_2^{CLS}, \cdots, h_{L_U}^{CLS})$. A second encoder, based on CNN, LSTM, or Transformer, encodes the dependency sequence at utterance level to produce higher-level hidden states $(m_1, m_2, \cdots, m_{L_U})$. Finally, a set of nine classifiers is used to predict the labels for each utterance, similar to the baseline model.

3.3. Speaker ID Graph

In this work, the discussion involves multiple speakers with arbitrary turns, hence, encoding the conversation without knowing the speaker's identity can lead to a suboptimal solution for several reasons. First, the discussion session involves multiple speakers. Their utterances are mixed up when presented in the chronological order of the transcript. As a result, without knowing the speaker

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	

Table 4: Performance of the models on the development and test sets of CAMAL dataset. The columns with ↑ show the improvements of the speaker ID graph over the base model.

identities of the utterances, the arguments are also mixed-up, causing the model to wrongly detect their moves. Secondly, an argumentation point can be scattered across multiple utterances. As such, tracking the development of the argumentation moves would prevent missing information in those fragmented arguments. To address this problem, we proposed using the speaker identities of the utterances to model their argumentation moves throughout the transcript. In particular, besides modeling the argumentation moves using chronological order, we modeled the argumentation moves based on the previous utterance of the same speaker, hence, avoiding interference with other speakers' utterances.

In particular, given a sequence of utterances $(U_1, U_2, \cdots, U_{L_U})$ which associate with speaker identities $(I_1, I_2, \cdots, I_{L_U})$. The speaker ID graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of nodes $\mathcal{V} = (\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_{L_U})$, which corresponds to the set of utterances. The set of edges contains two edge categories: global chronological edges and speaker edges as follows:

$$\mathcal{E} = \mathcal{E}_{chronological} \cup \mathcal{E}_{speaker}$$
$$\mathcal{E}_{chronological} = \{(U_i, U_{i+1}) | 0 < i < L_U)\}$$
$$\mathcal{E}_{speaker} = \{(U_i, U_j) | I_i = I_j)\}$$

Then, the utterances are encoded using a graph convolutional neural network (GCN) (Kipf and Welling, 2017) using the above speaker ID graph. We use the same utterance-level hidden states to feed the GCN module layers. Its output is concatenated to the final representations of the other baseline models. Figure 2 compares the speaker ID graph against other model architectures.

4. Evaluation

We evaluate the above models using the (macro) precision, recall, and F1 score metrics in this work.

4.1. Results

Table 4 reports the performance of the models on the development and the test sets of the CAMAL corpus. There are three significant observations from the table. First, comparing the naive baseline model against the sequence labeling model, the performance of the baseline MLP model (F1=50.7%) is significantly lower than the group of sequence labeling models (F1 scores > 60%). This clearly demonstrates that the task presented in the CAMAL dataset is complex and that understanding the content of an utterance is not enough to predict the argument moves presented in the utterance precisely. A more sophisticated model is needed to capture the information from the surrounding utterances such as BiLSTM, CNN, and Transformer.

Secondly, for the group of document-level sequence labeling models, CNN (F1=62.0%) performs better than the Transformer model (F1=61.5%), while BiLSTM is the worst among these (F1=61.0%). This indicates that short-term dependency is more critical in modeling the argument moves (in CNN model) than long-term dependency (in BiLSTM and Transformer). Moreover, modeling both short-term and long-term dependencies (in the CNN+LSTM model) achieves the highest performance (F1=64.2%).

Thirdly, models with the speaker ID graph (+Graph models) yield consistently superior performances against the base models without the speaker ID graph(BiLSTM, Transformers, CNN, and BiLSTM+CNN). The performance gains from $+0.9\%$ to $+2.8\%$ on the test set. This confirms the effectiveness of the speaker ID graph in modeling the dependencies between utterances in a multi-speaker discussion in the CAMAL dataset.

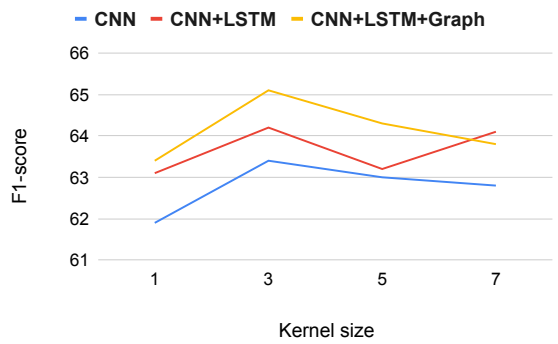

Figure 3: Performance comparison of CNN-based models based on context length.

4.2. Context length analysis

The above result has suggested that the surrounding context is important for the prediction of the argument moves. In this experiment, we aim to identify the extent of the context that the model should consider to obtain the best performance. To do that, we consider the performance of CNN-based models with different kernel sizes (1, 3, 5, 7). Figure 3 presents the performances of the examined models. Increasing the kernel size from 1 to 3 improves the performance of all three models. More importantly, the performances of all three models peak at a kernel size of 3. Increasing kernel size to 5 and 7 worsens the F1 score of all models with different patterns. The vanilla CNN model shows a slight decrease, the CNN+LSTM 's performance fluctuates, while the performance of the CNN+LSTM+Graph decreases at a steeper rate than the CNN model.

5. Related Work

Intent detection has been studied in NLP for various domains. Three most common corpora in intent detection are Banking77 (Casanueva et al., 2020), HWU64 (Liu et al., 2021), and CLINIC150 (Larson et al., 2019). However, these corpora only consider multiclass classification for intent detection. As such, the models that are designed for these datasets are mostly text-classification models (Casanueva et al., 2020; Papangelis et al., 2021). Transferring knowledge across known and unknown intents is important to enable intent detection in broader applications (Liu et al., 2019; Larson et al., 2019; Rastogi et al., 2020; Wu et al., 2021).

In educational assessment, intention detection has been used to collect features from student works or performance transcripts to support automatic scoring (Burstein et al., 1998; Sarker et al., 2019). On the one hand, our precision, recall, and F1-scores reported in Table 4 for CNN+LSTM+Graph model were better than the results of some prior studies (Cui, 2021; Lugini and Litman, 2018). On the other hand, our best results were lower than in some other investigations (Ariely et al., 2022; Sarker et al., 2019) To the best of our knowledge, our corpus is the first that shares data from teaching transcripts and develops NLP models to detect argument moves in a classroom discussion. Moreover, annotated data are not usually shared in educational measurement. That is one of the reasons we want to share our corpus to invite researchers to work on this teacher-student argument moves analysis problem with us.

6. Conclusion & Future Work

We described a corpus of argumentation-focused teaching transcripts generated from classroom discussions in a simulated environment along with a framework to annotate the data for argument moves that PST or student avatars took to move the discussion along. We showed the necessity of argument move analysis in automatic PST evaluation in discussion practices. We presented a human-annotated multi-label argument move analysis for teacher-student discussions. The experiment shows that adding the speaker ID graph into our baseline models helped improve the F1 score. The addition of the speaker ID graph to better encode the content of the conversation results in consistent improvement in all examined model architectures. The improvement of adding the speaker ID graph to the baseline models was quite notable. However, the results were far from the coding consistency of our human annotators. In this case, we believe that other researchers can offer more ideas to make the model we introduced in this paper better. Suppose that we can develop and deploy more effective models to detect PSTs and students' intents in simulated classroom discussions, the argument moves would be used to enrich the automatic scoring and feedback systems to provide timely and cost-effective feedback to PSTs.

Acknowledgements

We thank our colleagues at ETS for their participation in providing detailed annotations and insightful feedback throughout the work. This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI or the U.S. Government.

Bibliographical References

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2022. Machine learning and hebrew nlp for automated assessment of open-ended questions in biology. *International journal of artificial intelligence in education*, pages 1–34.

Farnaz Badiee and David Kaufman. 2014. Effectiveness of an online simulation for teacher education. *Journal of Technology and Teacher Education*, 22(2):167–186.

Eric R Banilower, P Sean Smith, Kristen A Malzahn, Courtney L Plumley, Evelyn M Gordon, and Meredith L Hayes. 2018. Report of the 2018 nssme+. *Horizon Research, Inc.*

Amanda Benedict-Chambers and Roberta Aram. 2017. Tools for teacher noticing: Helping preservice teachers notice and analyze student thinking and scientific practice use. *Journal of Science Teacher Education*, 28(3):294–318.

Jill Burstein, Lisa Braden-Harder, Martin Chodorow, Shuyi Hua, Bruce Kaplan, Karen Kukich, Chi Lu, James Nolan, Don Rock, and Susanne Wolff. 1998. Computer analysis of essay content for automated score prediction: A prototype automated scoring system for gmat analytical writing assessment essays. *ETS Research Report Series*, 1998(1):i–67.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Chazan and Patricio Herbst. 2012. Animations of classroom interaction: Expanding the boundaries of video records of practice. *Teachers College Record*, 114(3):1–34.

Zhongmin Cui. 2021. Machine learning and small data. *Educational Measurement: Issues and Practice*, 40(4):8–12.

Sibel Erduran, Shirley Simon, and Jonathan Osborne. 2004. Tapping into argumentation: Developments in the application of toulmin's argument pattern for studying science discourse. *Science education*, 88(6):915–933.

Elham Kazemi and Adrian Cunard. 2016. Orienting students to one another and to the mathematics during discussions. In *Qualitative Research in STEM*, pages 295–313. Routledge.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Marc Kleinknecht and Alexander Groschner. 2016. Fostering preservice teachers' noticing with structured video feedback: Results of an online- and video-based intervention study. *Teaching and Teacher Education*, 59:45–56.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 165–183. Springer.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium. Association for Computational Linguistics.

Jamie N Mikeska and Heather Howell. 2020. Simulations as practice-based spaces to support elementary teachers in learning how to facilitate argumentation-focused science discussions. *Journal of Research in Science Teaching*, 57(9):1356–1399.

JN Mikeska, H Howell, J Ciofalo, A Devitt, E Orlandi, K King, M Lipari, and G Simonelli. 2021. Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate

argumentation-focused discussions in mathematics: The mystery powder task. *Research Memorandum No. RM-21-06). ETS*.

J.N. Mikeska, J. Steinberg, T. Maxwell, A. Marigo, D. Pham, V. Lai, and T. Nguyen. 2023. Exploring the argumentation features of mathematics and science discussions within online simulated teaching experiences. *American Educational Research Association Annual Meeting*.

Joe Oyler. 2019. Exploring teacher contributions to student argumentation quality. *Studia paedagogica*, 24(4):173–198.

Alexandros Papangelis, Karthik Gopalakrishnan, Aishwarya Padmakumar, Seokhwan Kim, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Generative conversational networks. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–120, Singapore and Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Lauren Oropeza Snead and H Jerome Freiberg. 2019. Rethinking student teacher feedback: Using a self-assessment resource with student teachers. *Journal of Teacher Education*, 70(2):155–168.

Ting-Wei Wu, Ruolin Su, and Biing Juang. 2021. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4884–4896, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.