Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms

Abhijit Suresh[1,2], **Jennifer Jacobs**[2], **Margaret Perkoff**[1],
James H. Martin[1,2], **Tamara Sumner**[1,2]
[1]Department of Computer Science, [2]Institute of Cognitive Science
University of Colorado Boulder
FirstName.LastName@colorado.edu

Abstract

"Talk moves" are specific discursive strategies used by teachers and students to facilitate conversations in which students share their thinking, and actively consider the ideas of others, and engage in rich discussions. Experts in instructional practices often rely on cues to identify and document these strategies, for example by annotating classroom transcripts. Prior efforts to develop automated systems to classify teacher talk moves using transformers achieved a performance of 76.32% F1. In this paper, we investigate the feasibility of using enriched contextual cues to improve model performance. We applied state-of-the-art deep learning approaches for Natural Language Processing (NLP), including Robustly optimized bidirectional encoder representations from transformers (Roberta) with a special input representation that supports previous and subsequent utterances as context for talk moves classification. We worked with the publically available TalkMoves dataset, which contains utterances sourced from real-world classroom sessions (human- transcribed and annotated). Through a series of experimentations, we found that a combination of previous and subsequent utterances improved the transformers' ability to differentiate talk moves (by 2.6% F1). These results constitute a new state of the art over previously published results and provide actionable insights to those in the broader NLP community who are working to develop similar transformer-based classification models.

1 Introduction

There is a strong theoretical and empirical basis for encouraging students' active participation in inquiry-based and socially constructed classroom environments (Vygotsky, 1978; Webb et al., 2008). Numerous efforts exist to support teachers to become more purposeful and effective in their efforts to facilitate such environments (Herbel-Eisenmann, 2017; Chen et al., 2020). Most approaches to providing teachers with detailed feedback about their discourse strategies require highly trained human observers (Correnti et al., 2015; Wolf et al., 2005). However, recent research has shown that the development and training of deep learning models to automate and scale certain discourse analyses from instructional episodes is feasible (Song et al., 2021), effective (Demszky et al., 2021), and reliable (Donnelly et al., 2017; Jensen et al., 2020; Suresh et al., 2019).

Accountable talk theory offers well-defined, research-based practices for teachers to engage in high-quality instruction, including the use of specific talk moves that promote students' equitable participation in a rigorous learning environment (O'Connor et al., 2015; Resnick et al., 2018). By using talk moves, teachers place the "intellectual heavy lifting" and balance of talk toward students and help ensure that the discussions will be purposeful, coherent, and productive (Michaels et al., 2010). Talk moves support classroom discourse to move beyond the traditional Initiate, Response, Evaluate linguistic sequence (Mehan, 1979); namely, by replacing the act of evaluating with practices that support a collective understanding that builds on and extends mathematical ideas (Michaels and O'Connor, 2015).In this way, talk moves enable dialogue shifts from teacher directed recitation to true discussions in which knowledge is informally shared and constructed rather than transmitted.

This paper draws inspiration from speech recognition systems for spoken dialog systems to investigate the feasibility of applying a novel input representation that utilizes tokens from previous and subsequent utterances to classify teacher talk moves (Schukat-Talamazzini et al., 1994). We explore three different context setups: previous-only utterances, subsequent-only utterances, and both previous and subsequent utterances (equal numbers of each) with different window sizes. In addition to the longer dialog window experiments, we re-

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 71 - 81
July 15, 2022 ©2022 Association for Computational Linguistics

port findings from fine-tuning transformers such as BigBird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020) which are architected to support longer sequences. Similarly, we report findings from fine-tuning MathBERT, a transformer architecture that was trained to establish semantic correspondence between mathematical formulas and their corresponding context (Peng et al., 2021). For training and evaluation, we use the TalkMoves dataset comprising 567 lesson transcripts derived from video recordings of K-12 mathematics classrooms (Suresh et al., 2022). The main contributions of this work are summarized as follows:

- We provide evidence for improved performance when fine-tuning transfomers with longer dialog windows.
- We observed that transformer architectures designed to handle longer contexts such as Longformer do not provide any additional benefit in differentiating instructional strategies.
- We observed that math-based models pre-trained on mathematical formula understanding do not provide any improvement over the generic models.

2 Related Work

This section briefly describes the accountable talk theory framework, followed by a literature review on deep learning models for Natural Language Processing (NLP) focused on adding additional contexts and learning long-term dependencies.

2.1 Accountable talk theory framework

Accountable talk theory identifies and defines an explicit set of discourse moves intended to elicit a response within a classroom lesson (O'Connor and Michaels, 2019). These well-defined discursive techniques have been incorporated into various instructional practices and frameworks e.g., (Boston, 2012; Candela et al., 2020; Michaels et al., 2010). Their specificity makes talk moves well-suited for supervised multi-label sentence-pair classification. A number of research teams have made considerable progress in developing automated "intelligent agents" that are trained to emulate the role of the teacher. These agents prompt students to use designated aspects of accountable talk, such as revoicing and asking students to agree/disagree with another student. They typically act as facilitators or tutors during small group, text-based, online settings, taking part in and helping to focus the discussion at opportune moments e.g. (Adamson et al., 2013; Hmelo-Silver et al., 2013; Tegos et al., 2015). (Jacobs et al., 2022) and team developed an online application that provides personalized feedback to teachers on their classroom discourse practices, including the prevalence of talk moves. The system is fully automated and requires no human processing beyond the initial uploading of classroom recordings. Such education-focused NLP applications are in high demand to provide reliable feedback to teachers based on the accountable talk theory.

2.2 Transformers for additional context and long-term dependencies

The introduction of transformers has revolutionized the field of natural language processing. Unlike Recurrent Neural Networks (RNNs) and Long Short Term Memory networks (LSTMs), where training is performed sequentially, the design of transformer architecture enables parallel processing and allows for the creation of rich latent embeddings (Vaswani et al., 2017). Latent contextual representation of utterances through the self-attention mechanism makes transformers a powerful tool for various downstream applications such as question answering and text summarization (Devlin et al., 2018).

Research efforts to learn long-term dependencies with transformers were first introduced in Transformer-XL (Dai et al., 2019). Transformer-XL is a novel architecture that focuses on learning dependencies beyond the fixed length of vanilla transformers without disrupting the temporal coherence. This is achieved by saving the hidden state sequence of the previous segment to be used as context for the current segments, also known as the segment-level recurrence mechanism. In addition, to better encode the relationship between words, Transformer-XL uses relative positional embeddings. Results show that Transformer-XL can learn dependencies across the text with a window size of 900 words. Following Transformer- XL, (Yang et al., 2019) proposed XL-Net, which is a generalized autoregressive pretraining method that leverages the capabilities of Transformer-XL to solve the pre-train-finetune discrepancy commonly identified in early architectures such as BERT. XL-Net introduced two new developments. As an extension to the standard Causal Language Modeling (CLM), XL-Net uses permutation language modeling, which considers all possible permutations of the words within a sentence during the training phase. Also, XL-Net uses a secondary attention stream that focuses on the positional information of the predicted token. This additional attention stream led XL-Net to outperform many contemporary transformer architectures in downstream tasks, such as text classification. Similarly, to address the problem of processing long sequences with transformers, (Beltagy et al., 2020) introduced Longformer, which extends vanilla transformers with a modified self-attention mechanism to process long documents. The classic self-attention mechanism in BERT is computationally expensive, which explains the restriction of the maximum sequence length of 512 tokens. Instead, Longformer combines dilated sliding windows with global attention to achieve similar performance. As a result of reducing the computational complexity, Longformer can process long input sequences beyond the previously defined segment length of 512 tokens. Like Longfomers, Big-Bird (Zaheer et al., 2020) uses a sparse attention mechanism that includes a random attention component.

Over the past few years, we have seen an increasing trend in other approaches to supporting transformers to learn long-term dependencies, such as modifying pre-training methods and the classic attention mechanism. For example, to learn dependencies across documents, (Xie et al., 2020) adopted a simple approach to truncate the document used for classification. Similarly, (Joshi et al., 2019)) used a chunking approach where documents were broken down into multiple chunks, and the activations were then combined to perform the tasks. Another recent example is the BERT-Seq model for classifying Collaborative Problem Solving (Pugh et al., 2021). The BERT-Seq model uses a special input representation that combines embeddings from adjacent utterances as contextual cues for the model. Building on the prior work, we explored new ways to enrich transformers with additional contextual cues.

3 Current Work and Novelty

Currently, generating information about teachers' discourse strategies requires highly trained instructional experts to hand-code transcripts from classroom sessions (Correnti et al., 2015; Wolf et al., 2005), an approach that is expensive and not readily scalable. Encouragingly, a small number of researchers have recently trained computer models to automate and scale discourse analyses from instructional episodes, detecting educationally important discursive features such as instructional talk, authentic teacher questions, elaborated evaluation, and uptake (Dale et al., 2022; Demszky et al., 2021; Jensen et al., 2020). In prior work, (Suresh et al., 2021b,a) fine-tuned Roberta (Liu et al., 2019) to classify talk moves for each teacher utterance from a given classroom transcript. The input to Roberta was student-teacher sentence pairs, where the student sentence appeared immediately prior to the teacher's utterance. This paper builds upon the previous work to add contextual cues to transformers in various ways and evaluate their performance using the TalkMoves dataset. We experiment with modifying the input representation by combining multiple previous and subsequent utterances as context to classify teacher talk moves. This work serves as an example of how we can find new ways to use advances in natural language processing with classic ideas from speech recognition systems for spoken dialog system to capture the rich conversations between teachers and students in order to improve performance in applied domains such as education.

4 Method

This section discusses the different approaches we took to enrich contextual cues in the TalkMoves model in an effort to enhance performance.

4.1 Data

The TalkMoves dataset used in this study comprises 567 transcripts, including 174,186 teacher and 59,874 student utterances (Suresh et al., 2022). All the transcripts were human-generated from classroom audio and video recordings from K-12 mathematics classrooms. They were annotated for six teacher talk moves by two experts who established high inter-rater reliability (Suresh et al., 2021b, 2022). The talk moves in the dataset follow an uneven distribution, with certain moves being much more frequent than others (Figure 1). "Keeping everyone together" and "pressing for accuracy" are the most frequently used, whereas "getting students to relate" and "pressing for reasoning" are the least common. For training and testing split, we used the same split specified by (Suresh et al., 2022) in the TalkMoves dataset. Each teacher utterance in the TalkMoves dataset is annotated with one of six different teacher talk moves and "None". These talk moves are broadly classified into three categories based on their instructional purpose (Resnick et al., 2018): (1) accountability to the learning community, (2) accountability to content knowledge, and (3) accountability to rigorous thinking. See Table 1 for a brief description of each talk move, along with examples.

4.2 Research Motivation

In this study, we began working with transformers to classify talk moves. Prior attempts using non-transformers architecture achieved lower performance (65% F1 compared to 76.32% F1 with transformers) (Suresh et al., 2019, 2021b). The fine-tuned Roberta model proposed in (Suresh et al., 2022) employed a input representation of student-teacher sentence pairs to combine any given teacher utterance with the immediately prior student utterance (Suresh et al., 2021b). In order to understand the gaps in this model's performance, (Suresh et al., 2022) conducted an error analysis using a confusion matrix to consider examples where the Talk-Moves models were underperforming and often generated misclassifications. An initial analysis of those examples revealed several instances where the actual real-world context for the misclassified teacher utterance extended beyond the current representation of the previous student utterance. For example, consider the following dialogue "Student: Yes; Teacher: What do you think?". With limited context, it seems unclear if the teacher was relating to what a student said earlier or trying to prompt them to think. This challenge of limited context from prior work motivated us to find new ways to add contextual information to the existing models in order to improve performance.

4.3 Context-addition experiments

Constraints on the number of sequences in vanilla transformers, such as BERT and Roberta, prevents the direct application of transformers where there is a reliance on long-term dependencies. For example, consider a classroom session where a teacher encourages student X to think based on what student Y said earlier in the session. Without the expanded dialogue context, it can be challenging for transformers (and even humans) to classify the utterances. If we could expand the representation of available information such that it included the entire classroom session, the transformers may be more likely to learn to establish the long-term dependencies across the focal utterances or tokens. Given the importance of local context (Kovaleva et al., 2019), our input representation was modified from student-teacher sentence pairs to a fixed-size window surrounding each teacher utterance. This adjusted representation is atypical compared to the recommended input for fine-tuning, where a unique token separates two sequences (i.e., [SEP] in Bert and </s> in Roberta) (Devlin et al., 2018; Liu et al., 2019). There is a general notion that fine-tuning multiple utterances with multiple separator tokens, while theoretically possible, is not likely to work well. This notion was motivated by vanilla transformers, which were originally pre-trained on individual sentences or sentence pairs. We challenge this assumption by including additional past and future utterances in our adjusted input representation (Figure 2).

To establish a baseline performance level and generate information regarding the impact of context in classifying talk moves, we began with a simple input representation that includes only the target teacher utterance without any additional context. The output layer was a softmax over seven classes i.e., the six talk moves and "none" (no talk move). We also reproduced results from prior work on Roberta-base (Suresh et al., 2022). Following that, we experimented with three context setups: previous-only utterances, subsequent-only utterances, and both previous and subsequent utterances (equal numbers of each). In each setup, we evaluated several different window sizes. For example, the previous-only condition with a window size of three would have the immediately previous three utterances (with student(s) and/or the teacher as the speakers) serving as context cues for classifying the target utterance. If there was no prior utterance (such as at the start of a classroom session), we prepended empty strings. Similarly, given the previous and subsequent utterances condition with a window size of two, the target utterance would have two previous utterances prepended to the left and two subsequent utterances appended to the right. Separator tokens differentiated all of the utterances. As an additional preprocessing step, all utterances were truncated to 30 tokens long. The choice of truncation length was decided based on the distribution of sequence length (number of tokens) for all utterances in the dataset (see Figure 3). A token size of 30 accounted for more than 95% of the utterances in the dataset (two standard deviations from

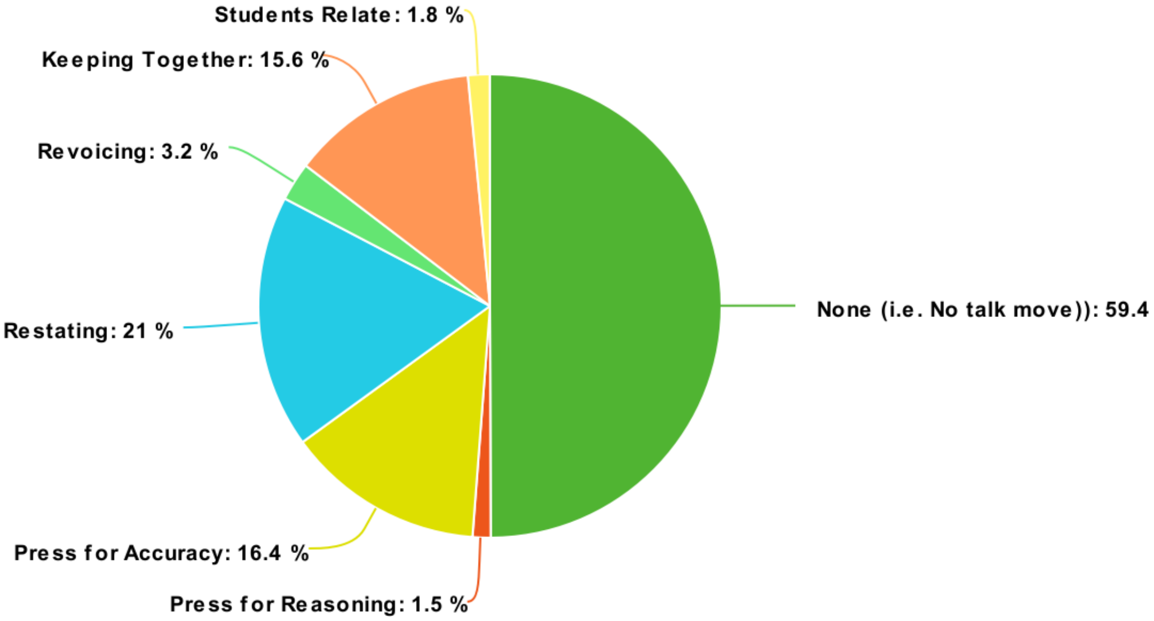

Figure 1: Distribution of teacher talk moves in the TalkMoves dataset

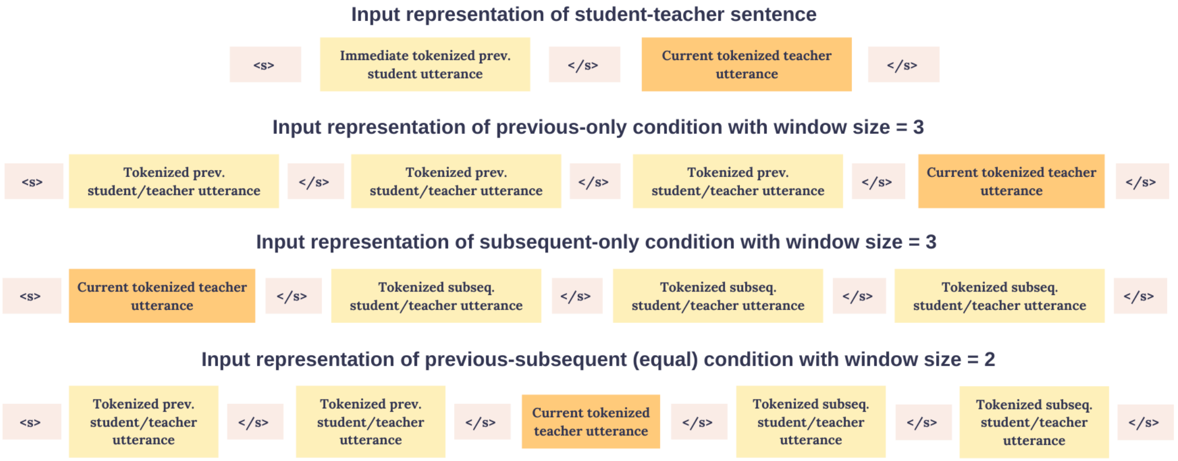
Input representation of student-teacher sentence
<s>
Immediate tokenized prev. student utterance
</s>
Current tokenized teacher utterance
</s>
Input representation of previous-only condition with window size = 3
<s>
Tokenized prev. student/teacher utterance
</s>
Tokenized prev. student/teacher utterance
</s>
Tokenized prev. student/teacher utterance
</s>
Current tokenized teacher utterance
</s>
Input representation of subsequent-only condition with window size = 3
<s>
Current tokenized teacher utterance
</s>
Tokenized subseq. student/teacher utterance
</s>
Tokenized subseq. student/teacher utterance
</s>
Tokenized subseq. student/teacher utterance
</s>
Input representation of previous-subsequent (equal) condition with window size = 2
<s>
Tokenized prev. student/teacher utterance
</s>
Tokenized prev. student/teacher utterance
</s>
Current tokenized teacher utterance
</s>
Tokenized subseq. student/teacher utterance
</s>
Tokenized subseq. student/teacher utterance
</s>

Figure 2: Modifying the input representation to support additional previous and subsequent utterances

the mean of the sequence length of seven tokens). We then fine-tuned transformers on the TalkMoves training set with different parameters using Amazon EC2 instances. We followed the recommended parameters from (Suresh et al., 2019, 2022) including learning rate (2e-5, 3e-5, 4e-5, 5e-5), number of epochs (3-6), batch size (4,8,16,32), warmup steps (0,100,1000) and maximum sequence length (512 for Roberta-like models) and (512,1024 for Longformer and BigBird). The performance on the testing set after fine-tuning is reported based on F1 measures and MCC (Suresh et al., 2021a). These measures work well for skewed datasets like TalkMoves (Chicco and Jurman, 2020; Suresh et al., 2021b). The code was implemented in Python 3.8

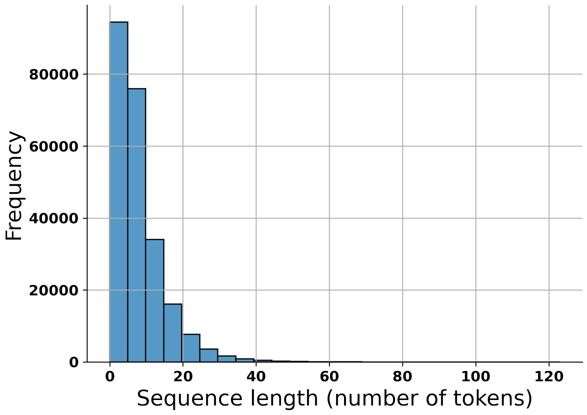

Figure 3: Number of utterances (frequency) vs sequence length (number of tokens) in TalkMoves dataset

Table 1: Teacher talk moves from TalkMoves dataset (Suresh et al., 2022)

Category	Talk move	Description	Example
Teacher Talk Moves			
Learning Community	Keeping everyone together	Prompting students to be active listeners and orienting students to each other	"What did Eliza just say her equation was?"
Learning Community	Getting students to relate to another's ideas	Prompting students to react to what a classmate said	"Do you agree with Juan that the answer is 7/10?"
Learning Community	Restating	Repeating all or part of what a student said word for word	"Add two here."
Content Knowledge	Pressing for accuracy	Prompting students to make a mathematical contribution or use mathematical language	"Can you give an example of an ordered pair?"
Rigorous Thinking	Revoicing	Repeating what a student said but adding on or changing the wording	"Julia told us she would add two here."
Rigorous Thinking	Pressing for reasoning	Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations	"Why could I argue that the slope should be increasing?"

with Pytorch and HuggingFace library (Wolf et al., 2019). In addition to the context-addition experiments with Roberta-base, we fine-tuned similar transformers architectures. XLNet, Longformer and BigBird are transformer architectures which support longer sequences. Since the TalkMoves dataset is composed of utterances from K-12 mathematics classrooms, we fine-tuned MathBERT, a pretrained architecture with focus on mathematical formula understanding.

5 Results

In this section, we present the results from our experiments that involved providing additional context to transformers to support the process of learning long-term dependencies. The experiments were repeated with ten random seeds, and the average score is reported (Table 2, 3). For brevity, we report performance only on Roberta-base (the best performing model from (Suresh et al., 2021b) as indicated in the first column of (Table 2) and transformers such as Longformer and Bigbird (Table 3). All the models are Base models (Large models are beyond the scope of this work). In the second column, we describe the context that was provided to the target teacher utterance for classification. For example, Previous 1 should be interpreted as a single previous utterance prepended to the target teacher's utterance. Similarly, Subsequent 1 should be interpreted as a single subsequent utterance appended to the target utterance. The third and final column describes the performance of the testing set.

For imbalanced datasets like TalkMoves, the Matthew Correlation Coefficient (MCC) and F1 measure are good indicators of model performance. An MCC score of +1 indicates a perfect correlation while 0 indicates a random correlation and -1 indicates a negative correlation. Similarly, the F1 score ranges from 0-100% where 100% indicates perfect performance. We begin with the No-Context condition which achieved a performance of 71.93% F1. On prepending the immediately prior or subsequent student utterance, the model achieved a performance of 76.32% F1 (Suresh et al., 2022). Next we turn to results from various context conditions with different window sizes followed by results from Longformer, BigBird, and other models. The maximum sequence length in most of these models was 512 with the exception of Longformer and Bigbird which had a sequence length upto 1024. The results presented in this work are comprehensive but not exhaustive since training and testing for all possible models and parameters is infeasible.

The results table clearly illustrates the importance of context in enhancing performance. Starting with Roberta-Base, the performance on the previous-only condition gradually increased with an increase in window-size and saturated for larger window-sizes. Similarly, we observed an improvement in performance for the subsequent-only condition. However, we did not see any significant improvement for larger window-sizes in this condition, possibly due to the negative impact in performance on "Revoicing" and "Restating" which rely on immediately prior student sentences. Moreover, the combination of previous and subsequent utterances resulted in the best performing model. The performance gradually increased proportionally with a window size up to 7 before saturating. Likewise, the performance on Longformer, XLNet and BigBird were comparable with similar input representation. The most surprising result was the performance on MathBert which was signficantly lower than other models. In summary, Roberta-Base with equal previous-subsequent condition ($size$ =7) outperformed rest of the models and constitutes the state-of-the-art results.

The primary motivation of the error analysis using a confusion matrix was to improve the performance on the under-performing talk move categories and identify patterns among the misclassfied utterances to be leveraged as features for the models. When comparing the confusion matrix from prior work (Suresh et al., 2022) (see Table 4), the current study shows a significant improvement in performance across all the teacher talk moves labels except "Restating" (see Table 5). With "Restating", we hypothesize that the decrease in performance was a result of supplementing additional context. Further analysis has to be performed in order to validate this claim.

6 Discussion

Based on the results from our experiments to improve the performance of a talk moves classifier using transformers, it is evident that longer dialog windows play an important role in differentiating talk moves. We successfully validated that the local discursive context is an important feature in classifying teacher talk moves. We generated a 4% F1 increase in performance when including a single additional utterance (either previous or subsequent) as compared to the no-context condition. Also, we observed that previous utterances are more impactful than future utterances for classifying talk moves. This finding is not surprising given that several talk moves, such as the teacher “restating” and “revoicing” what a student has already said, depend entirely on previous utterances as context. We also observed that context windows with a combination of previous and future utterances outperform either condition alone. Finally, we found that a window size of seven previous and subsequent utterances achieves the best performance. Beyond the identified size of seven, the performance decreases. It is possible that much earlier or much later utterances provide confusing or conflicting contextual information, which hinders model performance. It is equally likely that longer dialog windows could lead to overfitting.

Prior efforts to address the imbalanced nature of TalkMoves dataset through weighted loss resulted in reduced performance (Suresh et al., 2019). As an alternative, we attempted to generate synthetic samples of tokenized utterances through SMOTE (Synthetic Minority Oversampling Data) (Chawla et al., 2002). With SMOTE, it was challenging to retain the syntactic information of the generated examples. It was also difficult to generate the supporting contextual student and teacher utterances. Preliminary efforts did not yield any improvement in performance.

To further improve the performance, we have identified two future directions that appear worthwhile to consider: (1) experimenting with punctuation and other linguistic markers in the existing TalkMoves dataset and (2) collecting more training data. In the TalkMoves dataset, all the punctuation and other non-alphanumeric characters from the teacher and student utterances were removed. These text processing steps are typical for most text-based NLP applications to produce text that closely aligns with the output of Automated Speech Recognition (ASR) systems. However, we hypothesize that punctuation could play a significant role in differentiating one talk move from another. For example, “Agreed?” with a question mark can be considered an instance of “Keeping everyone together” whereas “Agreed” as a statement would be an instance of “None.” It remains to be determined the extent to which including punctuation markers might impact the performance of the models. Similarly, we can try incorporating speaker turns to indicate a student or teacher turn in previous and subsequent utterances as additional features to the model.

Table 2: Robert-Base performance with different window sizes

Model	Context	MCC	F1 (%)
Roberta-Base	No Context	0.7003	71.93
Roberta-Base	Immediate Student (Suresh et al., 2022)	0.7513	76.32
Roberta-Base	Previous 1	0.7460	76.01
Roberta-Base	Previous 5	0.7579	76.79
Roberta-Base	Previous 10	0.7615	77.08
Roberta-Base	Previous 15	0.7688	77.63
Roberta-Base	Previous 17	0.7657	77.35
Roberta-Base	Subsequent 1	0.7232	74.16
Roberta-Base	Previous 1 - Subsequent 1	0.7687	78.18
Roberta-Base	Previous 2 - Subsequent 2	0.7742	78.49
Roberta-Base	Previous 3 - Subsequent 3	0.7764	78.66
Roberta-Base	Previous 5 - Subsequent 5	0.7739	78.36
Roberta-Base	**Previous 7 - Subsequent 7**	**0.7805**	**78.92**
Roberta-Base	Previous 8 - Subsequent 8	0.7802	78.86

Table 3: Performance on classification of teacher talk moves on other models

Model	Context	MCC	F1 (%)
Roberta-Base	**Previous 7 - Subsequent 7**	**0.7805**	**78.92**
MathBERT	Previous 7 - Subsequent 7	0.6890	70.18
XLNet	Previous 7 - Subsequent 7	0.7709	78.06
Longformer	Previous 7 - Subsequent 7	0.7752	78.47
BigBird	Previous 7 - Subsequent 7	0.7694	77.89
BigBird	Previous 10 - Subsequent 10	0.7603	77.11

Another option that warrants consideration is supplementing data for the purpose of model pretraining. TalkMoves dataset (github.com/SumnerLab/TalkMoves) is a relatively small dataset for pretraining transformers when compared to Roberta which was pretrained on millions of data points. At the same time, we recognize the challenge in the collecting and annotating thousands of classroom transcripts. Moreover, there are important privacy concerns and other ethical considerations, given that these data involve minors, use proper names (which can be critical information for talk moves classification), and can be challenging to access in large quantities. We could potentially explore active learning to achieve greater accuracy with limited samples (Settles, 2009). Active learning is often sought as an option in machine learning applications where unlabeled instances are abundantly available (Schröder et al., 2021).

7 Conclusion

Documenting consequential elements of classroom instruction and providing teachers with feedback on their practices are critical endeavors in the education field. Taking into consideration the strong need to provide reliable feedback to teachers on productive classroom discourse, we need robust models to automatically classify teacher talk moves with high reliability. In this paper, we report on a number of experiments that involved providing longer dialog windows to the transformers in an effort to improve model performance. Based on these experiments, we generated a state-of-the-art 2.6% F1 improvement in performance (78.92% F1) over the previous models, primarily by adding a set number of previous and subsequent utterances to the input representation. Clearly, there are both challenges and opportunities for the development of innovative uses of AI techniques, particularly as they can be incorporated into tools that support teacher and student learning. The findings from this research open new avenues for exploration that can benefit both the education and NLP communi-

Table 4: Confusion matrix from Roberta-Base with Immediate student utterance as context

Roberta-Base (Immediate Student)		Actual							Precision	Recall	F1
0 - None	Predicted	42786	1779	67	54	232	1091	74	0.93	0.93	0.934
1 - Keeping Everyone together		1599	6549	106	139	99	518	30	0.73	0.72	0.73
2 - Getting students to relate		171	177	715	0	2	120	33	0.71	0.59	0.64
3 - Restating		112	18	3	932	21	12	0	0.79	0.85	0.82
4 - Revoicing		562	72	2	47	1063	44	0	0.72	0.59	0.62
5 - Pressing for accuracy		762	367	105	9	51	8289	669	0.82	0.86	0.84
6 - Pressing for reasoning		56	6	315	1	1	86	753	0.79	0.82	0.80

Table 5: Confusion matrix from Roberta-Base with Previous-7 and Subsequent-7 utterances as context. Compared to Table 4, we see an improvement in F1 score for almost all of the talk moves except Restating.

Roberta-Base (Previous 7 - Subsequent 7)		Actual							Precision	Recall	F1
0 - None	Predicted	14594	522	42	40	122	312	16	0.94	0.93	0.94
1 - Keeping Everyone together		512	2321	53	26	26	130	4	0.77	0.76	0.76
2 - Getting students to relate		31	23	206	0	0	37	9	0.64	0.67	0.65
3 - Restating		25	8	1	263	7	2	0	0.73	0.86	0.79
4 - Revoicing		179	24	0	25	326	7	1	0.66	0.58	0.62
5 - Pressing for accuracy		207	112	21	5	12	2678	41	0.84	0.87	0.85
6 - Pressing for reasoning		8	2	1	0	0	27	242	0.77	0.86	0.82

ties who might adopt our methods in applications where the local context may prove critical to improving performance.

Acknowledgements

The research team would like to thank Eddie Dombower and his team at Curve 10 for their contributions to the design and implementation of the TalkBack application. This material is based upon work supported by the National Science Foundation under Grant Numbers 1600325 and 1837986. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

David Adamson, Colin Ashe, Hyeju Jang, David Yaron, and Carolyn Penstein Rosé. 2013. Intensification of group knowledge exchange with academically productive talk agents. In *CSCL (1)*, pages 10–17.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Melissa Boston. 2012. Assessing instructional quality in mathematics. *The Elementary School Journal*, 113(1):76–104.

Amber G Candela, Melissa D Boston, and Juli K Dixon. 2020. Discourse actions to promote student access. *Mathematics Teacher: Learning and Teaching PK-12*, 113(4):266–277.

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.

Gaowei Chen, Carol KK Chan, Kennedy KH Chan, Sherice N Clarke, and Lauren B Resnick. 2020. Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. *Journal of the Learning Sciences*, 29(4-5):642–680.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Richard Correnti, Mary Kay Stein, Margaret S Smith, James Scherrer, Margaret McKeown, James Greeno, and Kevin Ashley. 2015. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. *Socializing Intelligence Through Academic Talk and Dialogue. AERA*, 284.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Meghan E Dale, Amanda J Godley, Sarah A Capello, Patrick J Donnelly, Sidney K D'Mello, and Sean P Kelly. 2022. Toward the automated analysis of teacher talk in secondary ela classrooms. *Teaching and Teacher Education*, 110:103584.

Dorottya Demszky, Jing Liu, Heather C Hill, Dan Jurafsky, and Chris Piech. 2021. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 218–227. ACM.

Beth A Herbel-Eisenmann. 2017. *Mathematics Discourse in Secondary Classrooms: A Practice-based Resource for Professional Learning: Facilitator Guide*. Math Solutions.

Cindy E Hmelo-Silver, Clark A Chinn, Angela M O'Donnell, and Carol Chan. 2013. The international handbook of collaborative learning.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*.

Emily Jensen, Meghan Dale, Patrick J Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hugh Mehan. 1979. *Learning lessons*. Harvard University Press Cambridge, MA.

Sarah Michaels and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue*, pages 347–362.

Sarah Michaels, Mary Catherine O'Connor, Megan Williams Hall, and Lauren B Resnick. 2010. Accountable talk® sourcebook. *Pittsburg, PA: Institute for Learning University of Pittsburgh. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF.*

Catherine O'Connor and Sarah Michaels. 2019. Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, 97:166–175.

Catherine O'Connor, Sarah Michaels, and Suzanne Chapin. 2015. Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. *Socializing intelligence through academic talk and dialogue*, pages 111–126.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*.

Samuel L Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela EB Stewart, Jessica Andrews-Todd, and Sidney K D'Mello. 2021. Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. *International Educational Data Mining Society*.

Lauren B Resnick, Christa SC Asterhan, and Sherice N Clarke. 2018. Accountable talk: Instructional dialogue that builds the mind. *Geneva, Switzerland: The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the United Nations Educational, Scientific and Cultural Organization (UNESCO)*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2021. Uncertainty-based query strategies for active learning with transformers. *arXiv preprint arXiv:2107.05687*.

E Schukat-Talamazzini, T Kuhn, and H Niemann. 1994. Speech recognition for spoken dialogue systems. In *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop, PAI*, volume 1, pages 110–120.

Burr Settles. 2009. Active learning literature survey.

Yu Song, Shunwei Lei, Tianyong Hao, Zixin Lan, and Ying Ding. 2021. Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, 59(3):496–521.

Abhijit Suresh, Jennifer Jacobs, Charis Clevenger, Vivian Lai, Chenhao Tan, James H Martin, and Tamara Sumner. 2021a. Using ai to promote equitable classroom discussions: The talkmoves application. In *International Conference on Artificial Intelligence in Education*, pages 344–348. Springer.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *13th International Conference on Language Resources and Evaluation (LREC 2022)*.

Abhijit Suresh, Jennifer Jacobs, Vivian Lai, Chenhao Tan, Wayne Ward, James H Martin, and Tamara Sumner. 2021b. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. *AAAI 2021 Spring Symposium on Artificial Intelligence for K-12 Education*.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9721–9728.

Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education*, 87:309–325.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children*, 23(3):34–41.

Noreen M Webb, Megan L Franke, Marsha Ing, Angela Chan, Tondra De, Deanna Freund, and Dan Battey. 2008. The role of teacher instructional practices in student collaboration. *Contemporary educational psychology*, 33(3):360–381.

Mikyung Kim Wolf, Amy C Crosson, and Lauren B Resnick. 2005. Classroom talk for rigorous reading comprehension instruction. *Reading Psychology*, 26(1):27–53.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.