# What Would a Teacher Do? Predicting Future Talk Moves

**Ananya Ganesh, Martha Palmer** and **Katharina Kann**
University of Colorado Boulder
{ananya.ganesh, martha.palmer, katharina.kann}@colorado.edu

## Abstract

Recent advances in natural language processing (NLP) have the ability to transform how classroom learning takes place. Combined with the increasing integration of technology in today's classrooms, NLP systems leveraging question answering and dialog processing techniques can serve as private tutors or participants in classroom discussions to increase student engagement and learning. To progress towards this goal, we use the classroom discourse framework of academically productive talk (APT) to learn strategies that make for the best learning experience. In this paper, we introduce a new task, called future talk move prediction (FTMP): it consists of predicting the next talk move – an utterance strategy from APT – given a conversation history with its corresponding talk moves. We further introduce a neural network model for this task, which outperforms multiple baselines by a large margin. Finally, we compare our model's performance on FTMP to human performance and show several similarities between the two.

## 1 Introduction

The field of natural language processing (NLP) has made rapid progress over the last few years (Wang et al., 2019). Success on natural language understanding, dialogue generation, and question answering tasks has spurred advances in NLP-based systems for educational applications. (McNamara et al., 2013; Litman, 2016; Burstein et al., 2020). Systems that can simulate human teachers in specific situations such as small-group discussions have the potential to aid learning by promoting student engagement.

Research has shown that deep conceptual learning is heightened when students are active participants in the classroom and contribute to discussions with their questions and ideas (McNamara, 2011; Bransford et al., 1999). However, large class sizes

| Scenario: Learning about proportional relationships in a classroom. The teacher gives an example of toasting two slices in a toaster, for 2 minutes. |
| --- |
| **Teacher:** So we've just seen that 2 slices of toast gets done in 2 minutes. (*None*) <br> **Teacher:** What if I had 3 slices of toast? (*Press for Accuracy*) <br> **Student:** 4 minutes! (*Wait*) <br> **Teacher:** Why would it take 4 minutes? (*Press for Reasoning*) <br> **Student:** Because you'd have to use the toaster twice. (*Wait*) |
| *FTMP: Getting Students to Relate* <br> *(e.g., who else agrees it would be 4?)* |

Table 1: Our proposed FTMP task; the teacher talk move corresponding to each utterance is shown in parentheses.

often make it difficult for all students to actively participate. Discussions in sub-groups increase each student's speaking time, but, in turn, make it impossible for a single teacher to guide all individual conversations. In this paper, we present a first step towards a system that can solve this problem by taking the teacher's role in facilitating sub-group discussions.

For this, we turn to a classroom discourse framework by Michaels et al. (2008) called academically productive talk (APT). This framework, which we describe in detail in Section 2, provides both teachers and students with a set of *talk moves* – a family of utterance strategies to use for productive and respectful in-class discussions. As a first step towards developing an NLP system that can guide academically productive discussions, we aim to design a model which can predict when which specific talk move is appropriate. Thus, we introduce the task of **future talk move prediction (FTMP)** – given a conversation history, the goal is to predict what the *next* teacher talk move should be. We formulate this as a multi-class classification problem, with the input being a sequence of previous utterances

and their corresponding talk moves, and the label being the next talk move.

We further propose a model for FTMP, which we call 3-E.[1] It consists of three recurrent neural network (RNN) encoders: one for individual utterances, one for utterance sequences, and one for talk move sequences. The model is trained on transcripts of classroom discussions where teacher utterances have been annotated for the talk moves they represent. We consider the actions of the teacher to be our gold standard data for FTMP. We show that our model strongly outperforms multiple baselines and that adding sentence representations from RoBERTa (Liu et al., 2019) or TOD-BERT (Wu et al., 2020) – a model trained on task-oriented dialogue – does not increase performance further.

Finally, we investigate the performance of human annotators on FTMP. Unlike the teacher, they do not have access to multi-modal signals, subject matter information, or knowledge of student behavior. This setting, which mimics the information available to our model, is significantly different from the teachers who generate the gold standard utterances captured in our data. We present a detailed analysis of their performance on a diagnostic test set, and highlight similarities to our model's performance. Our findings indicate that our model produces acceptable predictions a majority of the time. However, a gap between model and human performance on this task shows that there is still room for improvement.

## 2 Academically Productive Talk

In this section, we provide an overview of the APT discourse framework and introduce a new task within the broader research area of NLP for educational applications: FTMP.

### 2.1 Background on APT

Research in cognitive science and psychology highlights the importance of active participation as opposed to passive listening for achieving deep conceptual learning (McNamara, 2011; Bransford et al., 1999; Chi and Wylie, 2014). This can take the form of reflection on the lesson, as well as generation of new ideas, such as asking and answering questions, connecting concepts, and coming up with explanations. Chapin et al. (2009); Goldenberg (1992); Cazden (1988) discuss the importance

of classroom conversations in this process. Chapin (2003) present case studies that show how implementing structured discussions in classrooms over a period of two years results in measurable improvements in test scores in mathematics.

To formalize how such discussions can be facilitated, Michaels et al. (2008) present a classroom discourse framework called *academically productive talk* (APT; also called *accountable talk*). This includes strategies that teachers and students can use to promote engagement as well as deep conceptual learning through discussions.

**Facets** Michaels et al. (2008) present three facets of accountability that APT encompasses: *accountability to the learning community*, *accountability to standards of reasoning*, and *accountability to knowledge*. The first facet emphasizes the importance of listening to other students' contributions, and, subsequently, building on top of them. The second facet promotes talk that is based on evidence and reasoning, and involves getting students to provide explanations for their claims. The last facet covers talk which involves factual knowledge – such as introducing a new concept, or challenging a student's claim to correct misconceptions.

**Teacher Talk Moves** Michaels and O'Connor (2015) conceptualize the above facets as "tools" that can be used by teachers and students to engage in APT. For both teachers and students, these tools take the form of utterance strategies called *talk moves*, which they can employ in order to conduct meaningful discussions.

In this paper, we focus on the following six talk moves used by teachers: (1) ***Keeping Everyone Together*** refers to utterances that manage student interactions, and asks students to be active listeners; (2) ***Getting Students to Relate*** refers to utterances that ask a student to build on other students' ideas by agreeing, disagreeing, or following up; (3) ***Restating*** occurs when a teacher repeats a student's answer or claim verbatim with the purpose of ensuring it reaches the entire classroom; (4) ***Revoicing*** happens when a teacher paraphrases a student's ideas, but adds or removes information in order to correct a student or convey new knowledge; (5) ***Pressing for Reasoning*** refers to utterances that ask a student to explain a decision or to connect multiple ideas; and (6) ***Pressing for Accuracy*** refers to utterances that prompt for answers to a factual question, e.g., about a method or

---

[1]Code for all models is available at https://nala-cub.github.io/resources/

| Talk Move | Description | Example |
|---|---|---|
| Keeping Everyone Together | Ask students to be active listeners | Raise your hand if you know the answer |
| Getting Students to Relate | Ask students to contribute to another's ideas | Do you or agree or disagree with Michael? |
| Revoicing | Repeat what a student says with adding words or rephrasing | S: It had two T: So it had two edges |
| Restating | Repeat what a student says verbatim | S: Hexagon T: Hexagon! |
| Press for Accuracy | Prompt for an answer | What is this called? |
| Press for Reasoning | Prompt for explanation of thinking | How did you decide? |
| *None* | *Fits into none of the above* | *Good morning* |
| *Wait* | *Teacher says nothing while student speaks* | *S: It's the same shape* |

Table 2: An overview of all teacher talk moves, their purpose and an example utterance. *None* and *Wait* are not APT talk moves, and represent generic utterances and teacher pauses during student utterances, respectively.

a result.

*Keeping Everyone Together, Getting Students to Relate*, and *Restating* are part of accountability to the learning community; *Revoicing* and *Press for Reasoning* are part of accountability to standards of reasoning, and *Press for Accuracy* falls under accountability to knowledge. Examples for all talk moves are shown in Table 2.

**Student Talk Moves**   While we do not focus on student talk moves in this work, we summarize them here for completeness. Student talk moves can also be grouped into the same accountability facets as teacher talk moves (O'Connor and Michaels, 2019). Under accountability to the learning community, we have *Relating to Another Student* – building on a classmate's ideas or asking questions about them, and *Asking for more info* – requesting help from the teacher on a problem. Under accountability to knowledge, there is *Making a Claim* – providing an answer or a factual statement about a topic. Under accountability to standards of reasoning, we have *Providing Evidence/Explanation* – explaining their thinking with evidence.

## 2.2   Future Talk Move Prediction

In order to build a system that can facilitate in-class discussion in the way a human would, we aim at automatically answering the question *What would a teacher do?* at each point within a classroom conversation. Specifically, we define the task of future talk move prediction (FTMP) as choosing the next appropriate teacher talk move to make, given the history of what has been discussed so far.

Formally, the input for FTMP is a dialogue context $C = c_0, c_1, ..., c_t$, with each context element consisting of an utterance $u_i$, a binary variable $s_i$ indicating if the speaker is different from the previous utterance, and a teacher talk move label $t_i$, i.e., $c_i = (s_i, u_i, t_i)$. The goal then is to predict the next teacher talk move $t_{t+1}$ out of the possible talk moves defined above. Note that the future utterances are unseen; the prediction of the next talk move is to be made only based on the conversation history.

## 3   Related Work

### 3.1   Promoting APT with NLP Systems

Ideas from APT have been incorporated with success into intelligent tutoring systems (Dyke et al., 2013; Tegos et al., 2016; Adamson et al., 2014). These systems provide an environment to simulate classroom discussions, for instance, as small groups collaboratively solving problems with a shared textual chat interface for communication. The intelligent agent then plays a role similar to a teacher – it monitors the conversations and makes decisions about when to intervene in order to pro-

mote student engagement and learning.

Adamson et al. (2014) study the effects of two specific interventions: using the *Revoicing* talk move as well as an *Agree/Disagree* talk move (which corresponds to the *Getting Students to Relate* talk move in our above categorization). These interventions are made by matching student utterances in the chat to an annotated set of concepts and misconceptions for the topic being taught. Through multiple case studies, they show that interventions by the agent have a positive effect on learning, as measured by test scores before and after using the system. The agent interventions also prove useful in increasing student talk frequency. Similarly, Tegos et al. (2015) find that an APT-based intervention called Linking Contributions, similar to *Getting Students to Relate*, improves explicit reasoning as well as learning outcomes in students.

The findings of the above work provide a strong motivation for building a conversational AI system that can produce academically productive talk. Unlike the above systems, which focused particularly on accountability to the learning community, we attempt to predict opportunities for intervention across all talk moves described in Section 2. Since we do not have access to gold annotations of statements corresponding to concepts and misconceptions, we make use of transcripts of classroom discourse with annotations for talk moves used by the teacher.

## 3.2 NLP for Educational Applications

Our work is a first step towards improving in-class discussions with the help of an NLP system and, thus, to improve student learning and engagement. Prior work in understanding classroom discourse using NLP includes Suresh et al. (2019) and Donnelly et al. (2016). They propose an application where feedback can be provided to teachers by automatically classifying their utterances into talk moves. Other applications of NLP to education include language learning assistance (Beatty, 2013; Carlini et al., 2014; Tetreault et al., 2014), writing assistance (Dahlmeier and Ng, 2011; Chollampatt and Ng, 2018; Chukharev-Hudilainen and Saricaoglu, 2016), and automated scoring (Burstein et al., 1998; Farag et al., 2018; Beigman Klebanov and Madnani, 2020).

## 3.3 Dialogue Systems

Our work is further related to research on dialogue systems. Similar to talk moves, dialogue acts provide a categorization for utterances, but, in contrast to talk moves, they apply to general-purpose conversations (Stolcke et al., 2000; Calhoun et al., 2010). Examples include *Statement, Question, Greeting*, and *Apology*. Dialogue act tagging, which is sometimes called dialogue act prediction, is the task of classifying an utterance into the category it belongs to (Yu and Yu, 2019; Khanpour et al., 2016; Wu et al., 2020). Analogous to FTMP, future dialogue act prediction is the task of predicting what the next dialogue act should be, given a conversation history (Tanaka et al., 2019).

Pretrained models have been successfully adapted to the task of dialogue generation (Zhang et al., 2020; Wu et al., 2020; Adiwardana et al., 2020; Roller et al., 2020). However, if directly used in the classroom, these models could potentially produce harmful or unsuitable dialogue as they are trained on large datasets comprising conversations from the internet (Bender et al., 2021). Additionally, we want a system to facilitate structured conversations, and not cause further diversions – this is in contrast to many task-oriented or open-domain dialogue systems whose purpose is to entertain and appear personable to the user. Hence, we propose FTMP as a crucial first step towards an NLP system capable of facilitating classroom discussions.

## 4 Model

In this section, we describe our proposed model for FTMP, cf. Figure 1. Following Tanaka et al. (2019)'s model for future dialogue act prediction, its main components are three encoders. We hence name our model **3-E**. Our model predicts the next teacher talk move $t_{t+1}$, given the last $w$ context elements $c_{t-w+1}, \ldots, c_t$.

**Utterance Encoder** The first encoder – the utterance encoder – is a single-layer gated recurrent unit (GRU; Cho et al., 2014). It processes the sequence of vector representations $v(w_1), \ldots, v(w_m)$ of the words $w_1, \ldots, w_m$ that each utterance $u_i$ consists of and computes the last hidden state as a vector representation of $u_i$:

$$\hat{a}_i = \text{GRU}(v(w_1), \ldots, v(w_m)) \qquad (1)$$

Each utterance representation is then concatenated with a representation $s_i$ of the speaker role. This representation is either 1 or 0, depending on if the speaker has changed from the previous utterance:

$$a_i = \text{cat}(\hat{a}_i, s_i) \qquad (2)$$

Figure 1: Our proposed model for FTMP, consisting of separate encoders for utterances, past task moves, and the overall context.

**Dialogue Encoder** Next, the sequence of all $w$ utterance representations is passed to the dialogue encoder, which is also a single-layer GRU. The dialogue encoder processes the sequence, and we take the last hidden state as a representation of all utterances within our context window:

$$b_t = \text{GRU}(a_{t-w+1}, \dots, a_t) \qquad (3)$$

**Talk Move Encoder** The talk move encoder is a third single-layer GRU, which encodes the sequence of vector representations $v(t_{t-w+1}), \dots, v(t_t)$ of talk moves $t_{t-w+1}, \dots, t_t$:

$$d_t = \text{GRU}(v(t_{t-w+1}), \dots, v(t_t)) \qquad (4)$$

We obtain our final context representation $r_t$ by concatenating the representation of all utterances and all talk moves within the context window:

$$r_t = \text{cat}(b_t, d_t) \qquad (5)$$

Finally, we pass $r_t$ through a two-layer feed-forward network and a softmax layer to obtain a probability distribution over possible future talk moves.

### 4.1 Adding a Pretrained Sentence Encoder

**RoBERTa** Pretrained models define the state of the art on a large variety of NLP tasks (Wang et al., 2019). Thus, we additionally experiment with concatenating an utterance representation computed by RoBERTa (Liu et al., 2019) to the output of 3-E's utterance encoder. Equation 2 then becomes:

$$a_i^* = \text{cat}(\hat{a}_i, s_i, \text{RoBERTa}(w_1, \dots, w_m)) \qquad (6)$$

We call the model with additional RoBERTa representations **3-E-RoBERTa**.

**TOD-BERT** Since there is a domain mismatch between the text that RoBERTa is trained on and our data, we further experiment with including a model trained on task-oriented dialogue, called TOD-BERT (Wu et al., 2020). TOD-BERT differentiates between user utterances and system utterances using two special tokens, [USR] and [SYS]. Correspondingly, we use the [USR] token to indicate student utterances and the [SYS] token to indicate teacher utterances. We then concatenate a context of $w$ utterances, marked by speaker tokens when there is a change in speaker, to obtain $c_{tod}$. Finally, we encode $c_{tod}$ using the pretrained TOD-BERT model and concatenate it with the output of the dialogue encoder and talk move encoder. Equation 5 then becomes:

$$r_t = \text{cat}(b_t, d_t, \text{TOD-BERT}(c_{tod})) \qquad (7)$$

We call this model **3-E-TOD-BERT**. When pretrained sentence encoders are used, we use the respective BPE (Sennrich et al., 2016) tokenizer for each model.

### 4.2 Computing the Loss

We train all models using a cross-entropy loss. However, we observe a strong class imbalance in our training data, cf. Figure 2. Thus, we compute label weights inversely proportional to the frequency of a label's occurrence in the data and use them to weight the loss for each training example.

## 5 Experiments

### 5.1 Dataset

For our experiments, we make use of the dataset from Suresh et al. (2019). It consists of 216 annotated transcripts of classroom discourse collected in public schools in the US. The topic of instruction is mathematics. The transcripts have been collected from classes from kindergarten to grade 12 and are all in English. Each row in the transcripts consists of an utterance, the name of the speaker, and the talk move realized by this utterance.

The annotations assign each teacher utterance to one of the 6 APT talk moves described in Section 2. Utterances that do not fit into any talk move category are coded as *None*. In addition, we designate the teacher talk move corresponding to utterances made by a student as *Wait*. This category is needed as we eventually want to be able to detect when an in-class NLP system should remain quiet. The original annotations contain two additional categories that we remove due to sparsity: $Marking$ refers to repeated utterances, and we merge it with the $Restating$ category. Some student utterances are annotated as $Context$, which we merge with the $Wait$ category.

We create training, development and test data from $70\%$, $15\%$, and $15\%$ of the available documents, respectively. Thus, we have 151 documents for training, and 32 documents for each of development and testing. Our training set consists of over 63k utterances, and the distribution of talk moves in the training set is shown in Figure 2.

Since 3-E's utterance encoder operates on the word level, we split each utterance into words using the NLTK word tokenizer (Loper and Bird, 2002).

### 5.2 Baselines

We compare our model to three baselines.

**Random Baseline (RB)**   This baseline randomly selects one of the 8 talk moves for each input.

**Talk Move Bigram Model (TMBM)**   For this baseline, we compute the conditional probability of every talk move in the training set, given the talk move realized by the previous utterance. We then pick the talk move with the highest conditional probability.

**Talk Moves Only (TM-only)**   We further train a GRU model exclusively on the sequences of prior



Figure 2: Distribution of talk moves in our training set.

talk moves, i.e., this baseline has no access to actual utterances. We implement two variants of this baseline, one with class weights for training (TM-only-w), and one without (TM-only-z).

### 5.3 Metrics

Since the classes in our dataset are highly imbalanced, we do not evaluate using accuracy. Instead, we report precision, recall, and F1 score for all models. We compute F1 for all 8 classes individually, and additionally calculate macro-average F1 as an overall score for our dataset.

### 5.4 Results

Table 2 shows the performance of our proposed model 3-E as well as of 3-E-TOD-BERT, 3-E-RoBERTa, and all baselines. Looking at the macro-average F1 scores, we see that 3-E performs best with an F1 of 29.84. 3-E-RoBERTa, with 27.62 F1, performs worse; however, given that this model has more parameters and includes a strong pretrained component, this is an unexpected result.[2]

To reduce the domain mismatch between RoBERTa's training data and our classroom dialogue data, we substitute RoBERTa with TOD-BERT, which is also trained on dialogue. We see that, while 3-E-TOD-BERT performs better than 3-E-RoBERTa, 3-E still outperforms it. We also observe that on most individual talk move classes, 3-E-TOD-BERT performs equal to or better than 3-E. However, it does poorly on a few classes that

---

[2] We further experiment with directly finetuning RoBERTa on our task, but find its performance to be poor overall (around 17 F1). Hence, we do not report detailed results.

| Model | Prec. | Recall | F1 | None F1 | Wait F1 | Press Acc. F1 | Keep Together F1 | Revoicing F1 | Getting Students to Relate F1 | Restating F1 | Press Reasoning F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3-E | 35.67 | 30.38 | **29.84** | 72.72 | 75.70 | 24.25 | 13.31 | 20.27 | **3.25** | **18.45** | **10.77** |
| 3-E-ToD-BERT | 31.10 | 28.92 | 28.51 | **73.05** | **77.67** | **25.18** | 13.81 | 18.89 | 0.00 | 17.92 | 1.53 |
| 3-E-RoBERTa | 33.81 | 28.04 | 27.62 | 69.89 | 73.34 | 20.48 | **14.68** | 20.31 | 1.57 | 17.81 | 2.88 |
| TM-only-w | 28.95 | 22.66 | 20.40 | 72.14 | 52.75 | 13.19 | 1.94 | **21.04** | 0.68 | 0.00 | 1.45 |
| TM-only-z | 18.38 | 18.81 | 16.43 | 71.60 | 52.14 | 0.22 | 0.00 | 7.47 | 0.00 | 0.00 | 0.00 |
| RB | 12.25 | 11.74 | 8.50 | 15.50 | 19.82 | 10.52 | 2.93 | 2.28 | 2.81 | 11.58 | 2.57 |
| Majority | 6.46 | 12.50 | 8.52 | 68.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| TMBM | 13.18 | 17.74 | 15.09 | 49.59 | 71.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Annotator 1* | *37.57* | *29.29* | *24.50* | *31.46* | *42.55* | *15.69* | *18.18* | *29.51* | *5.13* | *53.52* | *0.00* |
| *Annotator 2* | *38.99* | *33.16* | *30.51* | *25.19* | *54.17* | *36.17* | *20.00* | *24.66* | *21.74* | *42.62* | *9.52* |
| *3-E* | *15.83* | *20.22* | *14.57* | *29.32* | *26.53* | *4.44* | *12.66* | *8.16* | *0* | *35.48* | *0.00* |

Table 3: Model and annotator performance on FTMP. Italics indicate results on a diagnostic test set of 300 examples taken from the development set.

| Label | Prec. | Recall | F1 |
|---|---|---|---|
| Macro average | 43.76 | 42.65 | 42.44 |
| None | 68.30 | 77.76 | 72.72 |
| Wait | 71.48 | 80.45 | 75.70 |
| Learning Community | 29.59 | 14.10 | 19.10 |
| Content Knowledge | 27.86 | 21.47 | 24.25 |
| Rigorous Thinking | 21.56 | 19.44 | 20.45 |

Table 4: Performance of 3-E evaluated on facets.

are less prevalent in the data. We hypothesize that small changes in the quality of the utterance representations have negligible effect on our model, since it gets a large amount of information from the sequence of prior talk move labels. This hypothesis is supported by the fact that all baselines which only receive prior talk move labels as input, i.e., TM-only-w, TM-only-z and TMBM, obtain F1 scores of 20.40, 16.42, and 15.09, respectively. All of them strongly outperform a random baseline with an F1 of 8.50. Comparing 3-E to our baselines, we see that our proposed model is indeed strong on FTMP: 3-E outperforms the best baseline, TM-only-w, by 9.44 F1.

# 6 Analysis

## 6.1 FTMP on the Facet Level

In some cases, the distinctions between different talk moves are subtle. For instance, both the *Keeping Everyone Together* and *Getting Students to Relate* moves, which fall under the facet of accountability to the learning community, are made when the teacher wants the students to actively listen and respond to statements made in the classroom. To understand how well the model can distinguish between different accountability facets, we evaluate our best model, 3-E, on the facet level by binning all predicted talk moves into their corresponding facets for the computation of the F1 score.

In Table 4, we see that performance goes up by 12.60 points in this setting, indicating that 3-E is able to distinguish between labels at a coarse-grained level, but struggles with fine-grained distinctions.

## 6.2 Window Size and Class Weights

We further investigate the effect of weighting the loss as described in Section 4.2 and the influence of different context window sizes. Full results on the development set can be found in Table 6 in the appendix, and we provide a summary of our findings here.

Varying the window size leads to small changes in F1. For smaller window sizes of 1 and 2, F1 is slightly lower at 28.64 and 27.62. When the window size is increased to 5, F1 increases to 29.83. However, when the window size is increased further to 7, F1 drops slightly, to 29.05. We thus choose a window size of 5 to train all our models, and conclude that very large window sizes are not beneficial. We hypothesize that this might be due to the most relevant information for FTMP being

Figure 3: Confusion matrix on the diagnostic test set for 3-E (left) and our two annotators (middle and right). The talk move labels in order are: *None, Wait, Press for Accuracy, Keeping Everyone Together, Revoicing, Getting Students to Relate, Restating*, and *Press for Reasoning*.

| Primary option: Annotators | |
|---|---|
| Inter-annotator agreement | 46% |
| Annotator 1–ground truth agreement | 29% |
| Annotator 2–ground truth agreement | 33% |
| Both Annotators–ground truth agreement | 17% |
| *Primary option: Model* | |
| Model–Annotator 1 agreement | 48% |
| Model–Annotator 2 agreement | 33% |
| Model–ground truth | 20% |
| *Acceptable options* | |
| Annotator 1's primary accepted by Annotator 2 | 94% |
| Annotator 2's primary accepted by Annotator 1 | 91% |
| Ground truth accepted by Annotator 1 | 72% |
| Ground truth accepted by Annotator 2 | 79% |
| Model predictions accepted by Annotator 1 | 90% |
| Model predictions accepted by Annotator 2 | 84% |

Table 5: Percentage agreement between our annotators, the ground truth, and the model's predictions on the diagnostic test set.

contained in the most recent dialogue history.

Further, class weighting during training increases 3-E's F1 score by 3, from 26.94 to 29.83. We thus, conclude that class weights are important to account for the label imbalance in our training set. [3]

### 6.3 Performance of Human Annotators

We further investigate (1) the difficulty of FTMP for human annotators, (2) the effect of multiple choices for the future talk move as opposed to a single answer, and (3) how annotator decisions differ from 3-E's predictions. While our gold standard

---

[3] We also experiment with downsampling the dominant classes and find its performance to be comparable to class weighting.

---

are actions of a teacher, who also represent human performance, an FTMP annotator is different from a teacher since the former is presented with the exact same information as our model. In contrast, we expect a teacher's decisions to be informed by background knowledge about the students, knowledge about the content being discussed, and multi-modal information.

We recruit two annotators who have extensive experience with linguistic annotation tasks, and are familiar with talk moves. We present them with a diagnostic test set of 300 examples from the development set. Similar to the model input, each example consists of the past 5 utterances, the corresponding talk moves, and speaker information. Both annotators then provide (1) the most likely future talk move, referred to as the 'primary' option, and (2) a set of all acceptable future talk moves given the conversation history. As with our modeling setup, we consider the ground truth for the primary option to be the talk move made by the teacher in the classroom transcript. Each talk move is equally distributed in the ground truth, with 37 examples each of talk moves *None, Wait, Restating, Revoicing*, and 38 examples each of the other talk moves.

**Primary Option** The last 3 rows in Table 3 show the performance of our annotators' primary option and the model on the diagnostic test set. There is a significant gap of 10 F1 and 15 F1 respectively between the performance of the model and the two annotators. However, there are similarities in the class-wise breakdown. Both the annotators and the model achieve a high F1 on the classes *None, Wait,*

and *Restating*, and perform poorly on *Press for Reasoning* and *Getting Students to Relate*. *Press for Reasoning*, and *Getting Students to Relate* are the least prevalent classes in the data. However, the annotators' performance on these talk moves suggests that these classes are intrinsically more difficult to predict based on conversational cues alone. On the other hand, the model's poor performance on categories like *Revoicing* and *Keeping Everyone Together* in comparison to the annotators indicates that there is still room for improvement for our model.

The similarities between the model's predictions and the annotators' primary option is further illustrated by the confusion matrices in Figure 3. Both the model as well as our annotators erroneously predict *None* when the true label is another category, both confuse *Restating* and *Revoicing* for each other, and both erroneously predict *Keeping Everyone Together* when the true category is *Getting Students to Relate* or *Press for Reasoning*.

**Acceptable Options**   Table 5 shows the percentage of responses for which the annotators agree with each other, the ground truth, and the model's predictions. On average, both annotators provide 3 acceptable options in addition to the primary – thus, roughly half the classes were viewed as acceptable for most examples. The impact of having a set of acceptable options in addition to a single correct option is evident here: while inter-annotator agreement is only around 46% on the primary option alone, the primary option of each annotator was one of the acceptable options by the other annotator in over 90% of the cases. Additionally, while agreement between the ground truth and the primary option is low with 29% and 33%, this increases to 72% and 79% when additional options are being considered.

Table 5 helps us contextualize our model's performance. Interestingly, the annotators agree with the predictions made by the model more often than they agree with the ground truth. This indicates that the model might truly be grasping overall patterns and cues from the training data, but probably struggles with finer-grained distinctions between the talk move classes. This is further substantiated by our analysis of how often the model's predictions featured in the set of acceptable options for each annotator. We find that the predictions were acceptable in 90% and 84% of all instances respectively for each annotator.

## 7   Conclusion

In this paper, we made use of the APT discourse framework to take a first step towards a system that can fill the role of a teacher in classroom discussions. We introduced the task of FTMP, which consists of predicting the next appropriate talk move given an in-class dialogue context. We then presented 3-E, a model for the task, which outperforms multiple baselines. Finally, we conducted an analysis of human performance on FTMP, and compared it to our model. Our results showed that, while the task is challenging, our model produces acceptable talk moves and can identify overall patterns, indicated by similarities with human performance.

## Ethics and Impact Statement

The data used in this paper was collected after obtaining informed consent from all participants, and this research was approved by the University of Colorado Boulder's Institutional Review Board (protocol #18-0432). All authors completed appropriate training prior to accessing the data. Since this dataset was provided to us by its owners purely for the purpose of the research reported in this paper, we are not making it publicly available here.

Further, this research is part of the NSF National AI Institute for Student-AI Teaming. Several learning scientists and educators are participating in the institute and advising us on avoiding unintended consequences and harms that could stem from NLP research. The work described in this paper is preliminary, and will be used to inform strategies for additional data collection and model design. Field trials in carefully controlled classroom settings will be conducted before wider deployment.

# References

David Adamson, Gregory Dyke, Hyeju Jang, and Carolyn Penstein Rosé. 2014. Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24(1):92–124.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Ken Beatty. 2013. *Teaching & researching: Computer-assisted language learning*. Routledge.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big.

John D Bransford, Ann L Brown, and Rodney R Cocking. 1999. *How people learn: Brain, mind, experience, and school*. National Academies Press.

Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch. 2020. Proceedings of the fifteenth workshop on innovative use of nlp for building educational applications. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Roberto Carlini, Joan Codina-Filba, and Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 1–12.

Courtney B Cazden. 1988. *Classroom discourse: The language of teaching and learning*. ERIC.

Suzanne H. Chapin. 2003. Classroom discussions: Using math talk to help students learn, grades 1-6.

Suzanne H Chapin, Catherine O'Connor, Mary Catherine O'Connor, and Nancy Canavan Anderson. 2009. *Classroom discussions: Using math talk to help students learn, Grades K-6*. Math Solutions.

Michelene TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Evgeny Chukharev-Hudilainen and Aysel Saricaoglu. 2016. Causal discourse analyzer: Improving automated feedback on academic esl writing. *Computer Assisted Language Learning*, 29(3):494–516.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 107–117.

Patrick J. Donnelly, Nathan Blanchard, Borhan Samei, Andrew M. Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K. D'Mello. 2016. Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, page 45–53, New York, NY, USA. Association for Computing Machinery.

Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. 2013. Towards academically productive talk supported by conversational agents. In *Productive multivocality in the analysis of group interactions*, pages 459–476. Springer.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Claude Goldenberg. 1992. Instructional conversations: Promoting comprehension through discussion. *The Reading Teacher*, 46(4):316–326.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021, Osaka, Japan. The COLING 2016 Organizing Committee.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diane Litman. 2016. Natural language processing for enhancing teaching and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.

Danielle S McNamara. 2011. Measuring deep, reflective comprehension and learning strategies: challenges and successes. *Metacognition and Learning*, 6(2):195–203.

Danielle S McNamara, Scott A Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, 45(2):499–515.

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, Catherine O'Connor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

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.

Koji Tanaka, Junya Takayama, and Yuki Arase. 2019. Dialogue-act prediction of future responses based on conversation history. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 197–202, Florence, Italy. Association for Computational Linguistics.

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.

Stergios Tegos, Stavros Demetriadis, Pantelis M Papadopoulos, and Armin Weinberger. 2016. Conversational agents for academically productive talk: A comparison of directed and undirected agent interventions. *International Journal of Computer-Supported Collaborative Learning*, 11(4):417–440.

Joel Tetreault, Martin Chodorow, and Nitin Madnani. 2014. Bucking the trend: improved evaluation and annotation practices for esl error detection systems. *Language Resources and Evaluation*, 48(1):5–31.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Dian Yu and Zhou Yu. 2019. MIDAS: A dialog act annotation scheme for open domain human machine spoken conversations. *CoRR*, abs/1908.10023.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

## A  Tuning of the Window Size

| Configuration | Prec. | Recall | F1 | Acc. |
|---|---|---|---|---|
| No weighting, window 5 | 32.31 | 25.67 | 26.94 | 63.44 |
| Class weighting, window 5 | 34.84 | 29.39 | 29.83 | 62.24 |
| Class weighting, window 1 | 31.17 | 29.18 | 28.64 | 62.61 |
| Class weighting, window 2 | 34.11 | 29.44 | 27.62 | 62.85 |
| Class weighting, window 3 | 29.83 | 29.29 | 29.09 | 58.07 |
| Class weighting, window 4 | 29.72 | 28.57 | 28.51 | 61.65 |
| Class weighting, window 6 | 32.55 | 30.56 | 28.95 | 63.76 |
| Class weighting, window 7 | 33.21 | 28.64 | 29.05 | 58.27 |

Table 6: Tuning experiments on the development set

## B  Training and Hyperparameters

3-E is implemented in PyTorch (Paszke et al., 2019). For training, we use an Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of $1e^{-4}$ and train for 30 epochs. The utterance encoder has an embedding size of 256, and a hidden layer size of 512. The talk move encoder has an embedding size of 32 and a hidden layer size of 64. The dialogue encoder has a hidden layer size of 1025. Finally, the feedforward layer uses a hidden layer size of 32. We do not use pretrained word embeddings in the 3-E model.

For the 3-E-RoBERTa model, we use the pretrained parameters of the Fairseq library's implementation of RoBERTa (Ott et al., 2019), and use representations with dimension 1024. For 3-E-RoBERTa, we further add dropout with a probability of 0.4 in two places to avoid over-fitting: on the layer where the two utterance representations are concatenated (c.f. Equation 6), and the layer where the utterance history and talk move history are concatenated (c.f. Equation 5).

For 3-E-TOD-BERT, we use the pretrained model provided by Wu et al. (2020) trained jointly on masked language modeling and response contrastive loss. We additionally add a dropout of 0.2 at the layer where the utterance representation is concatenated with talk move history (c.f. Equation 7).

The baseline models TM-only-w and TM-only-z are trained for 30 epochs using the same hyperparameters, with a batch size of 256 and a learning rate of 1e-4.