Future Generation Computer Systems 127 (2022) 194–207

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues

Jionghao Lin [a,b], Shaveen Singh [a,b], Lele Sha [a,b], Wei Tan [b], David Lang [c], Dragan Gašević [a,b,d,e], Guanliang Chen [a,b,*]

[a] Centre for Learning Analytics, Monash University, Australia
[b] Faculty of Information Technology, Monash University, Australia
[c] Graduate School of Education, Stanford University, United States
[d] School of Informatics, University of Edinburgh, United Kingdom
[e] Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia

ARTICLE INFO

Article history:
Received 7 February 2021
Received in revised form 20 August 2021
Accepted 3 September 2021
Available online 9 September 2021

Keywords:
Intelligent Tutoring Systems
Educational dialogue analysis
Tutoring strategies
Dialogue acts
Student performance
Learning analytics

ABSTRACT

To construct dialogue-based Intelligent Tutoring Systems (ITS) with sufficient pedagogical expertise, a trendy research method is to mine large-scale data collected by existing dialogue-based ITS or generated between human tutors and students to discover effective tutoring strategies. However, most of the existing research has mainly focused on the analysis of successful tutorial dialogue. We argue that, to better inform the design of dialogue-based ITS, it is also important to analyse unsuccessful tutorial dialogues and gain a better understanding of the reasons behind those failures. Therefore, our study aimed to identify effective tutoring strategies by mining a large-scale dataset of both successful and unsuccessful human–human online tutorial dialogues, and further used these tutoring strategies for predicting students' problem-solving performance. Specifically, the study adopted a widely-used educational dialogue act scheme to describe the action behind utterances made by a tutor/student in the broader context of a tutorial dialogue (e.g., asking/answering a question, providing hints). Frequent dialogue acts were identified and analysed by taking into account the prior progress that a student had made before the start of a tutorial session and the problem-solving performance the student achieved after the end of the session. Besides, we performed a sequence analysis on the inferred actions to identify prominent patterns that were closely related to students' problem-solving performance. These prominent patterns could shed light on the frequent strategies used by tutors. Lastly, we measured the power of these tutorial actions in predicting students' problem-solving performance by applying a well-established machine learning method, Gradient Tree Boosting (GTB). Through extensive analysis and evaluations, we identified a set of different action patterns that were pertinent to tutors and students across dialogues of different traits, e.g., students without prior progress in solving problems, compared to those with prior progress, were likely to receive more thought-provoking questions from their tutors. More importantly, we demonstrated that the actions taken by students and tutors during a tutorial process could not adequately predict student performance and should be considered together with other relevant factors (e.g., the informativeness of the utterances).

1. Introduction

Dialogue-based Intelligent Tutoring Systems (ITS), similar to conventional ITS like SQL-Tutor [1], Algebra Tutor PAT [2], and eTeacher [3], aim at helping students construct knowledge and skills of different subjects by providing them with immediate and personalized instructions or feedback. Compared to conventional ITS, dialogue-based ITS deliver instructions or feedback by having natural and meaningful conversations with students [4], and are expected to act as competent as human tutors to engage students and provoke more in-depth thinking and learning. Given the promising potentials, both academic researchers and industrial practitioners have put great efforts in building various dialogue-based ITS, among which CIRCSIM-Tutor [5], AutoTutor [6], BEETLE II [7], and Why2 [8] are notable representatives. Noticeably, these systems have been deployed for use in practice and have assisted millions of students with their learning.

Despite being popular, most of the existing dialogue-based ITS are plagued by their inability in delivering personalized learning experiences to students [9]. The current dialogue-based ITS, as yet, fail to achieve their full potential and are unable to act as

* Corresponding author.
E-mail address: guanliang.chen@monash.edu (G. Chen).

https://doi.org/10.1016/j.future.2021.09.001
0167-739X/

competently as human tutors [10]. One main reason is that these dialogue-based ITS, more often than not, lack sufficient pedagogical expertise as human tutors in guiding students [11,12]. That is, these dialogue-based ITS typically have little knowledge about the tutoring strategies that can be of use to facilitate the tutoring process [13]. For instance, questioning a student's progress of learning problems is a common strategy used to help tutors detect knowledge gaps of the student at the beginning of a tutorial session [14]. Then, in follow-up teaching activities, tutors can better direct their efforts, e.g., introducing relevant learning contents and designing appropriate teaching activities to enable students to develop mastery of those concepts. It should also be noted that successful applications of such a tutoring strategy often depends on (i) a tutor's experience (and domain/contextual knowledge) in applying the strategy (e.g., when to ask questions and what type of question should be asked) and (ii) information about students (e.g., mastery level and learning progress) [15–18].

Numerous studies have been conducted to investigate how dialogue-based ITS can be equipped with relevant pedagogical expertise to apply appropriate tutoring strategies [4,11,17,19–22]. In this strand of research, a recent trend is to mine large-scale data collected by existing dialogue-based ITS or generated between human tutors and students to discover effective tutoring strategies [11,12,19,23]. However, existing data-intensive studies typically focused on the analysis of successful tutorial sessions (i.e., those in which students successfully solved problems or achieved meaningful learning) and the identification of effective tutoring strategies that tutors should take. We argue that, to provide students with necessary help, tutors should also learn from unsuccessful tutorial sessions and gain a better understanding of the factors contributing to such failures. Therefore, unsuccessful tutorial sessions should also be analysed to better guide the design and development of future dialogue-based ITS.

This study aimed to identify the frequent tutoring strategies used by tutors in not only successful but also unsuccessful tutorial sessions by mining a large-scale human–human tutorial dialogue dataset. The study also aimed to examine the extent to which these identified tutoring strategies are predictive of students' problem-solving performance. Here, we described a tutoring strategy as the actions taken by a tutor in the tutorial process (e.g., asking thought-provoking questions and providing hints). Formally, our work was guided by three **R**esearch **Q**uestions:

- **RQ1** What actions are commonly taken by tutors and students during tutorial sessions?
- **RQ2** What patterns of actions, i.e., one or multiple consecutive actions, are associated with different levels of students' performance in solving problems in tutorial sessions?
- **RQ3** To what extent are the identified actions and action patterns predictive the problem-solving performance of students in tutorial sessions?

To answer the above questions, we first employed a widely-used dialogue act (DA) scheme (proposed by [24]) to characterize tutors' (as well as students') actions behind their utterances in a tutorial dialogue. Then, the derived actions were analysed by applying a sequence analysis to shed light on the frequent tutoring strategies employed by tutors, which were further used as input for a well-established machine learning method—Gradient Tree Boosting (GTB) [25]—to measure the contribution made by these strategies in predicting students' problem-solving performance. To our knowledge, our study is the first to take students' prior progress into account to reveal effective tutorial strategies in human–human online tutoring. By analysing a large corpus consisting of both successful and unsuccessful tutorial dialogues, our study contributed with an in-depth understanding of tutors' as well as students' behaviour in human–human online tutoring and offered empirical evidence to support existing good practices (e.g., providing timely feedback to students) for the development of dialogue-based ITS.

2. Related work

Tutoring strategies refer to principles and approaches employed by instructors to better assist students to learn in various educational settings [15,26], e.g., raising a question to trigger in-depth thinking and acknowledging students' achievement to motivate them to continue to learn. Effective tutoring strategies play essential roles in helping instructors better direct their teaching efforts and enabling students to construct meaningful knowledge, and thus have been investigated for years [26].

Given the increasingly important role of dialogue-based ITS, there have been growing debates and research endeavours on investigating to what extend and how dialogue-based ITS should be constructed to make use of effective tutorial strategies to students. As summarized in [15], there are three typical methods employed to develop dialogue-based ITS with tutoring strategies:

- *Observing from human expert instructors*. Researchers crafted a set of tutoring rules by observing from human tutors' effective tutoring practices, which were further incorporated into the development of ITS. Take *AutoTutor* [27] as examples, which were developed by applying rule-based algorithms to simulate human tutors to employ tutoring strategies such as asking questions and providing short feedback to students. The strategies applied in AutoTutor were demonstrated effective in assisting undergraduate students to develop the mastery in introductory computer literacy courses [27].
- *Deriving from learning theories*. The design and development of an ITS were guided by well-established learning theories. For instance, by building upon the adaptive control of thought theory [28], which describes the characteristics of human cognition in the process of memory, [29] developed tutoring strategies such as decomposing the problem into a set of sub-problems and providing timely feedback on errors. These strategies were adopted by *cognitive tutors* [29] and demonstrated benefits to promote high-school students' success rate in the studies of algebra, geometry, and computer programming.
- *Observing students*. Researchers developed tutoring strategies by observing how students of different characteristics (e.g., gender, age, and mastery level in the learning subject) responded to different teaching practices. For example, Arroyo et al. [30] analysed how elementary-school students of different gender and different levels of cognitive development reacted to various types of hints (e.g., hints provided in a form of numeric symbolic or concrete visual shape) when students attempted to solve mathematical problems.

It is worth noting that all of the methods described rely heavily on the collection and analysis of the data generated between human tutors and students or between ITS and students, especially the methods that involve *observing from human expert instructors* and *observing students* [11,12,23,24,31]. As the first step to reveal effective tutoring strategies, existing studies often developed a coding scheme to characterize the actions hidden behind the utterances made by tutors and students in a tutorial dialogue (i.e., DAs) [32–34]. One of the pioneering studies in this strand of research [35] proposed to use a five-step framework to describe and characterize the process of one-on-one tutoring including the following steps: (i) a tutor starts by asking a

question; (ii) a student attempts to answer the question; (iii) the tutor provides feedback on the quality of the answer; (iv) the tutor and the student collaborate to improve the quality of the answer; and (v) the tutor assesses the student's understanding of the answer. By analysing two samples of tutorial dialogues, this framework was demonstrated as effective in identifying frequent dialogue patterns that characterized the collaborative nature of the one-on-one tutorial process. The effectiveness of this five-step framework was also validated in several studies [6,36–38]. As another example, Hennessy et al. [34] developed a coding scheme to take interlocutors' sociocultural backgrounds into account (e.g., the relationship between the interlocutors), which consists of 33 dialogue act codes including `express or invite ideas` and `nake reasoning explicit`. Other representative studies on developing DA schemes include [24,39–41]. In particular, the scheme presented in [24] was developed by building upon several prior schemes, which consisted of two levels of DA tags, with the first level describing the general flow of a tutorial process (e.g., a tutor raised a question) and the second level capturing more fine-grained information of a specific action taken by a tutor/student (e.g., what type of question was asked by the tutor). This DA scheme has been widely adopted in recent studies [42–46] and demonstrated effective in revealing varying dialogue patterns.

With the DA tags determined for the utterances, effective tutoring strategies were investigated by analysing the relationship between students' learning performance and dialogue actions taken by tutors or students [11,24,47–52]. For example, Boyer et al. [47] demonstrated that the actions performed by tutors played a significant role in delivering different kinds of learning outcomes, e.g., actions offering encouragement to students were useful in bolstering self-efficacy, while actions providing positive cognitive feedback were helpful in boosting learning gains. In a similar view, Vail and Boyer [24] applied correlation analysis to reveal actions and action bigrams (i.e., two consecutive actions) that were indicative of student learning, e.g., the action bigram consisting of positive feedback given by tutors after a confirmation question from the student was positively correlated with the learning gain of students. In particular, there are three studies most relevant to our work [11,12,19]. All of these three studies made use of the DA scheme developed by Morrison et al. [39] to capture the interaction between human tutors and students recorded in a large-scale dataset consisting of over 19K tutorial dialogues. To locate the effective tutoring strategies adopt by tutors, Maharjan et al. [11,12] applied sequence analysis to characterize the significant action patterns displayed in successful tutorial dialogues. These patterns indicated that tutors tended to be more expressive and encouraging at the beginning of successful tutorial dialogues and used more scaffolding strategies (e.g., providing a series of hints to students) during the tutorial process.

Compared to the works described above, our work distinguished itself from several perspectives. Firstly, the studies presented in [11,12,19] relied on a large amount efforts of experienced coders to label the dialogues, which was rather costly and time-consuming. Our research, instead, applied the state-of-the-art pre-trained language model BERT with a small sample of labelled data as input to automatically infer the actions taken by tutors and students and tutors in all tutorial dialogues. Secondly, most of the existing studies have mainly focused on analysing successful tutorial dialogues, i.e., those in which students achieved meaningful learning. In contrast, we distinguished dialogues in a more fine-grained manner by considering both students' prior progress in solving a task before the start of a tutorial session as well as the students' ultimate problem-solving performance after the end of the session, and revealed action patterns that were specific to different categories of dialogues. Thirdly, to our knowledge, our work is the first study to apply a state-of-the-art machine learning method–GTB–to systematically quantify the power of tutorial actions and action patterns in predicting students' problem-solving performance in human–human online tutoring.

3. Methods

3.1. Dataset

With the ethics approval from Monash University for secondary data use (Project ID 26156), we used a deidentified tutorial dataset that was prepared by an educational technology company. The educational technology company provides a mobile phone application for tutors and students to work together to solve problems covering subjects like mathematics, chemistry, and physics. With the mobile application, a student could take a picture of an unsolved problem and initialize a request for help. Then, the application connected the student with an experienced tutor who guided the student to solve the problem by leveraging texts and images to communicate. According to the policy of the educational technology company, tutors should give their best to guide students to solve problems by themselves and are disallowed to directly share answers with the students. That means, the dialogues contained in this dataset detailed processes of how tutors and students collaborated to solve various problems.

Recall that this study aimed to reveal frequent tutoring strategies occurring in not only successful but also unsuccessful tutorial dialogues. Notice that students with different learning progress often requires different support to complete a learning task. A tutor is often suggested to select appropriate tutoring strategies by taking into account the prior progress the student has achieved before entering a tutorial session [53,54]. Therefore, we manually labelled the dataset in two steps. That is, for each dialogue, we first determined the level of prior progress that a student made in solving a problem before talking to a tutor, and then further characterized the level of problem-solving performance the student achieved after the end of a tutorial session, as described below.

Step 1: Prior Progress. At the beginning of a tutorial session, a tutor often raised questions to a student to learn about whether the student had made certain progress in solving a problem. For instance, a tutor might ask *"Can you tell me what you have tried so far?"*, and a student might answer *"I haven't tried anything"* or *"Yes"* and then described the progress she had made. By manually scrutinizing the first few utterances of a tutorial dialogue, we were able to determine a student's level of prior progress and, correspondingly, labelled the dialogue as either *With-Prior-Progress* or *Without-Prior-Progress*. In total, we recruited three human coders to label the whole dataset. Each dialogue was labelled by two coders and the disagreements between the two coders were resolved by the third coder. The overall agreement percentage score was 0.847 and the Cohen's κ score was 0.735, which demonstrated a substantial level of agreement.

Step 2: Problem-solving Performance. Different from previous studies [11,12,19,24], we distinguished the performance level of a student in a more fine-grained manner and labelled each dialogue in our dataset to one of the following categories:

- **Gap-clarified**: a tutor was able to identify the problem but uncertain whether the student had made any progress;
- **Gap-explained**: a tutor was able to identify the problem and help the student make certain progress, but the student had not identified a correct or full solution;

Table 1
The descriptive statistics of the dataset used in the study. With PP and Without PP denote *With Prior Progress* and *Without Prior Progress*, respectively. Mann–Whitney tests were applied to examine the difference (Rows 5–9) between any two of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress. All differences were significant ($p < 0.01$).

Metric	All	Gap-clarified		Gap-explained		Gap-bridged	
		With PP	Without PP	With PP	Without PP	With PP	Without PP
1. # total sessions:	14,562	1,203	1,302	1,255	1,931	4,482	4,389
2. # total uttrances:	1,216,784	31,014	30,128	78,575	113,099	475,849	488,119
3. # tutors:	116	92	96	98	99	110	106
4. # students:	5,165	763	962	908	1,419	1,800	2,168
5. Avg. Sess Dur (mins):	30.27 ± 30.66	10.55 ± 7.64	9.75 ± 7.21	25.94 ± 19.03	22.88 ± 18.05	37.78 ± 32.17	38.60 ± 37.37
6. Avg. # Uttr/Sess:	83.56 ± 81.05	25.78 ± 16.68	23.14 ± 14.92	62.61 ± 43.79	58.57 ± 42.73	106.17 ± 87.62	111.21 ± 93.70
7. Avg. # Words/Sess:	647.75 ± 596.12	201.62 ± 131.81	198.13 ± 134.44	524.09 ± 351.18	489.56 ± 346.82	807.46 ± 649.08	845.28 ± 675.05
8. Avg. % Uttr by tutors:	58.42 ± 7.86	53.95 ± 9.49	56.46 ± 9.51	58.68 ± 7.82	60.25 ± 7.77	58.03 ± 7.07	59.75 ± 6.94
9. Avg. % Words by tutors:	78.36 ± 9.10	74.32 ± 11.80	80.54 ± 10.20	78.87 ± 8.59	82.21 ± 7.96	76.09 ± 8.69	79.30 ± 7.81

- **Gap-bridged**: a tutor was able to identify the problem and guide the student to successfully solve the problem or a similar problem.

In this step, each tutorial dialogue was labelled by an independent educational expert employed by the educational technology company that collected the dataset. To validate the reliability of these expert-crafted labels, we randomly selected 500 tutorial dialogues and labelled them independently by using the same coding rules. Our labels reached a percentage agreement score of 0.884 and Cohen's κ score of 0.787 with those expert-crafted labels. We provided sample dialogues for each of these categories in an electronic appendix, which is accessible via https://github.com/bertDA/BertDA/blob/main/DA_Appendix.pdf

It is worth noting that the three dialogue categories were on an ordinal scale relative to the level of students' problem-solving performance, i.e., there was an increasing amount of problem-solving progress obtained by students from the Gap-clarified dialogues to the Gap-bridged dialogues. The dataset originally consisted of 18,203 dialogues. Since tutors were unlikely to deliver meaningful tutoring in short dialogues, we removed dialogues that (i) contained less than 10 utterances; (ii) lasted less than 1 min; (iii) were difficult to be determined whether a student had made any progress before the start of a tutorial session or during the tutorial session, e.g., those in which a student quit a session because no tutor was assigned to help the student or a student did not reply to a tutor at all in the whole tutoring process. After removal, there were 14,562 dialogues left, among which about 92% were related to math tutoring. The descriptive statistics of the dataset are given in Table 1. Most of the dialogues were of the category Gap-bridged (8,871, 60.9%), followed by Gap-explained (3,186, 21.9%) and then Gap-clarified (2,505, 17.2%). This suggests, in our case, more than a half of the students successfully solved problems. Besides, over 47% of the students had made certain progress before joining a tutorial session, and compared to their counterparts without any prior progress, these students were more likely to identify correct solutions (i.e., Gap-bridged dialogues), which was in line with our expectations. Also, it would be intuitive to assume that, the better problem-solving performance achieved by a student, the more efforts the student as well as the tutor had invested in a tutorial session. To corroborate this assumption, we depicted the characteristics of the dialogues in rows 5–7 of Table 1. As we can observe, there was a steady increase from Gap-clarified and Gap-explained to Gap-bridged in terms of the session duration and the number of utterances and words contained in a dialogue.

3.2. Dialogue act scheme and dialogue act labelling

In line with previous studies [24,39], we also characterized the underlying actions taken by tutors and students in a tutorial session by using the two-level DA scheme presented in [24], whose effectiveness in depicting online one-on-one tutorial process has been validated in several studies [43,55,56]. The structure of the DA scheme is detailed in Table 2. Specifically, there are 12 first-level DA tags in the scheme, which can be used to portray the general tutor–student interaction, e.g., tutors raised thought-provoking questions to students (i.e., the tag `Question`) and students answered the questions raised by the tutors (i.e., the tag `Answer`). To capture more fine-grained information from tutor–student interaction, the 12 first-level labels are further expanded to 31 second-level DA tags. For instance, the tag `Question` is extended to distinguish between types of questions raised by tutors, including the questions requiring students to recall specific learning concepts (`Factual Question`), the questions prompting students' critical thinking (`Probing Question`), and the questions encouraging students to reason and reflect (`Open Question`). Noticeably, some DA tags are pertinent to only tutors or students while the others are pertinent to both. For example, the tag `Request Feedback` can only be used to describe the utterances generated by students to seek feedback, while the tags `Positive Feedback` and `Negative Feedback` can only be used to describe the utterances generated by tutors to provide feedback to students. As for tags like `Acknowledge` (expressing agreement with or acknowledgement of their interlocutors) and `Correction` (correcting the typo errors made in their previous utterance) can be used to describe utterances made by both tutors and students. Recall that our dataset was obtained from a mobile phone application used for online one-on-one tutoring, which allows tutors and students to use not only texts but also images to communicate. Based on our observations on the dataset, tutors typically used images to provide hints to inspire students and students often used images to seek feedback from tutors for the partial or full solution developed by them. None of the existing tags in the DA scheme can be used to depict these actions. Therefore, to better capture the tutor–student interaction observed in our dataset, we added two new second-level tags to the DA scheme, i.e., `Hint by Image` within the first-level tag `Hint` and `Request Feedback by Image` within the first-level tag `Request Feedback`.

Considering the number of tags contained in the adopted DA scheme and the number of tutorial dialogues contained in our dataset, it would be a very time-consuming and costly process if we purely relied on human coders to identify the DA tags for the whole dataset. Inspired by Rus et al. [19,57], we labelled a subset of the original dataset, which was used to train a classifier by applying machine learning techniques and then further used the classifier to automatically infer the DA tags for the remaining data. Specifically, we recruited two educational experts who have been involved in teaching for years as coders to label 50 randomly-selected tutorial dialogues in our dataset, which contained a total of 3,629 utterances. It should be noted that an utterance often contained multiple sentences and different

Table 2
The description of the DA scheme developed in [24]. The DA tags marked with ♣ are added by us to better depict the tutorial process in our dataset. The column Role indicates whether a DA is only specific to tutors (T), students (S), or specific to both. (* `Operational Question` tag was Originally denoted as `Question` in the DA scheme in [24]. We described it as `Operational Question` to better illustrate the difference between this tag and other tags.)

First-level DA Tag	Second-level DA Tag	Role	Examples in our dataset
Hint	Information	T	*"It can be any one of the cards in the deck."*
	Hint by Image ♣		*[Image]*
	Observation	T&S	*"We have 80."*
Directive	Directive	T	*"Check this definition."*
Acknowledge	Acknowledge	T&S	*"Alright!"*
Request Confirmation	Evaluation Question	T	*"Does that make sense?"*
Request Feedback	Request Feedback by Image ♣	S	*[Image]*
	Confirmation Question		*"Would the answer be 30?"*
Positive Feedback	General Positive Feedback	T	*"Correct!"*
	Elaborated Positive Feedback		*"Your formula for period is correct!"*
Negative Feedback	Negative Feedback	T	*"No, it is incorrect."*
Lukewarm Feedback	Lukewarm Feedback	T	*"Almost correct, but the sign is missing."*
Correction	Correction	T&S	*"*We will"*
Question	Direction Question	S	*"How do I do that?"*
	Information Question		*"What are the units for W?"*
	Probing Question		*"How many options can it be?"*
	Open Question		*"What do you think we could try next?"*
	Factual Question	T	*"What is the value of x?"*
	Operational Question *		*"Any questions on this?"*
	Ready Question		*"Are you ready to begin?"*
	Extra Domain Question	T&S	*"How are you doing today?"*
Answer	Yes-No Answer	T&S	*"Yes, that would be very helpful."*
	WH Answer	T&S	*"It is 6."*
	Ready Answer	S	*"Yes, I'm ready."*
	Extra Domain Answer	T&S	*"I'm good."*
Statement	Explanation	T&S	*"The straight line is the line on the bottom."*
	Greeting	T&S	*"Hello!"*
	Extra Domain Other	T&S	*"Welcome to use this app!"*
	Reassurance	T	*"No problem, I will help you."*
	Understanding	S	*"Ok, got it."*
	Not Understanding		*"I don't know why."*

sentences could indicate different actions (i.e., sentences could be assigned with different DA tags), the labelling was performed on a sentence level. Also, to enable enough fine-grained information to be captured, we asked the coders to identify not only the first-level but also the second-level tags for each sentence. Before starting the labelling, the two coders were required to develop a clear understanding of each tag contained in the DA scheme and correspondingly crafted a set of labelling rules (e.g., sentences containing keywords like "*hello*" and "*welcome*" should be assigned with the tag `Greeting`). Then, the two coders started to annotate five tutorial dialogues together, through which the labelling rules were revised and expanded to facilitate the subsequent labelling. Then, each of the remaining 45 tutorial dialogues was labelled by the two coders independently and their overall agreement score was 0.77 (measured by Cohen's κ), which indicates a substantial agreement between the two coders and the derived labels were reliable. The cases with disagreements were resolved by inviting a third educational expert to discuss together with the original two coders. As we aimed to reveal the frequent actions and action patterns used in tutorial dialogues, only the sentences labelled with first-level tags which occurred in more than 5% of total sentences were further labelled with second-level DA tags.

3.3. Inferring educational dialogue acts

With the labelled data derived in Section 3.2, we aimed to construct a classifier to automatically infer the DA tags for the remaining dialogues. Driven by the great success achieved by pre-trained language models in deriving accurate representations of textual data [58], which can be further utilized to facilitate downstream prediction tasks (e.g., DA identification in our case), we also used BERT [59,60] in our study. Notably, BERT has been demonstrated as effective in various settings, even with a limited amount of labelled data [59]. Here, we concatenated a single classification layer as the task model on top of BERT's output for the [CLS] and [SEP], which are the special tokens used in BERT embedding to encode the information of the sentence segmentation from the whole input data. As indicated in [61], the assignment of a DA tag to a sentence often depends on the context in which the sentence was uttered. Therefore, in order to make use of the context related to a sentence for DA prediction, we concatenated the following information for each labelled sentence as input to train the classifier:

- The text of a sentence;
- The person who uttered the sentence (i.e., tutor or student), which enabled BERT to relate the linguistic difference of tutor/student-generated utterances to different tutor/student-specific DA tags;
- The order of a sentence in a tutorial dialogue, which enabled BERT to account for the occurrence likelihood of different DA tags throughout a tutorial process;
- The session ID, which enabled BERT to capture the overall context in which a sentence was uttered; and
- The text of the sentence preceding the current sentence, which enabled BERT to pay specific attention to the local context surrounding a sentence.

With the classifier built, the DA tags of the remaining dialogues were automatically inferred. Then, the distribution of

these DA tags in the whole dataset as well as in each category of dialogues were further analysed to answer RQ1.

3.4. Mining frequent action patterns

To answer RQ2, we employed the `TraMineR` package in R to identify the discriminant action patterns in our dataset. `TraMineR` is a popular tool used to mine, describe, and visualize discriminant sequences or discrete sequences of states or events in data. Though primarily developed to analyse biographical longitudinal data, `TraMineR` has been successfully applied to other kinds of categorical sequence data, including sequences of actions in tutorial dialogues [11,12]. Specifically, we used `TraMineR` as follows: (i) we first extracted the frequent action patterns by counting their occurrence frequency in all dialogues with the aid of the `seqefsub()` function of `TraMineR`; then (ii) these frequent action patterns were used as input to the `seqecmpgroup()` function of `TraMineR`, which applied the Pearson Independence Chi-squared test with Bonferroni correction to retrieve action patterns that can be used to discriminate dialogues with different levels of student performance. To depict how discriminative an action pattern is, we further computed the value of *Pearson Residual*, which is a statistic used to compare the dispersion of the observed action pattern with the expected occurrence and indicate the degree of its departure to the expected occurrence. A positive Pearson Residual indicates that the actual occurrence of an action pattern is higher than its expected occurrence, while a negative Pearson Residual indicates a lower actual occurrence than the expected occurrence. Here, we selected a p-value threshold of 0.01 so as to reveal the most likely distinctive action patterns for different categories of dialogues. It is worth noting that an action pattern is not necessarily a contiguous sequence of DAs observed in the data. Instead, the order of the observed DAs is preserved. For instance, *(Tutor, Information)-(Tutor, General Positive Feedback)* may be formed from the contiguous sequence of *(Tutor, Info)-(Student, Confirmation Question)-(Tutor, General Positive Feedback)*.

3.5. Predicting student performance

Prediction Model. For RQ3, we aimed to evaluate the effectiveness of the observed actions or action patterns in predicting the problem-solving performance of students, i.e., predicting which label (among Gap-clarified, Gap-explained, and Gap-bridged) should be assigned to a tutorial dialogue. Essentially, this can be treated as a multi-class classification problem. Examples of typical techniques used for multi-class classification problems are Naive Bayes, decision trees, and support vector machines, while recent studies suggested that techniques like Gradient Tree Boosting [62] can also be of use. Our recent study [63] utilized this technique for predicting students' satisfaction with a tutoring service by leveraging a set of different features derived from the dialogue discourse. GTB is designed based on the rationale of ensemble learning [64], which states that multiple predictors aiming to predict the same target variable are more likely to deliver better performance than any single predictor alone. In fact, GTB is highly similar to random forests, both of which construct multiple decision trees as the predictors, and the final prediction is generated by combining the predictions of all constructed predictors with techniques like weighted average and majority vote. It is worth noting that each decision tree is constructed with a random sub-sample of the data. By doing this, each decision tree is slightly different from the others and more importantly, these decision trees together can adequately capture the characteristics of the data and thus deliver better prediction performance. The main difference between GTB and random forests lies in that, the predictors in random forests are constructed independently, which means, there can be multiple predictors producing the same type of prediction error. On the contrary, the GTB predictors are built in a sequential manner, in which the errors produced by previous predictors can be corrected by the latter predictors, and thus GTB takes less time to reach close to actual labels. In particular, previous research has demonstrated the effectiveness of GTB in dealing with various types of feature data for a wide range of machine learning problems. Therefore, in line with [63], whose prediction task is highly similar to ours (i.e., classifying educational dialogues), we also adopted GTB as the predictive modelling method.

Feature Engineering. To measure the extent to which the observed actions and action patterns in a tutorial dialogue can indicate the problem-solving achievement accomplished by students, we engineered the following three groups of features:

- **# Individual DA**: the number of a specific DA made by a tutor/student in a dialogue;
- **% Individual DA**: the fraction of a specific DA made by a tutor/student in a dialogue (divided by the total number of the identified DA);
- **# Significant action patterns**: the number of significant action patterns appeared in a dialogue (as discovered by applying the method described in Section 3.4).

In total, we designed 218 features based on the DA produced by both tutors and students, including (i) 18 first-level DA (here 6 out of the 12 original first-level DA are shared between tutors and students); (ii) 41 second-level DA (here 10 out of the 31 original second-level DA are shared between tutors and students); (iii) the percentage of 59 first-level and second-level DAs; and (iv) 100 discriminative action patterns found by `TraMineR`. We denoted these features as *DA features*. We acknowledged that a student's prior progress in solving a problem might be beneficial in boosting the prediction performance. However, the acquisition of such information relied on the manual analysis of the first few utterances in a tutorial dialogue in our current study. As we aimed to develop a prediction model that can be deployed for real-time use in practice, i.e. the input features to the model should be directly and automatically engineered from the observed data, we did not incorporate this into the feature set. In the future, we plan to develop methods to automatically determine the prior progress of a student as a tutorial session proceeds, and further take this information into account for predicting problem-solving performance.

Though we mainly focused on DAs produced by tutors and students in this study, as student performance may not be solely determined by the DAs, it would be necessary to include other relevant features to quantify the role of DA features in predicting student performance. For instance, the informativeness and complexity of utterances expressed by tutors may be greatly related to student performance [63]. Therefore, in addition to the *DA features* described above, we further used [63] as a reference and engineered another 325 features from the dataset and used them for predicting the labels of dialogues. These features include:

- **Efforts**, i.e., the efforts that a tutor/student invested in a tutorial session, which were measured by calculating the duration of the tutorial session, the number of utterances uttered by the tutor/student, and the number of words contained in those utterances;
- **Informativeness**, i.e., the informativeness of the utterances uttered by a tutor/student, which was measured by calculating the number (or fraction) of unique words and concepts contained in those utterances;

Table 3
Top 10 most frequent DAs identified in our dataset (sorted according to the fraction of utterances associated with a specific DA in the whole dataset in a descending order, i.e., the column **All**). The top 3 largest fraction numbers in each column are in bold. **T** denotes tutors and **S** denotes students. **With PP** and Without PP denote *With Prior Progress* and *Without Prior Progress*, respectively. Mann–Whitney tests were applied to examine the difference between any two of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress. Except for the results marked with the same symbol in a row (e.g., ◊, †), the others were all significant ($p < 0.01$).

	Dialogue Act	Role	All	Gap-clarified		Gap-explained		Gap-bridged	
				With PP	Without PP	With PP	Without PP	With PP	Without PP
1.	General Positive Feedback	T	**10.16%**	**9.18%**	6.71%	**8.16%**	**7.73%**	**11.41%**	**11.16%**
2.	Information	T	**8.62%**	6.57%	8.03%	**9.05%**	**10.35%**	**7.77%**	**8.96%**
3.	Probing Question	T	**8.10%**	6.81%	7.23%	◊ 8.33%	**8.46%**	◊ 7.87%	**8.50%**
4.	Yes-No Answer	S	7.19%	8.22%	**9.49%**	6.66%	◊ 7.59%	6.49%	◊ 7.02%
5.	WH Answer	S	6.41%	6.64%	† 6.97%	◊ 6.78%	† 6.39%	◊ 6.15%	6.44%
6.	Acknowledge	S	5.67%	7.37%	7.12%	5.35%	◊ 5.93%	5.32%	◊ 5.34%
7.	Request Feedback by Image	S	5.10%	**8.60%**	6.34%	◊ 5.13%	3.84%	◊ 5.45%	3.95%
8.	Extra Domain Other	T	4.93%	**9.43%**	**11.13%**	5.34%	5.70%	3.46%	3.30%
9.	Confirmation Question	S	4.89%	◊ 5.39%	5.19%	4.65%	4.59%	◊ 4.92%	4.93%
10.	Operational Question	T	4.73%	7.93%	**8.53%**	4.12%	◊ 4.45%	3.89%	◊ 3.98%

- **Complexity**, i.e., the complexity of the utterances uttered by a tutor/student, which was measured by applying Flesch readability score [65];
- **Responsiveness**, i.e., the average amount of time that a student needed to wait before receiving a reply from a tutor after the student sent an utterance;
- **# Questions**, i.e., the number of questions asked by a tutor/student in a tutorial session;
- **Entrainment**, which calculates a score to describe the degree to which tutors' utterances and students' utterances were aligned with each other in a tutorial session;
- **Sentiment**, i.e., the overall sentiment polarity scores of utterances sent by a tutor/student in a tutorial dialogue;
- **Experience**, i.e., the number of tutorial sessions that a tutor/student had prior to the current one;
- **N-grams**. The top 100 most frequent unigrams, bigrams, and trigrams contained in the utterances of a dialogue.

3.6. Study setup

Model Training for DA Classification. We implemented the classifier by using a BERT pre-trained language model [60]. For classifying DA, the number of neurons contained in the classification layer coupled with BERT was set to 768 and *softmax* was selected as the activation function. The labelled sentences were randomly split to *training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%. We set the maximum sequence length to 512 and fine-tuned on a batch size of 32 for 6 epoches. AdamW optimizer was used with the learning rate of 2e-5 to optimize the training of the classifier. All experiments were implemented on Titan GTX 2080ti and 2.50 GHz Intel Xeon E5-2678 v3 CPU processor.

Model Training for Student Performance Prediction. For predicting student performance, we randomly assigned the dialogues to the *training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%. For comparison, we selected random forests as the baseline method to demonstrate the effectiveness of GTB. Both random forests and GTB were implemented with the aid of the `scikit-learn`[1] library in Python and their parameters were fine-tuned by applying grid search on the validation data, and then we evaluated the performance of the two methods on the testing data.

Evaluation Metrics. For both of the two classification tasks above, we adopted three representative metrics for measuring the competency of the classification models, i.e., Area Under the Curve (AUC), F1 score, and Cohen's κ coefficient (Cohen's κ). We also present the result of classification accuracy as a reference.

[1] https://scikit-learn.org/.

4. Results

With the method described in 3.3, we built a DA classifier which successfully assigned correct labels for 75% of the sentences in the labelled dataset. More specifically, the classifier achieved a performance of 0.742 and 0.828 in terms of F1-score and AUC, respectively. In particular, the classifier achieved a Cohen's κ of 0.735, which demonstrated a sufficient performance level, especially given the large number of DA contained in our dataset (i.e., 31 second-level tags). This pre-trained DA classifier is available at https://github.com/bertDA/BertDA.

In the following, we detail the results obtained in response to the three RQs raised in Section 1.

4.1. Results on RQ1

The top 10 most frequent second-level DAs are shown in Table 3. These DAs, in total, accounted for 65.80% of the sentences in the dataset. We can observe that `General Positive Feedback`, `Information`, and `Probing Question` were ranked 1st, 2nd, and 3rd in the table, respectively. These actions were often taken by tutors to give necessary hints (`Information`), to raise thought-provoking questions (`Probing Question`), or to offer positive feedback (`General Positive Feedback`) to acknowledge students' achievement. The high occurrence frequency of such tutor-specific DAs suggests that, in online one-on-one tutoring, tutors tended to take the lead role in this collaborative problem-solving process. On the other hand, the most frequent actions by students were `Yes-No Answers` (i.e., a tag used to annotate students' responses, which typically start with a "*yes*" or "*no*", to simple questions), `WH Answers` (i.e., a tag used to annotate students' responses to complex questions with starting words including "*what*", "*why*", and "*how*") and `Acknowledge` (i.e., a tag used to annotate students' statements made to express acknowledgement or agreement to the explanations provided by tutors), which ranked 4th, 5th, and 6th in the table, respectively. Again, this is not a surprising result given the large number of the probing questions raised by tutors.

By differentiating the levels of students' prior progress, we can observe several findings in Table 3. Firstly, the fraction of the DA tag `General Positive Feedback` given to *With-Prior-Progress* students was generally higher than that to *Without-Prior-Progress* students in all three session categories. This is not a surprising result, as indicated in Table 1, a student with prior progress was more likely to successfully solve a problem, and thus received more positive feedback from tutors. Also, in the dialogues where students successfully solved problems (i.e., Gap-bridged), tutors had the highest usage of `General Positive`

Table 4
The top 10 most frequent action patterns from each category of dialogues. The patterns are sorted according to their occurrence frequency in an descending manner. The action patterns that occurred in only one performance category of dialogues (i.e., Gap-clarified, Gap-explained, and Gap-bridged) are in bold, and the action patterns that occurred in only one prior-progress category of dialogues (e.g., With or Without Prior Progress) are marked with ♠. T denotes tutors and S denotes students. Here are the abbreviation of the DA tags: Oprt-Ques (`Operational Question`), Y-N-Ansr (`Yes-No Answer`), Req-Fdbk-Img (`Request Feedback by Image`), G-Pos-Fdbk (`General Positive Feedback`), and Prob-Ques (`Probing Question`).

	Gap-clarified	Gap-explained	Gap-bridged
With Prior Progress	(S, Req-Fdbk-Img)-(T, Greeting) **(S, Req-Fdbk-Img)-(T, Oprt-Ques)** **(T, Greeting)-(T, Oprt-Ques)** **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Oprt-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ **(T, Greeting)-(T, Greeting)** ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Greeting)** ♠ (S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(T, Prob-Ques) **(T, Greeting)-(T, Prob-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Prob-Ques)** ♠ (S, Req-Fdbk-Img)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ (T, Greeting)-(T, G-Pos-Fdbk) **(S, Req-Fdbk-Img)-(T, Information)**	(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(T, Greeting) (T, Greeting)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, G-Pos-Fdbk)** **(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** **(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** (T, Greeting)-(S, Y-N-Ansr) ♠ **(T, Greeting)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** (S, Req-Fdbk-Img)-(T, Prob-Ques)
Without Prior Progress	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(T, Oprt-Ques) (T, Greeting)-(T, Oprt-Ques) **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Oprt-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Oprt-Ques)-(S, Y-N-Ansr)** ♠ **(T, Oprt-Ques)-(S, Y-N-Ansr)** ♠ **(T, Greeting)-(T, Oprt-Ques)-(S, Y-N-Ansr)**	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) **(S, Req-Fdbk-Img)-(T, Information)** (S, Req-Fdbk-Img)-(T, Prob-Ques) ♠ **(S, Greeting)-(T, Information)** ♠ (S, Req-Fdbk-Img)-(T, Oprt-Ques) **(T, Greeting)-(T, Prob-Ques)** (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ (S, Greeting)-(T, Oprt-Ques)	**(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting) **(T, Greeting)-(T, G-Pos-Fdbk)** (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Prob-Ques) ♠ (S, Req-Fdbk-Img)-(T, Oprt-Ques) **(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** **(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** ♠ **(S, Y-N-Ansr)-(T, G-Pos-Fdbk)**

`Feedback`. This may indicate that positive feedback provided by tutors to students may be treated as a strong discriminator in revealing the problem-solving performance of students. Secondly, *Without-Prior-Progress* students, compared to *With-Prior-Progress* ones, received more `Information` and `Probing Question` from tutors in all three session categories. This indicates the extra scaffolding provided by tutors to assist students without making much progress in solving problems before entering a tutorial session. An interesting observation is that, compared to Gap-bridged students, Gap-explained students tended to receive more `Information` hints. Thirdly, it is noted that *Without-Prior-Progress* students in the Gap-clarified categories had high usage of `Yes-No-Answer` (9.49%). This suggests that tutors might have allocated extra efforts to engage and guide these students by asking simple questions. Among these questions, some utterances can be tagged as `Operational Questions`, e.g., *"Do you have any progress on it?"*. In addition, tutors in Gap-clarified dialogues used `Extra Domain Other` more frequently than their counterparts in the other two session categories.

4.2. Results on RQ2

To answer RQ2, we first extracted the frequent action patterns from all of the dialogues and counted their occurrence frequency in each category of dialogues (shown in Table 4). Here, we only considered patterns that appeared at least in 1% sentences in our dataset. In total, we identified 100 frequent action patterns. As action patterns consisting of only one DA tag, e.g., `Probing Question` and `Yes-No Answer`, were rather common in all categories of dialogues, we only present patterns that consist of at least two DA tags in Table 4. It is worth noting that all categories of dialogues have action patterns that are only specific to themselves, respectively. For instance, in Gap-clarified dialogues, a student's request for feedback (`Request Feedback by Image`) was often followed by responses from tutors who did not directly address the problem to be solved, such as `Operational Question` ("*Are you following me?*"). This, again, signified the extra efforts invested by a tutor to build the common problem-solving ground with a student. While scrutinizing the frequent patterns of Gap-bridged dialogues, we can easily observe that the same action taken by a student (i.e., `Request Feedback by Image`) was often followed by tutors' `General Positive Feedback`. As for the actions followed behind `Request Feedback by Image` in the Gap-explained dialogues, we can observe a high occurrence of `Information` (e.g., *"You should add the value of x"*) and `Probing Question` (e.g., *"How many elements do you get?"*), but not `General Positive Feedback`. These results, together, suggest that a student's problem-solving performance can largely be revealed by the varying usage of DA tags like `Operational Question`, `Information`, `Probing Question`, and `General Positive Feedback`. When comparing *With-Prior-Progress* with *Without-Prior-Progress* dialogues, we can see that students with prior progress received more `General Positive Feedback` from tutors while students without prior progress often faced more `Operational Questions` from tutors. This further corroborates our findings presented in Table 3.

With the aid of `TraMineR`, we identified a total of 100 discriminant action patterns that could be used to distinguish among the three categories of dialogues. The top 10 most discriminant patterns from the group *With Prior Progress* and *Without Prior Progress* are presented in Table 5. Interestingly, all of the top 10 patterns in both groups contained at least one of the following three DA tags pertinent to tutors, i.e., `General Positive Feedback`, `Probing Question`, and `Information`. This corroborates the findings presented in Table 4. This suggests that, when applying machine learning to predict the likelihood of a student successfully solving a task in a real-time manner, the thought-provoking questions asked by tutors and the hints and feedback provided by tutors can potentially be regarded as strong discriminators to distinguish different categories of tutorial sessions. Interestingly, `General Positive Feedback` is of a higher occurrence in the action patterns of *Without-Prior-Progress* dialogues that those of *With-Prior-Progress* dialogues in our case.

4.3. Results on RQ3

For RQ3, we aimed to investigate the extent to which DAs can be used to reveal students' problem-solving performance. It should be noted that, in the real-world tutorial scenarios, the earlier an unsuccessful session can be identified, the more interventions a tutor can take to help a student. Therefore, we were particularly interested in investigating whether the observed actions and action patterns displayed a varying predictive power as a tutorial session progressed. To this end, we selected the first N utterance in a tutorial dialogue to extract the DA features described in Section 3.5 as input for both GTB and random forests, and the results are depicted in Fig. 1 (the dash lines). We used $N \epsilon$ [5, 10, 15, 25, 30, 35, 40] as well as all of the available utterances as input for student performance prediction (denoted as ALL in Fig. 1). Based on Fig. 1, we can conclude that GTB was generally more effective than random forests in predicting student performance in terms of all the evaluation metrics, though the performance of these two models was relatively limited. For instance, the performance discrepancy between GTB and random

Table 5
The top 10 discriminant actions or action patterns in the group of *With Prior Progress* and *Without Prior Progress*. Action patterns that only occurred in either *With Prior Progress* or *Without Prior Progress* are in bold. Here, T denotes tutors and S denotes students. The value of Pearson Residual is used to compare the dispersion of the observed action pattern with the expected occurrence. The action patterns are sorted according to their Pearson Chi-square statistics in each group in a descending order, which indicate the extent to which an action pattern can be used to discriminate the three different categories of dialogues.

		Action Patterns	Pearson residual		
			Gap-clarified	Gap-explained	Gap-bridged
With prior progress	1	(T, Information)-(T, General Positive Feedback)-(T, General Positive Feedback)	-22.87	-3.34	13.62
	2	(T, Probing Question)-(T, General Positive Feedback)-(T, General Positive Feedback)	-19.78	-2.60	11.62
	3	(S, Request Feedback by Image)-(T, Information)-(T, General Positive Feedback)	-20.25	-0.17	10.59
	4	(T, Information)-(T, General Positive Feedback)	-20.25	-0.17	10.59
	5	**(T, General Positive Feedback)-(T, Information)-(T, General Positive Feedback)**	-22.48	-0.96	12.16
	6	(T, General Positive Feedback)-(T, Probing Question)-(T, General Positive Feedback)	-19.67	-1.46	10.97
	7	**(T, Information)-(T, Probing Question)-(T, General Positive Feedback)**	-22.80	-0.99	12.34
	8	**(T, Greeting)-(T, Information)-(T, General Positive Feedback)**	-20.21	-0.06	10.51
	9	**(T, Probing Question)-(T, Probing Question)-(T, General Positive Feedback)**	-21.09	-1.02	11.47
	10	**(T, Information)-(T, Information)-(T, General Positive Feedback)**	-23.41	0.01	12.13
Without prior progress	1	**(T, General Positive Feedback)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-24.35	-4.64	16.33
	2	(T, Probing Question)-(T, General Positive Feedback)-(T, General Positive Feedback)	-24.06	-3.97	15.73
	3	(T, Information)-(T, General Positive Feedback)-(T, General Positive Feedback)	-24.72	-3.97	16.09
	4	**(T, General Positive Feedback)-(T, General Positive Feedback)**	-20.99	-2.02	12.76
	5	**(S, Request Feedback by Image)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-20.98	-2.02	12.76
	6	(T, General Positive Feedback)-(T, Probing Question)-(T, General Positive Feedback)	-23.83	-2.60	14.70
	7	(T, Information)-(T, General Positive Feedback)	-22.13	0.08	11.99
	8	(S, Request Feedback by Image)-(T, Information)-(T, General Positive Feedback)	-22.12	0.09	11.99
	9	**(S, Yes-No-Answer)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-22.07	-2.61	13.75
	10	**(T, Probing Question)-(T, General Positive Feedback)**	-20.69	-0.13	11.35

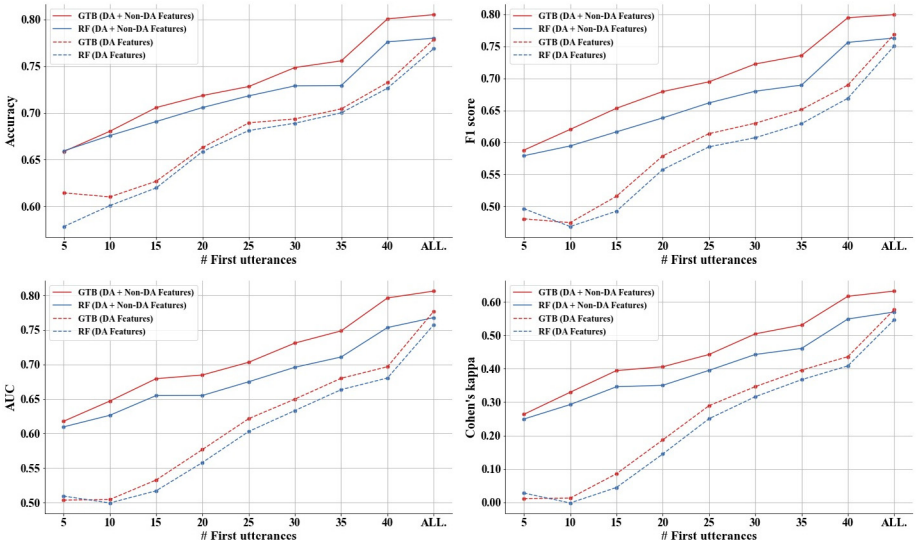


Fig. 1. The performance of GTB and random forests in predicting student performance in solving problems.

forests measured by the F1 score remained rather stable regardless of the number of input utterances. We can make similar observations when scrutinizing other evaluation metrics. When taking all of the available utterances as input, GTB achieved the performance of 0.779, 0.769, 0.577, and 0.777 as measured by accuracy, F1 score, AUC, and Cohen's κ, respectively. These results imply there is still space to further boost the prediction performance. Therefore, we further incorporated the non-DA features together with the DA features as input to the models for student performance prediction (the solid lines in Fig. 1). Unsurprisingly, the results indicate that both GTB and random forests achieved better prediction performance when taking both DA and non-DA features into account. If we take GTB ($N = 10$) as an example, we can see that the Accuracy was boosted from 0.610 to 0.680 and the F1 score was boosted from 0.475 to 0.620. The results of GTB for $N = ALL$ showed that the model achieved Cohen's κ score of 0.632, which indicates a substantial prediction performance. This suggests that, though being useful in characterizing different categories of tutorial sessions, tutors' tutoring actions and action patterns were insufficient in revealing students' problem-solving performance in online one-on-one tutoring. To gain a better understanding of the distinctive predictive power of different types of features, we further conducted an ablation test. That is, the contribution made by a feature is calculated as the difference between the prediction performance of a model when including the feature and that when excluding the feature [66]. Due to the limited space, we presented the results of the ablation test in the electronic appendix.[2] We found that the N-grams features (e.g., terms and phrases like "*great*" and "*good job*" in the positive feedback provided by tutors) were of particular importance in predicting students' problem-solving performance.

5. Discussion and conclusion

The construction of dialogue-based ITS with adequate pedagogic expertise is a longstanding task in the pathway towards

[2] Accessible via https://github.com/bertDA/BertDA/blob/main/DA_Appendix.pdf.

delivering on-time, personalized, and meaningful learning experiences to students. Though quite some studies have been carried out, these studies often ignored the analysis of unsuccessful tutorial sessions and seldom paid attention to the reasons behind these unsuccessful tutorial sessions. This motivated us to analyse a large-scale dialogue corpus (over 14 K), which consisted of both successful and unsuccessful online human–human tutorial sessions, to identify frequent tutoring strategies adopted by tutors and further use these tutoring strategies for predicting students' problem-solving performance. Through extensive analysis and evaluations, our study provided empirical evidence to support existing good practices for developing dialogue-based ITS and contributed to the research of educational dialogue analysis with the following main findings:

- Overall, tutors often took actions to provide feedback and information to students and to raise questions to guide students to solve problems. Correspondingly, students took more actions in answering questions or expressing agreement with tutors and acceptance of the provided explanations or solutions.
- In tutorial sessions where students delivered correct or partially correct solutions to the problems, tutors tended to ask more thought-provoking questions, offer more information hints, and pose less irrelevant statements or questions to students compared to tutorial sessions with lower problem-solving performance.
- If a student had made certain progress in solving a problem before entering a tutorial session, the student was likely to receive fewer hints and thought-provoking questions from a tutor, but still had a higher chance to successfully solve the task and received more positive feedback from the tutor.
- Positive feedback expressed by tutors can be used as a strong discriminator to differentiate dialogues of different student performance in solving problems.
- DA and DA patterns alone were insufficient to reveal the problem-solving performance of students and should be utilized together with other relevant factors (e.g., the informativeness, complexity, and the sentimental polarities of the utterance).
- We have released a DA classifier, which was constructed by applying state-of-the-art pre-trained language model BERT, to better support researchers for relevant research studies, which is accessible via https://github.com/bertDA/BertDA.

5.1. Implications

Firstly, utterances with tags such as `Information`, `Probing Question`, `Operational Questions`, and `General Positive Feedback` can be used to characterize tutorial sessions with different level of student performance. As reported in Table 3, the top three most frequently used tutorial dialogue actions taken by tutors in our dataset were `General Positive Feedback`, `Information`, and `Probing Question`. These tutorial actions, especially `General Positive Feedback` such as "*Correct!*" and "*Great!*", were more frequently observed in successful tutorial sessions (i.e., Gap-bridged), which is also evident in Tables 4 and 5. In fact, this is in line with the findings presented by previous studies [11,24,67], which suggested that the positive feedback provided by tutors played a vital role in verifying the correctness of the students' work, increasing a students' level of self-efficacy in accomplishing a learning task and encouraging the student to proceed with the remaining activities [67]. This implies that tutors providing online one-on-one tutoring service may consider, whenever it is appropriate, providing timely and adequate positive feedback to acknowledge students' achievements and further motivate them to deliver correct solutions. Though it might be possible that Gap-bridged students generally had a higher level of prior knowledge than their counterparts of the other two categories, which enabled them to be more likely to solve problems and thus received more positive feedback from tutors [68]. Future research should investigate the impact of students' prior knowledge on solving learning problems and what strategies should be utilized to better scaffold students with a low level of prior knowledge in our future research.

Secondly, the use of strategies like `Probing Question` and `Information` should be dependent on a student's level of prior progress. For instance, as shown in Table 3, *With-Prior-Progress* students generally received less `Information` hints than their *Without-Prior-Progress* counterparts. This observation is in line with the findings presented in [67], i.e., tutors should provide more scaffolding to students with little progress being made. Besides, we observed that, compared to Gap-explained dialogues, there was a lower usage of `Information` in Gap-bridged dialogues. Previous studies (e.g., [67,69]) showed that, to effectively engage a student in a learning task, tutors should avoid scaffolding the student with excessive information hints. Given that it might be rather challenging for tutors to determine a suitable amount of information hints to be provided to students in practice, we plan to develop automatic methods to measure the levels of both confusion and engagement of students and help tutors (or dialogue-based ITS) better direct their teaching efforts.

Thirdly, the failure of a tutorial session might not be entirely attributed to the extra use of DA tags that were not directly related to solving a learning problem. Table 3 shows that there was a higher occurrence of `Extra Domain Other` and `Operational Question` in Gap-clarified dialogues than the other two categories. To investigate the underlying reasons causing the use of such extra problem-solving-irrelevant utterances, we manually checked 200 randomly selected Gap-clarified dialogues which contained utterances tagged as `Extra Domain Other` and `Operational Question`, and found several issues. A common one is that tutors did not give enough time to a student to think and work on a problem and frequently asked operational questions like "*Are you working on the problem?*", which caused extra pressure to the student and further impeded her from solving the problem. Another common issue is that there were communication issues between tutors and students (e.g., "*I am sorry. I don't understand what you are trying to say*" and "*Let's clarify the information a bit*"). These communication issues, oftentimes, made a student quit a tutorial session before being able to receive any meaningful guidance. In addition, we observed there was a small portion of dialogues (about 12%) in which a student played against the rules (e.g., a student uttered "*Just give me the answer!*" and then a tutor replied with "*I know you might be frustrated, but handling out easy answers goes against our pledge and hurts you in the long run*") or a tutor did not provide timely responses (e.g., "*Sorry for the late reply*"), which also tended to make the student terminate the tutorial session before receiving help. These findings indicate that, when providing online tutoring service or designing dialogue-based ITS, appropriate methods should be developed to (i) remind tutors to allow students to have enough time to work on problems and provide timely feedback to them, (ii) facilitate the communication process between tutors and students (e.g., using a digital whiteboard), and (iii) provide students with a clear guideline (e.g., it is disallowed to directly share answers with students) to avoid potential misunderstandings held by certain students about the terms of service of the tutoring service and system.

Fourthly, GTB can be used to detect potentially successful tutorial sessions in real-world tutoring practices. As depicted in Fig. 1, GTB was capable of accurately classifying about 68% tutorial dialogues by only taking the first 10 utterances as input.

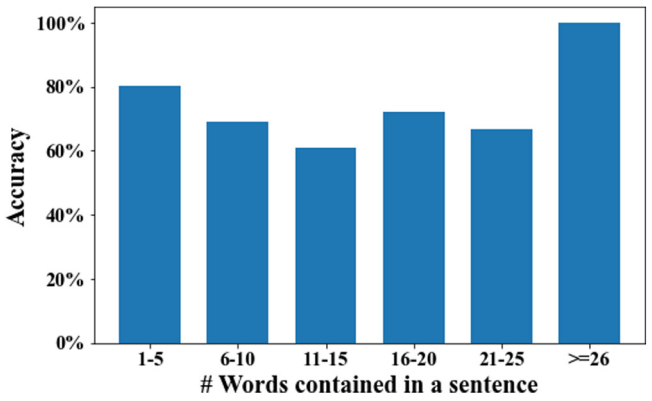


Fig. 2. The classification accuracy for sentences of different length.

With more utterances taken into account, the prediction accuracy kept increasing. This demonstrated that GTB is a reliable machine learning method that can be of practical use to locate potentially unsuccessful dialogues from effective ones. With these potentially unsuccessful dialogues detected in the early stage of a tutorial session, we expect that more interventions (e.g., changing a tutoring strategy) can be taken to better engage and assist a student.

5.2. Limitations

Misclassified Dialogue Acts. Recall that the DA classifier we presented in Section 3.3 was only able to deliver correct labels for 75% sentences in the labelled dataset, which suggests there is still certain improvement space. To gain a more in-depth understanding of the performance of the DA classifier, we calculated the prediction accuracy for each DA tag and found that some tags were more likely to be correctly identified by the classifier than others. For instance, the classifier was capable of distinguishing tags with characteristic keywords that rarely occurred in other tags, e.g., `Greeting` with keywords like "*hello*" and "*welcome*" and `General Positive Feedback` with keywords like "*great*", "*correct*", and "*awesome*". On the other hand, we noticed that tags with relatively fewer training samples were more likely to be misclassified. For instance, the DA `Elaborated Positive Feedback`, which was related to only 2.08% sentences in our labelled dataset, was often misclassified as `General Positive Feedback`. Though `Elaborated Positive Feedback` can also be characterized by keywords like "*great*" and "*correct*", sentences of this DA tag often contained more fine-grained information (e.g., "*Your formula for period is correct!*"). Our DA classifier was unable to distinguish this minor difference when there was a lack of enough training samples of `Elaborated Positive Feedback`. This suggests, for the future improvement of DA identification, it is worthwhile to label more tutorial sentences, especially those with tags of a low occurrence frequency, to enable the classifier to capture the fine-grained difference between various DA tags.

Next, as an initial step to investigate whether the amount of information contained in a sentence would likely impact the prediction performance of the classifier, we treated the number of words contained in a sentence as a proxy of the amount of information conveyed in the sentence, and plot the classification accuracy for sentences of different length in Fig. 2. Interestingly, we observed that, in our case, the classifier tended to deliver better performance when a sentence is particularly short ($[1, 5]$) or particularly long ($>= 26$) than the other sentences. However, considering that, as demonstrated before, the classifier tended to deliver varying performance when dealing with different DA tags, we could not conclude that the particularly short or long sentences were favoured by the classifier. In the future, it would be worthwhile to label more data for each DA tag, especially DA tags associated with sentences of varying lengths, and further investigate the impact of sentence length on the prediction performance of the classifier.

Table 6
Examples of the predicted DA tags delivered by the BERT-based classifier. Here are the abbreviation of the DA tags: Oprt-Ques (`Operational Question`), Extr-Dom-Othr (`Extra Domain Other Statement`), Y-N-Ansr (`Yes-No Answer`), and Ack (`Acknowledge`).

Role	Sentence	Actual	Prediction
Tutor	Anything else I can help you with?	Oprt-Ques	Oprt-Ques
Student	Awesome!	Extr-Dom-Othr	Y-N-Ansr
Student	That should be all thank you	Y-N-Ansr	Ack

Lastly, we observed that, for certain sentences, the wrong identification of DA tags could be explained by the lack of enough contextual information. As described in Section 3.3, when predicting the DA tag for a sentence, the text of the preceding sentence was incorporated as part of the input to capture relevant contextual information. However, as shown in Table 6, the contextual information related to a sentence might span more than one preceding sentence. That is, the student's response "*That should be all thank you*" was uttered to answer the question raised by the tutor ("*Anything else I can help you with?*"), but this response was preceded by another response uttered by the student (i.e., "*Awesome!*"). Therefore, the question raised by the tutor was not taken into account and the DA tag of "*That should be all thank you*" was misclassified as `Acknowledge`. This suggests that, to further improve the prediction performance of the DA classifier, it might be worthwhile to take additional contextual information (i.e., more than one proceeding sentence) into account.

Students' Performance Analysis *Firstly*, the categorical labels of the tutorial dialogues, i.e., Gap-clarified, Gap-explained, and Gap-bridged, were derived by one educational expert. Though a sanity check, which involved a second educational expert to use the same coding rules to label 500 dialogues randomly selected from the whole dataset, was conducted and a percentage agreement score of 0.884 was reached, future research efforts should be allocated to explore other methods to enhance the validity of the labelling results (e.g., employing crowd-sourcing workers to label the whole dataset). *Secondly*, as indicated before, students in certain dialogues played against the rules by asking tutors to directly share answers with them or students quit tutorial sessions because of not receiving timely responses from tutors. After manually scrutinizing 200 randomly-selected Gap-clarified dialogues, we found 12% were of this kind. We leave the automatic identification and exclusion of such dialogues for more fine-grained analysis for the future research. *Thirdly*, it has been widely recognized that the prior knowledge level of a student can significantly impact her performance in a learning task [70]. For instance, students with high prior knowledge, compared to those with low prior knowledge, are able to solve a learning problem with less information hints [71]. However, the information about students' prior knowledge level was not available in our dataset. As a remedy, we took into the level of prior progress achieved by a student before entering a tutorial session and demonstrated similar findings, i.e., students with prior progress were able to solve problems with less information hints compared to their counterparts without any prior progress. In the future, it would be worthwhile to further distinguish students' prior progress in a more fine-grained level (e.g., without progress, with small progress, and with much progress) and analyse its impact on

students' problem-solving behaviours and performance. *Fourthly*, we did not separate the analysis of dialogues of different subjects (i.e., math, chemistry and physics) in this study. Given that about 92% of the dialogues in our current dataset related to math tutoring, we plan to collect more dialogues of physics and chemistry to enrich the dataset, separate the dialogues of different subject areas for analysis, and further provide more in-depth insights to support the tutoring practices in different subject areas in our future work. *Fifthly*, though our work successfully revealed tutoring strategies that were frequently observed in both successful and unsuccessful tutorial dialogues, it still remains largely unknown when these tutoring strategies should be or should not be used. To further guide the development of future dialogue-based ITS, future research should focus on the content analysis of the utterances related to each DA tag and investigate the relationship between these DA tags and other relevant tags, e.g., whether students' statements specifying their confusion (i.e., the DA tag `Not Understanding`) always triggers an action from tutors to provide hints (i.e., the DA tag `Information`) and whether the continued use of such hint-providing actions likely promotes better student performance. For this purpose, causal models [72, 73] can be explored in future research. *Lastly*, though GTB was demonstrated to be effective in predicting students' problem-solving performance, there is still space to further improve the prediction performance. Given the wide success achieved by deep neural networks in tackling various types of tasks, especially those in the field of natural language processing, future research should investigate methods based on deep neural networks to deliver more accurate student performance predictions.

CRediT authorship contribution statement

Jionghao Lin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Shaveen Singh:** Conceptualization, Methodology, Validation, Formal analysis, Data curation. **Lele Sha:** Software, Validation, Data curation. **Wei Tan:** Software, Validation. **David Lang:** Resources. **Dragan Gašević:** Conceptualization, Validation, Supervision, Writing – original, Writing – review & editing, Project administration, Funding acquisition. **Guanliang Chen:** Conceptualization, Methodology, Validation, Supervision, Writing – original, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Mitrovic, Learning SQL with a computerized tutor, in: Proceedings of the Twenty-Ninth SIGCSE Technical Symposium on Computer Science Education, SIGCSE '98, ACM, 1998, pp. 307–311.

[2] S. Ritter, J. Anderson, M. Cytrynowicz, O. Medvedeva, Authoring content in the PAT algebra tutor, J. Interact. Media Educ. 1998 (2) (1998).

[3] S. Schiaffino, P. Garcia, A. Amandi, eTeacher: Providing personalized assistance to e-learning students, Comput. Educ. 51 (4) (2008) 1744–1754.

[4] V. Rus, S. D'Mello, X. Hu, A. Graesser, Recent advances in conversational intelligent tutoring systems, AI Mag. 34 (3) (2013) 42–54.

[5] M.W. Evens, J.A. Michael, One-on-One Tutoring by Humans and Computers, 2006.

[6] A.C. Graesser, P. Chipman, B.C. Haynes, A. Olney, AutoTutor: an intelligent tutoring system with mixed-initiative dialogue, IEEE Trans. Educ. 48 (2005) 612–618.

[7] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, G. Campbell, BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics, IJAIED 24 (3) (2014) 284–332.

[8] K. VanLehn, A.C. Graesser, G.T. Jackson, P.W. Jordan, A. Olney, C.P. Rosé, When are tutorial dialogues more effective than reading? Cogn. Sci. 31 (1) (2007) 3–62.

[9] A. Almasri, A. Ahmed, N. Al-Masri, Y.S.A. Sultan, A.Y. Mahmoud, I. Zaqout, A.N. Akkila, S.S. Abu-Naser, Intelligent tutoring systems survey for the period 2000- 2018, Int. J. Acad. Eng. Res. 5 (3) (2019) 21–37.

[10] A. Alkhatlan, J. Kalita, Intelligent tutoring systems: A comprehensive historical survey with recent developments, Int. J. Comput. Appl. 181 (43) (2019) 1–20.

[11] N. Maharjan, V. Rus, D. Gautam, Discovering effective tutorial strategies in human tutorial sessions, in: The Thirty-First International Flairs Conference, 2018.

[12] N. Maharjan, V. Rus, A tutorial Markov analysis of effective human tutorial sessions, in: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 30–34.

[13] A.C. Graesser, K. VanLehn, C.P. Rosé, P.W. Jordan, D. Harter, Intelligent tutoring systems with conversational dialogue, AI Mag. 22 (4) (2001) 39–51.

[14] K. Cotton, Classroom questioning, School Improv. Res. Ser. 5 (1988) 1–22.

[15] B. Du Boulay, R. Luckin, Modelling human teaching tactics and strategies for tutoring systems: 14 Years on, IJAIED 26 (1) (2016) 393–404.

[16] V.J. Shute, SMART: Student modeling approach for responsive tutoring, User Model. User-Adapt. Inter. 5 (1) (1995) 1–44.

[17] S. Katz, P.L. Albacete, A tutoring system that simulates the highly interactive nature of human tutoring., J. Educ. Psychol. 105 (4) (2013) 1126–1141.

[18] J. Paladines, J. Ramirez, A systematic literature review of intelligent tutoring systems with dialogue in natural language, IEEE Access 8 (2020) 164246–164267, http://dx.doi.org/10.1109/ACCESS.2020.3021383.

[19] V. Rus, N. Maharjan, L.J. Tamang, M. Yudelson, S. Berman, S.E. Fancsali, S. Ritter, An analysis of human tutors' actions in tutorial dialogues, in: The Thirtieth International Flairs Conference, 2017.

[20] K. VanLEHN, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educ. Psychol. 46 (4) (2011) 197–221.

[21] S. Sundararajan, S. Nitta, Designing engaging intelligent tutoring systems in an age of cognitive computing, IBM J. Res. Dev. 59 (6) (2015) 10:1–10:9.

[22] V. Rus, R. Banjade, N. Niraula, E. Gire, D. Franceschetti, A study on two hint-level policies in conversational intelligent tutoring systems, in: Innovations in Smart Learning, Springer, 2017, pp. 175–184.

[23] K.E. Boyer, R. Phillips, A. Ingram, E.Y. Ha, M. Wallis, M. Vouk, J. Lester, Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach, Int. J. Artif. Intell. Educ. 21 (1–2) (2011) 65–81.

[24] A.K. Vail, K.E. Boyer, Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes, in: ITS, 2014.

[25] M. Kumar, M. Kumar, et al., XGBoost: 2D-object recognition using shape descriptors and extreme gradient boosting classifier, in: Computational Methods and Data Engineering, Springer, 2021, pp. 207–222.

[26] A.C. Ornstein, T.J. Lasley, Strategies for Effective Teaching, Harper & Row New York, 1990.

[27] A.C. Graesser, N. Person, D. Harter, T.R. Group, et al., Teaching tactics in AutoTutor, Modell. Hum. Teach. Tact. Strateg. IJAIED 11 (2000) 1020–1029.

[28] J.R. Anderson, The Architecture of Cognition, Vol. 5, Psychology Press, 1996.

[29] J.R. Anderson, A.T. Corbett, K.R. Koedinger, R. Pelletier, Cognitive tutors: Lessons learned, J. Learn. Sci. 4 (2) (1995) 167–207.

[30] I. Arroyo, J.E. Beck, B.P. Woolf, C.R. Beal, K. Schultz, Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism, in: International Conference on Intelligent Tutoring Systems, Springer, 2000, pp. 574–583.

[31] J.G. Cromley, What do reading tutors do? A naturalistic study of more and less experienced tutors in reading, Discourse Process. 40 (2) (2005) 83–113.

[32] J.C. Marineau, P.M. Wiemer-Hastings, D. Harter, B.A. Olde, P. Chipman, A. Karnavat, V. Pomeroy, S. Rajan, A. Graesser, Classification of Speech Acts in Tutorial Dialog, 2000.

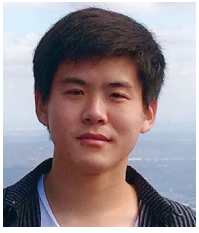
[33] R. Pilkington, Analysing Educational Discourse: The DISCOUNT Scheme, University of Leeds, Computer Based Learning Unit, 1999.

[34] S. Hennessy, S. Rojas-Drummond, R. Higham, A.M. Márquez, F. Maine, R.M. Ríos, R. García-Carrión, O. Torreblanca, M.J. Barrera, Developing a coding scheme for analysing classroom dialogue across educational contexts, Learn. Cult. Soc. Interact. 9 (2016) 16–44.

[35] A.C. Graesser, N.K. Person, J.P. Magliano, Collaborative dialogue patterns in naturalistic one-to-one tutoring, Applied Cognitive Psychology 9 (6) (1995) 495–522.

[36] K. Forbes-Riley, D. Litman, Investigating human tutor responses to student uncertainty for adaptive system development, in: Conference on ACII, Springer, 2007, pp. 678–689.

[37] K.E. Boyer, R. Phillips, A. Ingram, E.Y. Ha, M. Wallis, M. Vouk, J. Lester, Characterizing the effectiveness of tutorial dialogue with hidden markov models, in: ITS, Springer, 2010, pp. 55–64.
[38] A.C. Graesser, S. Lu, G.T. Jackson, H.H. Mitchell, M. Ventura, A. Olney, M.M. Louwerse, AutoTutor: A tutor with dialogue in natural language, Behav. Res. Methods Instrum. Comput. 36 (2) (2004) 180–192.
[39] D. Morrison, B. Nye, B. Samei, V.V. Datla, C. Kelly, V. Rus, Building an intelligent pal from the tutor. com session database phase 1: Data mining, in: Educational Data Mining 2014, Citeseer, 2014.
[40] S. D'Mello, A. Olney, N. Person, Mining collaborative patterns in tutorial dialogues, J. Educ. Data Min. 2 (1) (2010) 1–37.
[41] M.T. Chi, S.A. Siler, H. Jeong, T. Yamauchi, R.G. Hausmann, Learning from human tutoring, Cogn. Sci. 25 (4) (2001) 471–533.
[42] F.J. Rodríguez, K.M. Price, K.E. Boyer, Exploring the pair programming process: Characteristics of effective collaboration, in: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, 2017, pp. 507–512.
[43] A.K. Vail, J.F. Grafsgaard, K.E. Boyer, E.N. Wiebe, J.C. Lester, Predicting learning from student affective response to tutor questions, in: ITS, Springer, 2016, pp. 154–164.
[44] A. Ezen-Can, K.E. Boyer, A tutorial dialogue system for real-time evaluation of unsupervised dialogue act classifiers: Exploring system outcomes, in: AIED, Springer, 2015, pp. 105–114.
[45] J.F. Grafsgaard, et al., Multimodal affect modeling in task-oriented tutorial dialogue., 2014.
[46] P. Robe, S.K. Kuttal, Y. Zhang, R. Bellamy, Can machine learning facilitate remote pair programming? Challenges, insights & implications, in: 2020 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, IEEE, 2020, pp. 1–11.
[47] K.E. Boyer, R. Phillips, M.D. Wallis, M.A. Vouk, J.C. Lester, Learner characteristics and feedback in tutorial dialogue, in: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, ACL, 2008, pp. 53–61.
[48] M.G. Core, J.D. Moore, C. Zinn, The role of initiative in tutorial dialogue, in: EACL, 2003.
[49] K. Forbes-Riley, D. Litman, A. Huettner, A. Ward, Dialogue-learning correlations in spoken dialogue tutoring, in: Proceedings of the 2005 Conference on AIED: Supporting Learning Through Intelligent and Socially Informed Technology, 2005, pp. 225–232.
[50] S. Katz, G. O'Donnell, H. Kay, An approach to analyzing the role and structure of reflective dialogue, IJAIED 11 (2000) 320–343, Part I of the Special Issue on Analysing Educational Dialogue Interaction.
[51] D.E. Meltzer, Relation Between Students ' Problem-Solving Performance and Representational Format, 2005.
[52] H. Muhonen, E. Pakarinen, A.-M. Poikkeus, M.-K. Lerkkanen, H. Rasku-Puttonen, Quality of educational dialogue and association with students' academic performance, Learn. Instr. 55 (2018) 67–79.
[53] F. Yang, F.W. Li, Study on student performance estimation, student progress analysis, and student potential prediction based on data mining, Comput. Educ. 123 (2018) 97–108.
[54] F.J. Dochy, G. Moerkerke, R. Martens, Integrating assessment, learning and instruction: Assessment of domain-specific and domaintranscending prior knowledge and progress, Studi. Educ. Eval. 22 (4) (1996) 309–339.
[55] A. Ezen-Can, J.F. Grafsgaard, J.C. Lester, K.E. Boyer, Classifying student dialogue acts with multimodal learning analytics, in: LAK, 2015, pp. 280–289.
[56] J.F. Grafsgaard, J.B. Wiggins, A.K. Vail, K.E. Boyer, E.N. Wiebe, J.C. Lester, The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring, in: ICMI, 2014, pp. 42–49.
[57] V. Rus, N. Maharjan, R. Banjade, Dialogue act classification in human-to-human tutorial dialogues, in: Innovations in Smart Learning, Springer, 2017, pp. 185–188.
[58] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, Trans. Assoc. Comput. Linguist. 8 (2021) 842–866.
[59] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[60] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on EMNLP and the 9th International IJCNLP, 2019, pp. 3606–3611.
[61] H. Khanpour, N. Guntakandla, R. Nielsen, Dialogue act classification in domain-independent conversations using a deep recurrent neural network, in: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2012–2021.
[62] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[63] G. Chen, D. Lang, R. Ferreira, D. Gasevic, Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues, in: EDM, 2019.
[64] T.G. Dietterich, et al., Ensemble learning, in: The Handbook of Brain Theory and Neural Networks, Vol. 2, MIT press Cambridge, MA, 2002, pp. 110–125.
[65] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL-Int. J. Appl. Linguist. 165 (2) (2014) 97–135.
[66] J. Lin, S. Pan, C.S. Lee, S. Oviatt, An explainable deep fusion network for affect recognition using physiological signals, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, ACM, 2019, pp. 2069–2072.
[67] V.J. Shute, Focus on formative feedback, Rev. Educ. Res. 78 (1) (2008) 153–189.
[68] E.R. Fyfe, B. Rittle-Johnson, M.S. DeCaro, The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters, J. Educ. Psychol. 104 (4) (2012) 1094.
[69] J. Van de Pol, M. Volman, J. Beishuizen, Scaffolding in teacher–student interaction: A decade of research, Educ. Psychol. Rev. 22 (3) (2010) 271–296.
[70] T. Hailikari, N. Katajavuori, S. Lindblom-Ylanne, The relevance of prior knowledge in learning and instructional design, Am. J. Pharm. Educ. 72 (5) (2008).
[71] S. Narciss, S. Sosnovsky, L. Schnaubert, E. Andrès, A. Eichelmann, G. Goguadze, E. Melis, Exploring feedback and student characteristics relevant for personalizing feedback strategies, Comput. Educ. 71 (2014) 56–76.
[72] S. Athey, G.W. Imbens, Machine learning methods for estimating heterogeneous causal effects, Stat 1050 (5) (2015) 1–26.
[73] Y. Luo, J. Peng, J. Ma, When causal inference meets deep learning, Nat. Mach. Intell. 2 (8) (2020) 426–427.



Jionghao Lin is a Ph.D. student in the Centre for Learning Analytics at Monash University, Melbourne, Australia. His primary research interests focus on the areas of learning analytics, natural language processing, and affective computing. Currently, Jionghao is mainly working on applying artificial intelligent technologies to understand and optimize the learning environment. He received his B.E. degree from Jianghan University, China, and Master degree in Data Science from Monash University, Australia.

Shaveen Singh is a Research Fellow at the Centre of Learning Analytics at Monash University. His research interests include the design and deployment of technology to increase the understanding and improve digital learning experiences. More specifically, his work examines the areas of learning analytics, personalized active learning, and building tools for teacher support. Shaveen is currently pursuing his Ph.D. at Monash University.

Lele Sha is a second-year Ph.D. student in the Centre for Learning Analytics at Monash University. His main research interest centres on applying Machine Learning and Natural Language Processing techniques to automatically processing educational forum posts. Specifically, he is focusing on improving model performance by applying extensive feature engineering and sentence embeddings. Before starting his Ph.D., Lele also worked in several software-as-a-service projects on learning management systems, which were successfully deployed to production and currently offering hundreds of online courses on its interactive training platform for Australian students.

Wei Tan is a Doctoral Researcher who studies the cutting-edge machine learning algorithm in Data Science. He specializes in Active Learning that optimize the labelling budget and time for the human annotator. His Ph.D. project is funded by Google Turning point. The aim is to develop the Surveillance System that will enable capture of a more complete set of coded ambulance data relating to SITB, mental health, and AOD attendances to inform policy, practice and intervention. He holds a master's degree from Monash University, and has expertise in analytics design for the social media platform.

David Lang is a doctoral student in the Economics of Education program and an IES Fellow. He graduated from UCLA in 2008 with a B.A. in Economics & a B.S. in Actuarial Mathematics. Prior to his doctoral studies, David worked for five years as a research analyst at the Federal Reserve Bank of San Francisco. His research interests include higher education, online education, and quantitative methods in education research. At Stanford, David also obtained a master's degree in Management Science and Engineering.

Dragan Gašević is Distinguished Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. As the past president (2015–2017) and a co-founder of the Society for Learning Analytics Research, he had the pleasure to serve as a founding program chair of the International Conference on Learning Analytics and Knowledge (LAK) and a founding editor of the Journal of Learning Analytics. His research centres on self-regulated and social learning, higher education policy, and data mining. He is a frequent keynote speaker and a (co-)author of numerous research papers and books.

Dr. Guanliang Chen is serving as a Lecturer in the Faculty of Information Technology, Monash University in Melbourne, Australia. Before joining Monash University, Guanliang obtained his Ph.D. degree at the Delft University of Technology in the Netherlands, where he focused on the research on large-scale learning analytics with a particular focus on the setting of Massive Open Online Courses. Currently, Guanliang is mainly working on applying novel language technologies to build intelligent educational applications. His research works have been published in international journals and conferences including AIED, EDM, LAK, L@S, EC-TEL, ICWSM, UMAP, Web Science, Computers & Education, and IEEE Transactions on Learning Technologies. Besides, he co-organized two international workshops and has been invited to serve as the program committee member for international conferences such as LAK, FAT, ICWL, etc.