# Automated Session-Quality Assessment for Human Tutoring Based on Expert Ratings of Tutoring Success

Benjamin D. Nye*
The University of Memphis
Memphis, TN 38152
benjamin.nye@gmail.com

Donald M. Morrison
The University of Memphis
Memphis, TN 38152
dmmrrson@memphis.edu

Borhan Samei
The University of Memphis
Memphis, TN 38152
bsamei@memphis.edu

## ABSTRACT

Archived transcripts from tens of millions of online human tutoring sessions potentially contain important knowledge about how online tutors help, or fail to help, students learn. However, without ways of automatically analyzing these large corpora, any knowledge in this data will remain buried. One way to approach this issue is to train an estimator for the learning effectiveness of an online tutoring interaction. While significant work has been done on automated assessment of student responses and artifacts (e.g., essays), automated assessment has not traditionally automated assessments of human-to-human tutoring sessions. In this work, we trained a model for estimating tutoring session quality based on a corpus of 1438 online tutoring sessions rated by expert tutors. Each session was rated for evidence of learning (outcomes) and educational soundness (process). Session features for this model included dialog act classifications, mode classifications (e.g., Scaffolding), statistically distinctive subsequences of such classifications, dialog initiative (e.g., statements by tutor vs. student), and session length. The model correlated more highly with evidence of learning than educational soundness ratings, in part due to the greater difficulty of classifying tutoring modes. This model was then applied to a corpus of 242k online tutoring sessions, to examine the relationships between automated assessments and other available metadata (e.g., the tutor's self-assessment). On this large corpus, the automated assessments followed similar patterns as the expert rater's assessments, but with lower overall correlation strength. Based on the analyses presented, the assessment model for online tutoring sessions emulates the ratings of expert human tutors for session quality ratings with a reasonable degree of accuracy.

## Keywords

Automated Assessment, Tutoring Dialog, Dialog Acts, Dialog Modes, Natural Language Processing, Educational Data Mining

*Corresponding Author

## 1. INTRODUCTION

As online learning has expanded, computer-mediated tutoring and help-seeking has become more prevalent and accessible. This tutoring occurs in a variety of forms, ranging from large commercial platforms employing certified teachers down to ad-hoc peer tutoring in rudimentary learning management systems (LMS). These systems generate a wealth of data about human tutoring interactions that can provide significant insights into the processes of online learning, the space of effective tutoring strategies, and the effectiveness of different platforms and contexts for tutoring. However, to study successful tutoring, tools are needed that can help distinguish between more and less successful sessions.

Quality ratings for tutoring sessions are often only available from self-reports by the tutor and student. However, these ratings have significant problems. Students typically have limited metacognitive skills and need training to assess their own learning [17]. Tutors can be more effective judges of learning, but a tutor's assessments of their students' learning can be biased and hard to compare due to these rating biases. Some of these biases may be individual variation (easy vs. hard raters), while others are systematic, such as less-expert tutors reporting higher average learning from their sessions. Other tutoring session sources have no real quality measure. For example, peer tutoring often lacks any assessment of the quality of the tutoring session, and hand-tagging these sessions for quality measures would be very time-consuming.

A standardized, automated estimator for the effectiveness of online tutoring sessions is arguably essential to the analysis of large-scale transcript corpora. Such a tool can be used to identify especially high-rated sessions, to track the results of improvement efforts, and to identify patterns in associated metadata. Also, differences between the automated estimator and tutors' self-reports could be used to identify new features that indicate effective tutoring strategies (i.e., an active learning approach). As such, the iterative improvement of a session success indicator would provide new insights into the features of effective tutoring and how they relate to other sets of data.

In this work, we have used a two-step supervised learning approach to train an estimator for session effectiveness. This

estimator was trained on a corpus of 1438 human-to-human tutoring sessions, where each session was rated in terms of two quality measures and each statement was annotated with a dialog act tag (e.g., *Confirmation:Positive*) and a dialog mode (e.g., Scaffolding). Based on the quality ratings assigned by independent expert tutors, features related to tutoring session success were identified using sequential pattern mining and statistical analysis of high-level session features (e.g., duration). Second, regression models that employed these features were trained to rate the quality of the tutoring sessions. Finally, this model was applied to a large sample of 246k tutoring sessions to examine the consistency of these ratings against metadata associated with each session, such as the original tutor's rating of student learning and of the student's knowledge of necessary prerequisites.

## 2. BACKGROUND AND RELATED WORK

Studying strategies and patterns in tutoring transcripts is a longstanding research area with roots in speech act theory [21]. Key techniques from this literature include dialog act classification [8], identifying dialog modes [1], and identifying statistically significant sequence patterns [3]. Our research described here relies on the use of all three levels of analysis to identify significant features that can be used to assess session quality. Dialog act classification involves binning each tutor or student statement into distinct taxonomy categories, which represent the functional purpose of the statement (e.g., an "Assertion" that states a fact). Dialog act taxonomy distinctions vary depending on the research focus, such as question types [8], higher-level dialog acts and feedback [1], and finer-grained pedagogical acts [3]. Our research extended this prior work in several ways, including a highly granular coding scheme, developed in collaboration with professional online tutors, which will be discussed later.

Dialog modes are a more recent area of focus for machine learning, but their theoretical underpinnings for studying learning are equally mature. In our work, modes represent shared understandings regarding hidden, higher-order dialog states with associated roles and expectations concerning the likelihood and appropriateness of particular dialog acts given that state [16]. In tutoring research, theoretically-based modes typically represent pedagogical strategies, such as Modeling, Scaffolding, and Fading. More recent studies of modes have used unsupervised approaches, such as Hidden Markov Models to detect patterns of dialog acts that match such theoretical modes [1]. However, such discovered states are not always guaranteed to be modes as we frame them here: others likely represent intermediate structures, such as adjacency pairs (e.g., a question followed by an answer). As such, in this research, we have relied on human-tagged modes and supervised mode-classifiers based on such modes, so that each mode can be linked more clearly to theoretical descriptions of pedagogy.

Finally, this research relies on features extracted using sequence data mining. A good review of prior work for sequence mining tutoring transcripts is presented by D'Mello and Graesser [3], which outlines conventional approaches (e.g., association rule mining) as well as a novel method based on transition likelihoods. In general, traditional analyses of tutoring sessions focus on identifying frequent or distinctive dialog act transitions and subsequences. However, where supervised labels exist (e.g., quality tags), alternative sequence analysis techniques can be applied to identify sequences that occur more frequently in certain session types. This type of analysis detects distinctive subsequences, which discriminate between one group of sequences versus another group of sequences [5].

Since online human tutoring is a dyadic interaction, it also has similarities with computer-supported collaborative learning (CSCL). CSCL analysis often considers higher-level constructs related to collaboration, such as reaching consensus and division of tasks [13]. Many of these constructs are less central to a professional tutoring process, which has predefined roles (tutor vs. student) and associated cultural expectations for dialog behavior. However, aspects of these more general interactions were incorporated, such as dialog management (a "Process Negotiation" mode) and interpersonal relationships (a "Rapport Building" mode).

The quality of a tutoring session can be measured in two ways: "objective" assessments, such as tests given to the student [1], or "subjective" assessments, based on expert ratings or tags assigned to the session. However, even objective assessments require subjective decisions about their criteria. Additionally, expert raters can often provide higher granularity for tagging events during the tutoring process. As such, process-focused machine learning often focuses on building classifiers and estimators trained on expert tags and ratings [18]. Our research builds on this approach, so our automated assessments model how expert tutors *perceive* session quality rather than necessarily the resulting learning gains. In future work, we feel that there would be great value in contrasting a session quality assessment trained on tested learning gains against the one developed in this paper. Such an assessment might identify session features that help identify when illusions of mastery and other rating biases occur [6].

## 3. DATA SET

This research analyzes a full data set of 246k online human-to-human tutoring transcripts from a major commercial tutoring service (Tutor.com). Thousands of different tutors, and tens of thousands of different students participated in these sessions, but all focused on Algebra and Physics topics. As an on-demand service, each session was initiated by a student who requested help on a problem or concept (e.g., at an impasse). Of these transcripts, approximately 4k were excluded from analyses on the full data set due to missing data or formatting issues. Each session contained a timestamped line-by-line text transcript of the statements typed by the student, the tutor, and system messages (e.g., file uploads). Every session was also associated with metadata collected before and after the session. This metadata included the tutor's assessment of evidence of learning during the session (EL1) and the tutor's assessment of the student's level of prerequisite knowledge (PREREQ). Metadata was also available for a subset of tutors, which included their "Tutor Level," an internal performance level that ranged from "Probationary" (0) to "Level III" (Highest). The tutor level was determined by each tutor's mentor, based on internal reviews of the tutor's sessions, and is correlated with experience. On average, Level III tutors had five years experience, Level II had two to four years, and Level I had a little over

a year. Probationary tutors averaged 6 months.

Of the total set of transcripts, 1438 sessions were annotated by a panel of 19 subject matter experts (SMEs), selected from a pool of some 2,800 Tutor.com tutors using a rigorous screening process, which included analysis of answers to a set of survey questions designed to gather initial expert opinion about tutoring, and also to assess the respondents' ability to critique session transcripts. The training process and details on inter-rater reliability are described in more detail in related work [15]. As part of the annotation process, the SMEs rated each session on two scales: evidence of learning (EL2) and educational soundness (ES). Annotators were instructed to consider different criteria for each: EL2 targets outcomes (i.e., did the student learn) and ES targets process (i.e., did the tutor use good tutoring strategies). This is important because sometimes good tutoring steps can still fail to produce learning for a given student. EL1, EL2, ES, and PREREQ were all rated on a 0-5 scale, where zero represents a low rating and five represents a top rating.

Each line in the tutoring session was also tagged for a dialog act and was also part of a dialog mode. Given the size of the taxonomies (126 dialog acts and 16 dialog modes), a full review of each tag would be infeasible, so specific tags that showed value as features will be noted as they are discussed. The taxonomy of dialog acts included 126 distinct tags, organized into 15 main categories. At a macro-level, these categories focus on traditional dialog act classes such as Questions, Assertions, Requests, Directives, and Expressives [21]. Within the tutoring context, these categories tend to be used to provide information (Answer, Assertion, Clarification, Confirmation, Correction, Expressive, Explanation, Reminder), asking for information (Hint, Prompt, Question), and managing the tutoring process (Directive, Promise, Request, Suggestion). Within each of the 15 main categories, subtypes capture key differences such as positive versus negative feedback (e.g., *Expressive:Positive* vs. *Expressive:Negative*).

Annotators also tagged student or tutor contributions that signaled the start of a dialog mode, or a switch from one dialog mode into another. The 16 included modes associated with classic tutoring strategies (Fading, Modeling, Scaffolding, Sensemaking, Session Summary, Telling), identifying the problem (Method Identification, Problem Identification) or learner prerequisites (Assessment), interpersonal strategies (Metacognitive Support, Rapport Building), and session process (Process Negotiation, Opening, Closing, Method Road Map, Off Topic). The time spent in each mode was far from uniform. Tutoring strategy modes, particularly Scaffolding, accounted for a majority of most sessions. Session process modes were also significant, such as Process Negotiation (i.e., getting on the same page), Openings, and Closings. Other modes were fairly rare, such as Method Identification.

Based on these annotated tags, complementary research on this data set developed a logistic regression dialog act classifier [20] and a conditional-random fields (CRF [11]) mode classifier [19]. This tagging methodology followed similar principles to Moldovan et al. [14]. These classifiers approached the level of reliability shown by independent tagging by human experts, as noted in Table 3. The figures in this table show the best performance by both the human taggers (i.e., their final inter-rater reliability tests) and the performance of the classifiers used for automated tagging in this paper. Machine tagging statistics shows cross-validation results. As can be observed, the classification of the main dialog acts (15 categories) and full set of sub-acts (126 categories) approximated human inter-rater tagging fairly closely. Classifying modes was fairly effective also, but lost nearly half of its accuracy the tagger trained on human speech act tags was applied to the machine-labeled dialog acts (29% accuracy). Retraining on machine tags before testing on machine tags improved overall accuracy, but still produced a significantly lower kappa (43% and 0.21, respectively, as shown in parentheses), as compared to training and testing on human tags. As such, mode tags will be less accurate for machine-tagged sessions.

**Table 1: Reliability Scores for Tagging**

| | Main Act | | Sub-Act | | Mode | |
|---|---|---|---|---|---|---|
| Tagger | Acc | Kappa | Acc | Kappa | Acc | Kappa |
| Human | 81% | 0.77 | 65% | 0.63 | 56% | 0.47 |
| Machine | 77% | 0.71 | 53% | 0.50 | 57% (43%) | 0.52 (0.21) |

From the standpoint of analysis, the 1,438 human-tagged training set was used for initial feature identification and training of the session quality assessment model. The full set of 242k machine-tagged sessions were then treated as a second data sample for analysis, which included the original training set but tagged using the automated dialog act and mode classifier models. This research builds on the prior research that developed dialog act classifiers [20] and mode classifiers [19], as well as development of a taxonomy for speech acts and modes in human tutoring [15]. The novel contributions reported in this paper include identifying patterns in speech acts and modes (subsequence analysis), identifying features that help estimate tutoring session quality, training machine learning models that estimate tutoring session quality, examining the strength of features in these models, and examining the correlation between estimated session quality against other indicators of session quality (e.g., the original tutor's rating of learning during the session). This work was done to target the research questions described in the following section.

# 4. RESEARCH METHODOLOGY

Based on these data sets, this work approaches five primary research questions:

1. How closely can we model expert judgments about session quality, based on domain-independent dialog acts and modes?
2. What models show the most promise for assessing session quality?
3. What features are the strongest predictors in these models?
4. What features lose predictive power when trained on machine tags rather than human tags?

5. How closely do the results from machine quality tags correlate with metadata on the full corpus (e.g., EL1), as compared to the training corpus?

To examine these questions, a session quality classifier was trained using a two-step process of feature selection followed by supervised learning. First a set of high-level features was selected that correlated with the rater's evidence of learning (EL2) and educational soundness (ES). These features included the duration of the session, the average number of words typed by the student per contribution (verbosity), the number of dialog acts typed by the tutor and by the student, and the number of short and long pauses between dialog acts. Additionally, the counts of each mode tag and of each individual dialog act by a given speaker were used as features (e.g., *Confirmation:Positive [Tutor]*).

Next, to capture more complex features of the tutoring process, sequence pattern mining was applied to tutoring sessions to identify subsequences of dialog acts or dialog modes that help distinguish between excellent and poor tutoring sessions. For this analysis, two subsets of human-annotated tutoring sessions were selected that included the most successful sessions (N=261, where $ES = 5$ and $EL2 = 5$) and the least successful sessions (N=93, where $ES <= 2$ and $EL2 <= 2$). Subsequences of dialog modes consider dialog mode switches, where there was a change from one mode to another. This is important because modes often span multiple dialog acts.

The subsequence analysis used the TraMiner package for sequence analysis [5], which contains an algorithm for detecting discriminant event subsequences between two groups of sequences. At a high level, this algorithm calculates the frequency of all subsequences up to a given length for each group of sequences, then applies a Chi-squared test (Bonferroni-adjusted) to identify subsequences that are statistically more (or less) frequent in each group. In this context, a subsequence must be distinguished from a substring: subsequences are ordered, but do not necessarily have to be contiguous. Three sets of distinctive subsequence analyses were performed: 1) dialog act subsequences, 2) mode subsequences, and 3) dialog acts within each type of mode. Any subsequence which was distinctive at the $p<0.4$ level was included as a candidate feature. The $p<0.4$ cutoff was selected to allow a large set of candidate features, while still likely performing better than chance. This analysis was performed on the human-annotated tags. Each subsequence was treated as a feature whose incidence would be counted within a session (i.e., a count of the number of times that tags occurred in that order, without reusing any tags).

Four algorithms were trained to estimate the average of ES and EL2 based on the full feature set: linear regression with feature selection, support vector machine (SVM) regression [10], and additive regression based on decision stumps [4]. In general, these algorithms were selected and tuned to try to avoid over-fitting: the final number of active candidate features was 1465, which was comparable to the number of training sessions (1438). Ridge regression reduces the number of parameters by penalizing additional factors. Support Vector Machines are resistant to overfitting because they regularize the space solution space. Additive regression (also called Stochastic Gradient Boosting) uses smoothing that reduces the impact of each additional factor. Each algorithm was evaluated using 10-fold cross validation, using Weka [9]. After evaluating the effectiveness of each algorithm on the human-annotated data, the best of these algorithms was then tested on the machine-tagged sessions to examine performance. The best algorithm was re-trained using machine-tagged sessions, to test if calibrating to the dialog act and mode classifier outputs would improve performance.

Finally, the full set of 242k tutoring sessions was tagged using the best-fit model for session quality. These quality tags were correlated against session metadata available for the larger corpus of sessions: the original tutor's evidence of learning (EL1), the original tutor's assessment of the student's prerequisite knowledge (PREREQ), and the level of the tutor (Tutor Level). These correlations were compared against the correlations observed between the automated assessments and these same metadata variables for the training set. The goal of this analysis was to examine the consistency of the automated assessment with other ratings of session quality that were available for all tutoring sessions.

# 5. RESULTS AND DISCUSSION

The results from each step are discussed in this section, including sequence mining for session features, training and evaluating the session assessment model, and applying this model to a large corpus of online tutoring session transcripts. For the sake of brevity, dialog acts in this section are displayed using the shorthand form <Main Dialog Category>: <Sub Act> [<Speaker>], such that *Expr:Praise [T]* means "expressive praise from the tutor."

## 5.1 Sequence Pattern Mining

Discriminate sequence analysis that compared the most successful and least successful tutoring sessions identified 1151 better-than-chance ($p<0.4$) distinctive subsequences from 2 to 7 elements long. The majority of these sequences were sequences of dialog acts (1062) and a significant number of these sequences captured variations on similar patterns. Due to the granularity of the taxonomy, distinctions occurred such as *Assertion:Calculation [S]⇒ Expressive: Confirmation: Positive [T]* versus *Assertion:Calculation [S]⇒ Confirmation:Positive [T]*, where the only difference was whether the tutor's feedback took the form of an Expressive. Moreover, such distinctions sometimes showed slightly higher distinctiveness. For example, in the above case, *Expressive:Confirmation: Positive* feedback (e.g., *"Great!"*) was a stronger indicator of session success than *Confirmation: Positive* (e.g., *"Right"*).

A total of 89 distinctive mode subsequences were identified as candidate features that distinguished between session quality. Many of these were variants of eight patterns that were supported by Bonferroni-adjusted Chi-squared tests at the $p<0.05$ level. Six of these patterns were indicators of positive sessions. 1) Successful sessions almost always ended with a Closing/WrapUp, suggesting that both the tutor and student are satisfied with the progress. 2) Successful sessions had more Fading. The existence of even one Fading segment was an indicator of success, though Scaffolding preceding Fading was a better indicator; 3) Successful ses-

sions tended to have repeated Scaffolding or Sensemaking segments (the conceptual equivalent of Scaffolding), where Scaffolding was interleaved with other modes. 4) Successful sessions were more likely to have late-session Rapport Building is after Scaffolding or Fading, but preceding the Closing. 5) A Telling mode (i.e., mini-lecture) before Rapport Building was also a positive feature, which likely indicates that a summary is positive. 6) The presence of a single Opening mode was also an indicator of a good session, where less-successful sessions skipped the Opening greetings and moved immediately to Problem Identification.

Two patterns of mode subsequences tended to be associated with less successful tutoring sessions. 1) Unsuccessful sessions tended to have repeated Modeling mode cycles. While a single Modeling mode segment was not indicative of a poor session, two or more in series was associated with worse ratings. 2) Unsuccessful sessions were also indicated by repeated Process Negotiation, particularly if Process Negotiation alternated with Modeling (the tutor solving the problem) or Problem Identification (figuring out what problem the student has). It was also a negative indicator when Process Negotiation started early in a session sequence. Process Negotiation is a mode that is associated with discussing the tutoring process itself, which includes figuring out who should be speaking or addressing technical issues. Process Negotiation itself was not a bad mode, and was also present in many good characteristic sequences. In these good sequences, it tends to occur late in the session (preceding a Closing) rather than early-on. In general, long or early cycles of Process Negotiation likely indicate that the student is unable to contribute meaningfully to the problem due to lack of prerequisites, technical issues, or poor dialog coordination (e.g., student interrupting).

From aligning these distinctive subsequences, an ideal path of modes for a session might be framed as: Opening ⇒ ProblemID ⇒ Scaffolding ⇒ Fading⇒ ProcessNegotiation ⇒ Telling ⇒ RapportBuilding ⇒ Closing, where some modes (e.g., Scaffolding and Fading) optimally alternate multiple times. This successful mode sequence shows some similarities and differences when compared to Graesser et al.'s 5-step frame for in-person tutoring, which can be described as: [Tutor poses a question]⇒ [Student attempts to answer]⇒ [Tutor provides brief feedback]⇒ [Collaborative interaction]⇒ [Tutor checks if student understands] [7]. The final two frames align well with Scaffolding⇒ Fading⇒ ProcessNegotiation pattern observed in the successful online sessions. The main differences likely stem from the tutoring context. The Graesser tutoring frame assumes a tutor-driven process in which the student is attempting to answer a question, typically conceptual, posed by the tutor. In our data, the student is typically coming to the tutor for help on a specific problem, and the session is in this sense student-driven. As such, Problem Identification occurs first, instead of the tutor posing an initial question.

The insights from the dialog act sequences for successful versus less successful sessions show similar patterns as those based on sequences of modes. However, they are more granular and some of the distinctive sequences tend to be longer or repeating (e.g., repeated answers by a student alternating with *Confirmation:Positive* by the tutor are better). These patterns match loosely to the learning-relevant affective states noted by D'Mello and Graesser [2], which were: Achievement, Engagement, Disengagement, Confusion / Uncertainty, and Frustration. Evidence of achievement (i.e., answers that received positive feedback, explanations followed by expressions of understanding) corresponded with higher session ratings. Likewise, engagement (student answer attempts and sequences with multiple student statements) were positive.

Disengagement indicators, such as questions followed by *Expressive:LineCheck* (e.g., *"Are you there?"*) and *Expressive:Neutral* statements by the student (e.g., *"ok"*) were associated with lower ratings. Raters likely interpreted neutral responses as indicating that the learner was passively processing the session. By comparison, tutor questions that transitioned to *Confirmation:Understanding:Negative* (e.g., *"No, I don't understand"*) were not strong indicators of an unsuccessful session. Frustration was not significantly observed in the corpus, in part due to a lack of taxonomy tags devoted to detecting it and in part due to a relatively low prevalence of obvious frustration within the training corpus. While taxonomy acts for confusion and uncertainty were available in the taxonomy, these were less common and did not have a clear correlation to successful or unsuccessful sessions. This is somewhat expected, since a limited amount of confusion tends to be productive [2], but a large amount can lead to unproductive frustration. More nuanced techniques might be needed to monitor these cycles in tutoring sessions.

## 5.2 Automated Assessment Models

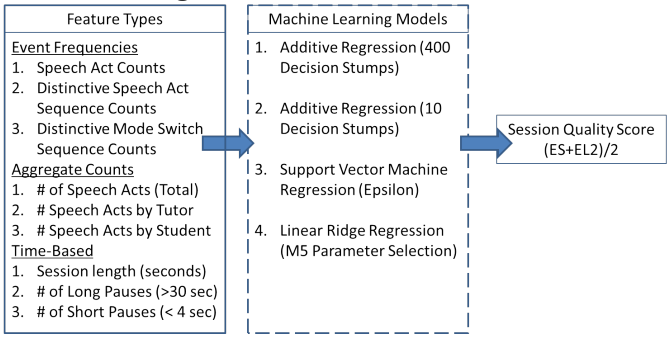
The total feature set was used to train a series of machine-learning models: linear ridge regression with parameter selection (Linear), SVM regression (SVM), and additive regression with decision stumps (Add.). The outcome variable for this training was a unified quality score based on the average of the rater's assessment of educational soundness (ES) and evidence of learning (EL2). The process for training these models is outlined in Figure 1. The results of 10-fold cross-validation for the best-fit models are presented in Table 5.2, in terms of the correlations between the machine-generated tags and the hold-out folds. Additive regression outperformed the other models, even with a fairly small number of decision branches (10). However, it improved significantly when allowed to use additional decisions (400). From examining the decision stumps, these additional stumps allowed it to incorporate additional factors and also form piecewise curves for some of the strongest factors.

**Table 2: Regression Fits for (ES+EL2)/2 (10-fold CV)**

| | Linear | SVM | Add. (10) | Add. (400) |
|---|---|---|---|---|
| Human Tags | 0.24 | 0.55 | 0.62 | 0.69 |
| Machine Tags | 0.24 | 0.49 | 0.52 | 0.56 |

The linear model performed very badly, despite parameter selection: it tended to overfit the data and did not seem to model the expert ratings very well. SVM performed slightly better, but was not the best model overall. The Additive model, which was based on decision thresholds, worked best

**Figure 1: Model Data Flow**




out of the three. This may indicate that the human raters tended to implicitly use heuristics such as "too many Modeling modes," or "not enough Student contributions." The nature of features was also a factor, since many features were relatively sparse in each session (e.g., only occurred once or twice within an average session), which lends itself to rules related to the existence of a feature (i.e., N > 0).

Models trained on the machine-generated tags followed a similar pattern, but with slightly worse estimates. Retraining the classifiers on the machine-labeled tags did not significantly improve estimates based on those tags. When applying the model trained on human tags to the training set with machine tags, the model fit is R=0.54, as compared to R=0.56 for the cross-validated model built on the machine tags. As such, the machine tags appear to lose certain information, rather than simply categorizing it differently.

Since the smallest Additive Regression model worked so effectively, it is worthwhile to examine the features that were included. These models differed slightly when trained on human tags versus machine-labeled tags. The top features for this model on human tags vs. machine-labeled tags are shown in Table 5.2, in order of their importance (note: *Confirmation* is shortened to *Conf*). The presented analysis used non-standardized data, which is reasonable partly because the length of Tutor.com sessions tends to be fairly regular (i.e., a typical session is 15-25 minutes). Normalization would likely be needed to apply this to significantly different corpora. In general, many of the same patterns are important for both the human and machine tagged models. At least some of the judgments are based on a required minimal session length (e.g., # of Tutor Acts). Certain features appear to target evidence of learning (EL2), such as tutor actions that indicate the student has provided correct answers (*Confirmation:Positive*,*Expr:Praise*) and not passive in the tutoring session (*Expr:Neutral*, *Expr:LineCheck*). Other features appear to be associated with educational soundness (ES) for tutoring process (e.g., existence of a Closing, Scaffolding, and no excessive Modeling). Machine tagging appears to lose some of these nuances with respect to modes, probably due to the significantly lower accuracy for classifying modes.

Overall, the model appears to capture evidence of learning (EL2) better than educational soundness (ES). When trained on the full training data set (human tags), the Additive Regression (400) correlates with the average of ES and EL2 at R=0.8. By comparison, the correlation to these estimates is R=0.76 for EL2 versus R=0.63 for ES. Clearly, this is not the result of the outcome variable itself, which is a straight average of the two ratings (R=0.93 with EL2 and R=0.92 with ES). Instead, this indicates that the features for evidence of learning are more easily detected using the available taxonomy tags and features. This limitation was amplified when using the machine-generated tags, where the fit to (ES+EL2)/2 was R=0.54 but the correlation with the components was R=0.55 for ES2 and R=0.38 for ES. As such, improving the automated tagging of dialog modes would improve the automated assessments significantly.

**Table 3: Top-10 Features in Additive Regression**

| Trained on Human Tags | Trained on Machine Tags |
|---|---|
| Closing > 0 | # of Tutor Acts > 11 |
| *Expr:Conf:Positive [T]*⇒ *Expr:Conf: Positive [T]* > 0 | RapportBuild ⇒ Closing > 0 |
| Scaffolding > 0 | *Expr:Conf:Positive [T]*⇒ *Expr:Conf: Positive [T]* > 0 |
| Closing > 0 | *Assertion:Concept [T]* < 18 |
| *Expr:Apology [T]* = 0 | # of Tutor Acts < 12 |
| # of Tutor Acts > 6 | # of Tutor Acts > 5 |
| ProcessNegotiation⇒ Modeling ⇒ Modeling⇒ Modeling < 4 | *Request:Conf: Understanding [S]* < 3 |
| | |
| *Expr:Praise [T]* > 0 | Scaffolding⇒ Scaffolding⇒ Closing > 4 |
| *Expr:LineCheck [T]* = 0 | # of Tutor Acts < 12 |
| *Expr:Neutral [S]* > 15 | *Expr:Conf:Positive [S]* > 1 |

## 5.3 Tagging Large Tutoring Data Set

To examine the consistency of this assessment model on out of sample data, it was applied to a corpus of 242k machine-tagged sessions. The features for each tutoring session were extracted from parsing the transcript. Metadata about the session and the tutor were collected and aligned to the automated session assessments for analysis. The correlations between the Automated Estimates (Estimates), EL1, and PREREQ were available for almost the full corpus of 242k sessions. Other metadata was not always complete (e.g., not all tutor level data was available), so each pairwise correlation may have a slightly different N. However, all comparisons involve thousands of values and are statistically significant at the p<0.01 level.

**Table 4: Correlations of Quality Scores with Session Metadata**

| | Estimate | (ES+EL)/2 | EL1 | PREREQ |
|---|---|---|---|---|
| (ES+EL)/2 | 0.54 | - | - | - |
| EL1 | 0.45 | 0.56 | - | - |
| PREREQ | 0.39 | 0.49 | 0.87 | - |
| Tutor Level | 0.05 | 0.11 | -0.02 | -0.04 |

Table 5.3 shows the correlations between the automated estimate of session quality (Estimate), the average quality score for human raters (ES+EL2)/2 (available for the training set only), the original tutor's ratings for evidence of learning (EL1) and the learner's prerequisite knowledge (PREREQ), and the Tutor Level. The first two columns of this ta-

ble show that the estimate maintains similar correlations to those for the ratings that it was based on, across the larger data set, but slightly weaker overall. For example, the session tutor's rating of learning for the student correlates at R=0.56 (N=1438) for the training tags, but only R=0.45 (N=242k) for the automated tags across the full session data. With that said, the automated session rating maintains a similar pattern as the supervised tags across the full corpus. This indicates that the automated assessment captures significant information from the original expert raters, but with additional noise due to the machine-tagging process (particularly for modes).

This table also indicates why an external rating source can be important for evaluating the quality of tutoring sessions, even for well-trained professional tutors. Despite being rated independently by tutors with no knowledge of the original tutor, a higher Tutor Level correlated with significantly higher external quality ratings (R=0.11, N=1328). However, these more-expert tutors rated both the learning (R=-0.02) and the prerequisite knowledge (R=-0.04) lower than lower-level tutors. Or, put another way, less-expert tutors probably over-estimate both the learning and initial understanding of their students.

Moreover, it may be difficult for session tutors to provide ratings for the session that capture distinct features. For example, the original tutors expressed an R=0.87 (N=242k) correlation between learning (EL1) and and prerequisite knowledge (PREREQ). While one would expect these factors to be related, that level of correlation is nearly identical. By comparison, the external quality ratings correlated with the PREREQ assessments much more loosely (R=0.49, N=1438) and the automated assessments shadow this pattern (R=0.39, N=242k). So then, this automated rater provides a unique source of information modeled after the judgments of the external raters, which can be complementary to other sources of information about tutoring session quality.

# 6. CONCLUSIONS AND FUTURE WORK

This research has offered some insights into the five primary research questions posed earlier in Section 4. First, this work demonstrates the feasibility of an automated assessment model that models human expert judgments about the learning that took place during an online human-to-human tutoring session, at a level of R=0.54. While room for improvement exists, this model is already functionally useful. At least in this work, non-linear meta-models based on decision stumps (e.g., Additive Regression) outperformed more linear approaches such as Linear Regression and SVM Regression. This finding indicates that Random Forests [12] and similar algorithms are probably also promising for this type of problem. The strongest predictors of session quality in these models tended to be features where the tutor confirmed the accuracy of the student's responses, the session process indicated that progress was occurring (e.g., Scaffolding, Fading), or a consensus about successful learning was reached (i.e., a mutually-agreed Closing). Of these features, modes were fragile when machine tags were used: the level of noise in the mode classification appears to wash out information that is needed to evaluate the tutoring process. Finally, the resulting model was shown to follow similar patterns to the original training ratings, even over a much larger data set. This indicates that the automated assessments offer a reasonable proxy for expert human assessment when needed.

Notably, these ratings are calculated without a domain model that can directly assess the quality of students' answers. Instead, the model captures more general features of the tutoring interaction that relate to engagement and consensus between the tutor and student about learning accomplishment. As such, this model should be effective across a variety of tutoring domains beyond those analyzed in this work (Algebra and Physics). These session features are, in principle, domain-independent: they are based on classifications of tutoring dialog acts and modes.

However, this is also a limitation. Since the automated assessment system lacks the ability to assess the correctness of student input, it relies significantly on the session tutor's domain knowledge and basic capabilities to provide correctness feedback. As such, the session assessments can detect aspects of the pedagogy and student progress, but are unlikely to work appropriately if the tutors are entirely unqualified. This is, in part, because the training corpus includes only professional tutors who are rated and evaluated for quality. As such, additional quality-rated corpora might be needed to transition this estimator to other tutoring contexts where session quality assessments are important (e.g., peer-tutoring).

Additionally, significant drops in performance were observed when using machine-annotated sessions instead of human-annotated sessions. These drops were particularly severe for mode classifications, which had a direct impact on the ability of the session quality estimates to model the educational soundness of a session. This functionality would be helpful, as it allows credit for "good process" even when strong learning outcomes are not observed. Improving the accuracy of dialog mode classification would significantly strengthen the assessment of tutoring sessions, and is an important area for further research. One way to approach this problem would be to use active learning where machine-annotated transcripts are corrected by human taggers.

Finally, an important next direction for this research would be to train a similar tutoring session assessment model based on pre-test and post-test assessments, such as the approach taken by Boyer et al. [1]. This step would enable a comparison between the features underlying our expert ratings of session quality against the features associated with measured learning gains. This work may show notable qualitative differences related to not only the key features, but also the algorithms involved (e.g., discontinuous algorithms such as Additive Regression might not be as dominant). Features associated with learning gains that are not associated with human ratings might also help detect illusions of mastery or expert blind spots. Likewise, integrating both approaches for analysis of tutoring sessions would offer the potential to identify authentic "Eureka moments" where the learner's sense of sudden understanding can be shown to correlate with subsequent performance on a similar problem. In the long term, the process of maintaining and improving this model should provide insights into new features of successful tutoring that may even be more valuable than the automated assessments calculated by the model.

## 7. REFERENCES


[1] K. E. Boyer, R. Phillips, A. Ingram, E. Y. Ha, M. Wallis, M. Vouk, and J. Leste. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of AI in Education*, 21(1-2):65–81, Jan. 2011.

[2] S. D'Mello and A. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[3] S. D'Mello, A. Olney, and N. Person. Mining collaborative patterns in tutorial dialogues, Dec. 2010.

[4] J. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

[5] A. Gabadinho, G. Ritschard, N. S. Muller, and M. Studer. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37, 2011.

[6] A. Graesser, S. D'Mello, and W. Cade. Instruction based on tutoring. In *Handbook of Research on Learning*, pages 408–426. 2011.

[7] A. Graesser and N. K. Person. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6):495—-522, 1995.

[8] A. C. Graesser and N. K. Person. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137, Jan. 1994.

[9] M. Hall, E. Frank, and G. Holmes. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[10] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18—-28, 1998.

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.

[12] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[13] A. Meier, H. Spada, and N. Rummel. A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, 2(1):63–86, Feb. 2007.

[14] C. Moldovan, V. Rus, and A. Graesser. Automated speech act classification for online chat. In *Midwest Artificial Intelligence and Cognitive Science Conference*, pages 23–29, 2011.

[15] D. Morrison, B. D. Nye, and V. Rus. Tutorial dialogue modes in a large corpus of online tutoring transcripts. In *Artificial Intelligence in Education (AIED) 2015*, Under review.

[16] D. Morrison and V. Rus. Defining the nature of human pedagogical interaction. In R. A. Sottilare, X. Hu, H. Holden, and K. Brawner, editors, *Generalized Intelligent Framework for Tutoring Systems, Volume 2: Pedagogical Strategies*, pages 217–224. 2014.

[17] I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger. Metacognitive practice makes perfect: Improving students' self-assessment skills with an intelligent tutoring system. In A. Biswas, G and Bull, S and Kay, J and Mitrovic, editor, *AIED 2011*, volume 6738 of *LNAI*, pages 288–295, 2011.

[18] C. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271, Jan. 2008.

[19] V. Rus and N. Niraula. Automated labeling of dialogue modes in tutorial dialogues,. Technical report, The Language and Information Processing Lab, The University of Memphis, 2014.

[20] B. Samei, V. Rus, B. D. Nye, and D. M. Morrison. Hierarchical dialogue act classification in online tutoring sessions. In *Educational Data Mining (EDM) 2015*, In Press.

[21] J. Searle, F. Kiefer, and M. Bierwisch. *Speech act theory and pragmatics*. 1980.