Exploring real-time student models based on natural-language tutoring sessions

A look at the relative importance of predictors

Benjamin D. Nye*
The University of Memphis
Memphis, TN 38152
bdnye@memphis.edu

Mustafa Hajeer
The University of Memphis
Memphis, TN 38152
mhhajeer@memphis.edu

Carol M. Forsyth
The University of Memphis
Memphis, TN 38152
cmfrsyth@memphis.edu

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

Keith Millis
Northern Illinois University
DeKalb, IL 60115
kmillis@niu.edu

Xiangen Hu
The University of Memphis
Memphis, TN 38152
xhu@memphis.edu

ABSTRACT

Natural language tutoring systems generate significant data during their tutoring sessions, which is often not used to inform real-time, persistent student models. The current research explores the feasibility of mapping concept-focused tutoring sessions to knowledge components, by breaking sessions down into features that are integrated into a session score. Three classes of tutoring conversation features were studied: semantic match of student contributions to domain content, tutor support (e.g., hints and prompts), and student verbosity (i.e., word counts). Analysis of the relative importance of these features and the ability of these features to predict later task performance on similar topics was conducted. Reinforcing prior work, semantic match scores were a key predictor for later test performance. Tutor help features (hints, prompts) were also useful secondary predictors. Unlike some related work, verbosity was a key predictor even after accounting for the semantic match.

Keywords

Student Modeling, Natural Language Processing, Educational Data Mining, Trialogs

1. INTRODUCTION

Adaptive, natural-language intelligent tutoring systems (ITS) have been a research goal since the early days of AI. The current research focuses on natural language conversation features from tutoring sessions instead of more explicit methods such as multiple choice questions. In this work, we associate each natural language tutoring conversation with a knowledge component, then score the session. This approach can be used in a pure natural language tutoring system to build a persistent, component-based student model that resembles those more commonly found in problem-based tutors. Effectively, tutoring conversations can provide a quasi "stealth assessment" [12] where the intelligent tutor assesses performance without breaking out from the tutoring session.

This study looks at the relative importance of tutoring session features and models a reasonable set of regression weights that could be used to inform an efficient, useful, and interpretable real-time student model. Focus was placed on preparing this model to transfer across domains. This is because the model is ultimately intended for a new tutoring system focusing on natural language tutoring for Algebra I, but it is being trained on data for a system focusing on research methodology. Additionally, a long-term goal of this model is to use it inside a generic persistent student model for AutoTutor-style tutoring [4], which is one major framework for conversational tutoring.

*Corresponding Author

2. BACKGROUND AND RELATED WORK

Forbes-Riley and Litman [2] compared multivariate regression models to determine the relative usefulness of hundreds of features from sessions of the ITSPOKE system, a natural language ITS for physics. A variety of features were evaluated, which were categorized as either shallow features (e.g., student verbosity), semantic features (e.g., concept keywords/# of words), pragmatic features specific to ITSPOKE (e.g., number of goals to retry), discourse structure specific to ITSPOKE (e.g., depth), or local context for dialog acts (e.g., bigram speech act pairs). On holdout test data sets, models with semantic features consistently outperformed any model without these features, with R^2 values ranging from 0.338 to 0.524 depending on the data sets used. So then, semantic features appear to be the most pivotal. Pragmatic features, discourse structure, and local context all showed some evidence of usefulness for training,

and on some test conditions, but none had the consistency of semantic features. Shallow features were not effective and were dropped when fitting regressions using all features. Follow-up research based on corpora from both ITSPOKE and BEETLE II, an ITS for electronic circuit analysis, also found meaningful correlations with posttest scores based on the number of semantically-relevant patterns (e.g., stemmed keywords) expressed by the student [9].

While the work with ITSPOKE indicated a potential for overfitting when basing models on tutoring behavior (e.g., ITSPOKE pragmatics) [2], later research on problem-based ITS indicates that pragmatic features may still be valuable. A study on continuous Bayesian knowledge tracing found that calculating partial credit for each problem based on the number of hints and bottom-out answers improved estimates of student performance, even though penalty weights for hints and other support were ad-hoc [13]. This research noted that greater numbers of hints correlated negatively with later performance. Since high-performing students are less likely to need hints, this is somewhat intuitive.

Overall, this review of the literature found that semantic features are essential predictive features and support (e.g., hints) may also be valuable. Verbosity during a session (e.g., average words per statement) was also considered as possibly useful, based on earlier analysis of AutoTutor data. Prior to the analysis, it was anticipated that higher semantic match scores would correlate positively with later performance, while support would correlate negatively with performance. Higher verbosity was expected to be associated with better performance, but was not necessarily expected to add value beyond the semantic match.

3. DATA SAMPLE AND FEATURES

This study analyzed a corpus of data collected from the Operation ARA (Acquiring Research Acumen), which is Pearson Education's commercial version of the desktop tutor Operation ARIES (Acquiring Research Investigative and Evaluative Skills) [10]. Both ARA and ARIES use natural language conversations based on the AutoTutor conversational framework. Operation ARA tutors research methodology using multiple methods, including three-way conversations between a human student and two or more artificial agents (e.g., trialogs). Operation ARA has three phrases (Training, Proving Ground, and Active Duty) and 11 chapters. These phases were preceded by a pretest (2 per chapter, 22 items total) and followed by a posttest (also 2 per chapter, 22 items total). Each chapter focuses on a particular concept related to research methodology (e.g., correlation vs. causation). Only sessions from the Training phase were used as inputs, since these dialogs are most representative of generative natural language tutoring (i.e., where the student attempts to explain concepts).

A set of 192 students across 11 classrooms and 9 instructors was analyzed in the current study. These subjects were from a pool of 462 students from 12 class sections in an undergraduate Psychology course at Northern Illinois University. Unfortunately, many students were excluded due to lack of required consent forms (190), missing pre/post tests (42), or unreasonably fast response times on pre/post tests (38). The latter group answered test questions in under 5 seconds, corresponding to a reading speed of over 10 words/second, far faster than is reasonable to read and methodically select an answer.

Pretest results were considered as a predictive feature, to determine the relative effectiveness of the tutoring conversation features against a traditional assessment. The final posttest results were treated as performance outcomes. Two chapters were excluded from the reported analysis because post-test results correlated poorly with pre-test results, raising some uncertainty over item equivalence. This may just be due to by-topic differences, as observed in an earlier data set collected with ARIES [3]. A follow-up analysis including these chapters found no overall change to model fit or conclusions: one topic raised fit, the other reduced it slightly.

Four features were extracted from each session: the average semantic match score during a session (i.e., average quality of student responses), the verbosity, and two "help" features (the number of hints and of prompts). For human statements answering a tutor or computer student question, the semantic match score and verbosity (number of words) were extracted. AutoTutor calculates and logs these semantic match scores as the session unfolds, using Latent Semantic Analysis [7, 4]. The average semantic match (S) and average verbosity (V) for student contributions was calculated. Statements were only included in this average if an open-ended response would be expected. As a substitute for average verbosity, a logarithm for verbosity was also calculated ($ln(1+V)$). This transform captures the inherent nonlinearity of verbosity: a single word is qualitatively different than a blank response, but a single word onto a 100-word statement is seldom important.

Computer agents' statements were classified by their type, such as a hint or prompt. For each session, multiple hints (H) and multiple prompts (P) may exist. While the semantic match and verbosity of student statements partially influence the ITS to produce these speech acts, the specific rule sets and depth of content for tutoring also influences these values. Effectively, these capture the interaction of student input with the author's expert model for when feedback was needed.

Finally, while not a dialog feature, a student model needs to establish the relative importance of more recent dialog sessions as compared to earlier sessions. Since tutoring increases the student's understanding, knowledge levels are inherently non-stationary and have a certain degree of (hopefully positive) drift. The magnitude of this drift will determine the optimal update rate (λ_u) for weighting more recent sessions. For this study, an exponential moving average was applied (e.g., $\overline{X_t} = \lambda * x_t + (1 - \lambda) * \overline{X_{t-1}}$, where $\overline{X_1} = x_1$) [1]. In a simulated exploration, update rates between 0.4 to 0.8 had higher average model fits, with $\lambda_u = 0.5$ being representative. While this update rate is not claimed to be optimal, it was a reasonable starting point. This exponential moving average smoothed and summed each feature into a single feature score for a given concept (i.e., chapter).

4. DATA MINING AND MODELING

The analysis presented here had main goals: 1) Determine which tutoring discourse features contribute unique predic-

tive value and 2) Determine the relative importance of these factors for predicting later test performance. Across the 192 students and 9 chapters analyzed, 1067 student-chapter combinations had at least one tutoring session during the Training phase and were used as the sample. The majority of pairs had only one ((656) or two (333) sessions, with a handful having three (75) or four (3). To note, since lower-performing students received more tutoring sessions, this data may over-represent lower-performing students. With that said, the number of tutoring sessions was not correlated with posttest scores for each chapter.

First, correlations were calculated between the posttest and time-averaged predictors. Next, three regression models were fitted to predict posttest outcomes: pretest only, the full set of tutoring session features, and all features plus the pretest. These were performed using 10-fold cross validation and on the full data set using Weka [6]. Then, the LMG (Lindeman, Merenda, & Gold) method for relative importance of linear regressors [8] was applied, as implemented in the R *relimpo* package [5]. LMG calculates the average R^2 contributed by each factor (e.g., the variance explained) across all orderings and combinations of regressors. Relative importance regressions often produce regression weights that generalize to new data. A second set of relative importance Pratt regression coefficients was also calculated [11]. Pratt weights are standardized coefficients (i.e., Beta weights) multiplied by the correlation between the predictor and outcome (i.e., $\beta * r_{i,out}$).

5. RESULTS

The correlations between factors followed the expected patterns. Table 5 shows Pearson correlations between the posttest results and predictors. The pretest (Pre), postest (Post), semantic match (S), and verbosity ($ln(1+V)$) all have highly-significant, positive correlations with each other ranging from weak to moderate. Hints (H) and prompts (P) correlate positively with each other, but negatively with the other variables. The logarithm of verbosity had much higher correlations with other variables than raw verbosity. For example, raw verbosity correlated 0.05 (p=0.08) with the posttest, compared to 0.17 (p<0.001) for the logarithmic transform.

Table 1: Posttest and Predictor Correlations

	Post	Pre	S	H	P	ln(1+V)
Post		0.14[c]	0.14[c]	-0.04	-0.08[a]	0.17[c]
Pre	0.14[c]		0.06[a]	-0.04	-0.06[a]	0.08[b]
S	0.14[c]	0.06[a]		-0.62[c]	-0.69[c]	0.55[c]
H	-0.04	-0.04	-0.62[c]		0.58[c]	-0.31[c]
P	-0.08[a]	-0.06[a]	-0.69[c]	0.58[c]		-0.43[c]
ln(1+V)	0.17[c]	0.08[b]	0.55[c]	-0.31[c]	-0.43[c]	-

[a]p<0.05; [b]p<0.01; [c]p<0.001

5.1 Linear Weights

All features (pretest, semantic score, hints, prompts, and verbosity) improved the linear model during 10-fold cross validation and on the full data set. Three key models are shown in Table 2: pretest only, all tutoring features, and the combined set of predictors (pretest and tutoring). The low variance explained by the pretest demonstrates the noise in the data, since pretest values often account for the majority of the explained variance [2]. In this data set, dialog features are significantly more predictive than the pretest alone.

Table 2: Variance Explained by Linear Regressions

Predictor Set	R^2 (Training)	R^2 (Cross Val.)
Pretest Only	0.020	0.017
Tutoring Only	0.037	0.028
Combined	0.054	0.043

While the overall variance explained is modest, very limited data was available for each combination of student and chapter. Most pairs contain only a single session and the average session had 2.4 student contributions. To look at the added value for additional sessions, the data was split into two subsets: single-session ($N_S = 1$) and multiple-sessions on a chapter ($N_S > 1$). Table 3 shows the model fits for this split. Discourse features were more predictive when multiple sessions were available. The combined model with two or more tutoring sessions outperforms any other model, and accounts for 8% of training variance and 5.7% for cross-validation. 81% of the $N_S > 1$ subset have two sessions, so even adding one additional session captures 1.4% to 2.6% of the remaining component variance.

Table 3: Impact of the Number of Sessions on Variance Explained

Predictor Set	R^2 (Training)	R^2 (Cross Val.)
Pretest Only ($N_S = 1$)	0.027	0.022
Pretest Only ($N_S > 1$)	0.012	0.005
Tutoring Only($N_S = 1$)	0.028	0.018
Tutoring Only($N_S > 1$)	0.072	0.053
Combined ($N_S = 1$)	0.052	0.038
Combined ($N_S > 1$)	0.080	0.057

The remainder of the analysis focuses on the tutoring features only. While pretests have predictive value, they are content-specific and are unlikely to be transferable to a new domain. Table 4 shows three sets of relative importance weights for each feature: LMG (contribution to R^2), Pratt (standardized, meaningfully-signed coefficients), and a set of interpretable unstandardized weights generated by transforming the Pratt weights. LMG and Pratt weights were similar in relative magnitude, with the exception that the Pratt weights are signed. Verbosity and semantic match scores dominate in both cases, with the influence of hints and prompts almost an order of magnitude lower. With that said, hints and prompts are still significant predictors and improve the model fit.

Table 4: Relative Importance Weights

Predictor	LMG	Pratt	Interpretable
Semantic	0.0124	0.0183	1.00
Hints	0.0019	-0.0024	-0.078
Prompts	0.0022	-0.0026	-0.013
ln(1+V)	0.0208	0.0242	0.28
R^2	0.037	0.037	0.031

Interpretable weights were generated in a two-step process. First, each Pratt weight was divided by the sample standard

deviation for that variable. Second, all of these weights were divided by the semantic match score weight. These resulting weights retain the majority of the predictive value on the training data, so long as predictions are clipped to fit in [0,1]. Rescaling the weights until the intercept is zero tended to offer a higher fit than other scalings. This occurred when the semantic match coefficient was close to 1, as displayed in the above weights (1.045 for the 9-chapter set and almost exactly 1 when the an additional chapter was considered). As such, it appears that the semantic match score acts like a de-facto intercept value. This model was used to predict the sum of student posttest scores across their included chapters (i.e., any chapter with a tutoring session). The Pearson correlation between the sum of each student's posttest scores and the sum of predicted knowledge levels was fairly strong (R^2=0.388, p<0.001, N=192).

6. CONCLUSIONS AND FUTURE WORK

The logarithm of verbosity and the semantic match score were the primary predictors, performing better than even the pretest items. The high importance of verbosity was somewhat surprising, given prior work which found little value for surface features [2]. The difference may have been caused by this study focusing on the average verbosity on a particular concept, rather than overall word counts. Additionally, the logarithmic transform improved verbosity from a fairly weak correlate to a powerful one. Support such as hints and prompts was also predictive, and negatively related to later performance on the posttest. This weaker importance is probably caused by causation from poor semantic match (poor answers make hints more likely) and learning due to hints (learning from hints offsets worse knowledge).

The model explained significant variance in the overall posttest score (R^2=0.39), but modest variance for each component. Given that most components had only one short tutoring session (about 2.4 student contributions) to predict a pair of posttest multiple choice questions, this is fairly promising. Using only a handful of student utterances, this model outperformed balanced pretest items for predicting posttest component performance. Since even a single additional session significantly increased the variance explained, more sessions per concept should improve predictions. Additionally, the Proving Ground and Active Duty phases add noise between the Training phase and posttest.

With that said, these results are drawn from tutoring dialogs on a single domain with a fairly small number of topics. Follow-up research will test the model on a new domain (Algebra I), with larger numbers of tutoring sessions per concept. This evaluation will occur during the next year, and should provide useful information about the transferability of this tutoring session scoring model to a new domain. Future studies will focus on the effectiveness and limitations of a student model for classifying student performance, once pilot and evaluation data have been collected.

7. ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research STEM ITS Grand Challenge, the Institute for Education Sciences (R305B070349), and the National Science Foundation (HCC 0834847). The authors alone are responsible for statements in this paper.

8. REFERENCES

[1] R. G. Brown. *Smoothing, forecasting and prediction of discrete time series.* Dover Publications., Mineola, New York, 2004.

[2] K. Forbes-Riley, D. Litman, A. Purandare, M. Rotaru, and J. Tetreault. Comparing linguistic features for modeling learning in computer tutoring. pages 270–277, June 2007.

[3] C. Forsyth, P. J. Pavlik, A. C. Graesser, Z. Cai, M.-l. Germany, K. Millis, R. P. Dolan, H. Butler, and D. Halpern. Learning gains for core concepts in a serious game on scientific reasoning. In *Educational Data Mining (EDM) 2012*, pages 192–195. International Educational Data Mining Society., June 2012.

[4] A. Graesser, P. Chipman, B. Haynes, and A. Olney. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618, Nov. 2005.

[5] U. Grömping. Relative importance for linear regression in r: the package relaimpo. *Journal of Statistical Software*, 17(1):1–27, 2006.

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

[7] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[8] R. Lindeman, P. Merenda, and R. Gold. *Introduction to bivariate and multivariate analysis.* Scott Foresman, Glenview, IL, 1980.

[9] D. Litman, J. Moore, M. O. Dzikovska, and E. Farrow. Using natural language processing to analyze tutorial dialogue corpora across domains modalities. In *AIED 2009*, pages 149–156, Amsterdam, The Netherlands, 2009. IOS Press.

[10] K. Millis, C. Forsyth, H. A. Butler, P. Wallace, A. C. Graesser, and D. F. Halpern. Operation ARIES! a serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, and J. Lakhmi, editors, *Serious Games and Edutainment Applications*, pages 169–195. Springer, London, UK, 2011.

[11] J. Pratt. Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics.*, pages 245–260, Tampere, Finland, 1987. University of Tampere.

[12] V. Shute. Stealth assessment in computer-based games to support learning. In S. Tobias and J. D. Fletcher, editors, *Computer games and instruction*, pages 503–524. 2011.

[13] Y. Wang and N. Heffernan. Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Artificial Intelligence in Education (AIED) 2013*, pages 181–188, Berlin, Germany, 2013. Springer.