Predicting Student Learning from Conversational Cues

David Adamson[1], Akash Bharadwaj[2], Ashudeep Singh[3], Colin Ashe[4], David Yaron[1], and Carolyn P. Rosé[1]

[1] Carnegie Mellon University, USA
[2] National Institute of Technology Karnataka, India
[3] Indian Institute of Technology Kanpur, India
[4] Indiana University of Pennsylvania, USA

Abstract. In the work here presented, we apply textual and sequential methods to assess the outcomes of an unconstrained multiparty dialogue. In the context of chat transcripts from a collaborative learning scenario, we demonstrate that while low-level textual features can indeed predict student success, models derived from sequential discourse act labels are also predictive, both on their own and as a supplement to textual feature sets. Further, we find that evidence from the initial stages of a collaborative activity is just as effective as using the whole.

Keywords: Computer-Supported Collaborative Learning, Discourse Analysis, Educational Data Mining.

1 Introduction

Intelligent tutoring and computer-supported collaborative learning can both provide cognitive, metacognitive, and social benefits to learners [13, 22, 27]. These systems also offer a wealth of process data to researchers and developers. This windfall can be used to analyze learning and other behavioral processes, and opens the door to automatic moment-to-moment formative assessment and support. The recent boom in massive and open online courses, with their similarly massive student-to-human-teacher ratios, has underlined both the need and the potential for such data-driven assessments and interventions. In this paper, we present multiple sources of predictive features from the chat transcripts of a collaborative learning scenario. As a baseline, we show that features based on the lexical and syntactic contents of student contributions in chat are predictive. We then supplement those features by paying attention to the sequence and structure of dialogue at the discourse level, and demonstrate that these features can anticipate student learning.

The remainder of this paper is organized as follows: In Section 2, we review relevant literature and establish a theoretical framework for our contribution. In Section 3, we describe the collaborative learning context which we analyze according to the methods presented in Section 4. We present our results in Section 5, and offer some in-depth interpretation. We end with a look forward, to future applications and extensions of this work.

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 220–229, 2014.

2 Background

This paper grounds itself in the fields of Educational Data Mining and Computer Supported Collaborative Learning. In particular, we build upon prior work that has successfully employed a variety of methods for feature extraction and pattern learning to predict affective, collaborative, and learning outcomes from discourse.

Linguistic analysis methods for studying both individual learners and small groups [11] have been be used to assess cognitive and meta cognitive knowledge [10], critical thinking, knowledge construction [9] and consensus building techniques [16]. In many cases [5, 26], methods for automatically labeling these features are developed hand-in-hand with their application to a prediction task. Analysis applied to course message boards has shown it is possible to detect unresolved questions [12] in asynchronous discussions, and that patterns of interaction and participation can be used to predict final learning outcomes [21]. In the context of a single-user conversational tutor, a set of conversational features, including measures of the quality and content of student answers as derived from Latent Semantic Analysis [15], have been successfully applied to predict the moment-to-moment affect of the learner [5].

In intelligent tutoring systems with a conversational component, automated analysis methods may be employed as formative assessments, predicting student learning or collaborative performance. These predictions can be used to inform a tutor's interventions during future learning experiences, or to provide moment-by-moment facilitation in response to continuous assessment [1]. Recent work has demonstrated the power of data mining for building moment-to-moment models of student learning [2], although as this work was situated in a non-conversational tutoring system, it did not leverage linguistic features to anticipate learning. Fully automated coding and modeling methods have been used to successfully predict the outcome of a facilitated civil-dispute negotiation [26]. Models of conversational trajectory have also been developed as a source of feedback for learners and their human instructors, using a set of features describing conversational attributes derived from per-turn coding of a conversation [3, 4]. In that work, each coded move contributes to one of four underlying conversational dimensions (conformity, creativity, elaboration, and initiative), allowing concrete quantitative measures to power a qualitative analysis of group state.

Hidden Markov Models [20] trained on sequences of student-selected sentence-opener moves have been used to classify and describe groups of collaborative learners as more or less productive [24, 25]. HMMs have also been applied to surveys of participant emotion, to draw inferences about underlying affective or cognitive state [6]. However, such work has relied on participants selecting their next move or observed state from a limited set of options. More recent work has used n-grams or stretchy patterns [8] over discourse act labels to model local conversational structure and predict group task success [19]. Although this body of work illustrates the potential of sequential models for understanding student state, their suitability as a method for assessing individuals within an unconstrained multiparty discourse has not been fully explored.

3 Context and Corpus: College Chemistry Collaboration

We conduct our study data collected from a small-group chat-based collaborative task in the domain of college chemistry. The participants in this study were first-year undergraduate students in an introductory chemistry course, during a unit on intermolecular interactions. Students were randomly assigned to groups of three or four. Participation in the exercise was voluntary, and students had the option of not consenting for their data to be included in our research. Altogether, our analysis includes data from 50 consenting students from 16 different groups - with a mean of 93 messages per student, or 292 per group. Students were administered a pre-test the day before they completed the task, and completed a post-test the day after. Two test forms were randomly counter-balanced by student between pre- and post-test.

This task and chat environment have been used before to study methods for automatic discussion facilitation [1]. The 90-minute task focuses on intermolecular forces and their influence on the boiling points of liquids. The task was framed as a collaborative data analysis activity, where the students in each group were assigned to read individually about one of three classes of molecules, and the factors most likely to influence their boiling point. This division also provided intrinsic motivation for collaboration, as the task could not be completed without knowledge from each of the student experts. A conversational agent [14] facilitated the activity for each group, presenting the series of exercises to the group and prompting them to explain their reasoning to each other.

4 Methods: Predicting Learning from Conversation

We aim to capture the properties of conversation that are distinctive of more (or less) successful learners. Low-level lexical and syntactic features are examined alongside higher-order representations of discourse, and evaluated as candidates for automating future formative assessment. In order to assess individual learning, we first build a linear model, predicting student post-test score from pre-test score alone. This model accounts for 61% of variance in student performance. The impact of collaboration, if any, might be found in the remaining unaccounted-for variance. Thus, we use as our target the residual from this regression in the remainder of the analysis.

4.1 Baseline Textual Features

Especially in unstructured conversational data, the success of a machine learning algorithm is tied to the feature representation of the contents of that data. We first use **"bag-of-words"** features, which represents only the vocabulary used in a conversation (including both content words and function words). We then present a second model, based on **"complex language"** features. This model contains a superset of the bag-of-words feature set. Adjacent pairs of words (bigrams) and local syntactic part-of-speech bigrams are added as features.

In addition to a single student's language (the **Student Only** condition above), much of her learning may be tied up in her interactions with peers. We therefore introduce an additional text representation (**Whole Group**), including features for all students in each conversation as a second feature set. These new features are represented as distinct - thus any unigram may appear twice in an instance as a distinct feature, once if spoken by the student of interest, and again if spoken by any of her groupmates.

Finally, in order to evaluate our methods' suitability for mid-activity formative assessment, we also test the condition where only features from the first third of each student transcript are used for prediction (**Start Only**), stopping at the end of the first phase of the activity described in Section 3.

We train a Naive Bayes classifier to differentiate groups with a positive residual (learning more than the pre-test would suggest) from those with a negative residual. To avoid overfitting (identifying the peculiarities of individual groups, rather than overall trends in student behavior), results of our machine-learning experiments are presented from 16-fold leave-one-group-out cross-validation. In this arrangement, models are trained on 15 groups of 46 or 47 students, and tested on the remaining group of 3 or 4 students. Reported performance is averaged across groups. The model is limited to using the top 100 most predictive language features on each training fold, using χ^2 feature selection [7].

4.2 Active Learning Annotation

To represent features above the contributions of individual lines of dialogue, we refer to established frameworks for conversational analysis. In Barros et al.'s work, a set of attributes for qualitative conversational analysis is proposed [4] based on a set of six sentence-opening moves. This is similar to the scheme used by Soller [23]. We combine Barros' two types of proposal and consider just five types of "Active Learning" moves:

- Proposals (**PR**) begin a sequence and introduce a new concept or idea.
- Questions (**QU**) target proposals and question them.
- Clarifications (**CL**) are elaborations on proposals, or answers to questions.
- Agreements (**AG**) show agreement or assent between speakers in a sequence.
- Remaining contributions are Comments (**CM**); including topic statements, floor grabbing moves, pauses, etc.

In earlier works, assignment of turn labels relied on student inputs being constrained to a fixed set of sentence-openers. In our approach, the students are not thus fettered, and we instead rely on annotation of free text. To allow this flexibility, we adapted a coding manual based on the systemic functional linguistics "Negotiation" framework [17], describing the flow of information and action within a conversation. Recent work has shown that Negotiation annotation can be automated for freeform chatroom conversations [18]. With an eye toward such future automation, we adapted Mayfield's coding manual, converting Negotiation labels to Active Learning moves using heuristics. This manual was first validated on separate pilot data. For this data, three transcripts (about 2000 turns of conversation) were coded by both annotators to check reliability, and the rest were each coded by a single annotator. Resulting reliability was high for Active Learning annotations, $\kappa = 0.75$.

From these annotations, we can now represent sequences of labeled turns as inputs to our machine learning algorithms. As a starting point (**Active Learning Trigrams**), we use sequences of three consecutive labels, extracted from the sequence of labeled turns, as a feature for our group and student tasks. In the case of per-student outcome prediction, each tag is differentiated based on who (relative to the student in question) is speaking - either the student herself, or another participant. For example, $\mathbf{PR}_s$ is a proposal issued by this student, $\mathbf{CL}_o$ is a clarification by another student, and so on. We consider this representation both on its own and as a supplement to our textual features.

As in Section 4.1, we train a Naive Bayes classifier with these features and report results from 16-fold cross-validation. As an additional experiment, we also evaluate a single classifier trained on the combined feature set of **Active Learning Trigrams** and **“complex language”** features.

4.3 Predicting Learning with Contrastive Hidden Markov Models

As a more sophisticated differentiator of conversational structure, we use Hidden Markov Models [20] to model variation between successful and unsuccessful students. HMMs are a sequential labeling algorithm, where observed behaviors are assumed to be a result of an unobserved, hidden state. In this case, states may correspond to a student’s intention when contributing a new turn to the dialogue. By analyzing sequences of observed labels, HMMs can discover these unobserved states statistically.

Following Soller et al. [24], we train two HMMs with four hidden states, on sequences drawn from subsets of the corpus - one using the sequences from the four students with the highest residuals, the other using the four students with the lowest residuals. The resulting models should distinguish the sequential behaviors of unusually high- and low-performing students. We make no presumptions about the meanings of specific hidden states [6], although we expect to see meaningful patterns relevant to collaborative discourse.

As with our textual experiments, we use leave-one-group-out cross-validation, so no student transcript is evaluated on a model trained on a member of that transcript. For each held-out student in the test group, we calculate the normalized sequence likelihood of their entire transcript for each model, and use the likelihoods that the two models assign to the held-out data as features for a linear model performing binary classification. To mirror the **Start Only** conditions above, we also apply the same procedure to only the first third of the Active Learning sequences in each transcript, to assess this method’s suitability for in-process formative assessment.

5 Results and Discussion

Results for the classification experiments using textual features are presented in Table 1. In general, we find that richer text features and including context from group members' posts both contribute to performance well above an individual student's vocabulary alone, and their benefit is somewhat additive. Further, in the more complex model we find that using only features from the starting section of each transcript perform statistically indistinguishably from models built on the entire transcript, suggesting that such methods may enable mid-activity formative assessments based on conversational features.

Table 1. Predicting individual learning above or below expected levels with textual features alone, based on raw accuracy (%) and Cohen's kappa. **Bold** represents a marginal improvement over baseline accuracy, $p < 0.1$.

	Student Only		**Whole Group**		**Start Only**	
Feature Set	%	κ	%	κ	%	κ
Bag-of-words	0.58	0.14	0.64	0.25	0.49	-0.01
Complex language	0.64	.025	**0.70**	**0.38**	0.68	0.38

In Table 2, we see the impact of Active Learning sequential features. Active Learning trigrams appear to offer additive benefit alongside textual features, improving our ability to predict student over- or underperformance. Using the more sophisticated contrastive HMM model, we are able to replicate this performance by only modeling states based on sequences of Active Learning tags. Table 3 lists a few features from this combination model that are highly predictive of high and low residual scores.

Table 2. Predicting individual learning above or below expected levels with sequential dialogue features. **Bold** represents marginal improvement over baseline, $p < 0.1$.

Sequence Representation	%	κ
Active Learning Trigrams	0.66	0.30
Trigrams + Textual Features	**0.72**	**0.43**
Contrastive HMMs	**0.72**	**0.44**
Contrastive HMMs (Start Only)	0.64	0.28

5.1 Qualitative Analysis of Contrastive HMMs

The output of the contrasting HMMs can be used to gain insight into the conversational habits of more (or less) successful students. Figure 1 illustrates the difference in transition patterns between student with higher and lower residual scores. Note that although the learned states were not predetermined, fairly consistent groupings emerge between models. In the model for higher scores, we see

Table 3. Representative features for high or low residual scores

Feature	Actor(s)	Class	Example
thinking	student	high	i'm thinking in between
ADV WH	other	high	so why not higher?
$\mathbf{CL}_o\mathbf{PR}_s\mathbf{QU}_s$	both	high	yep - the dipole moment is what's different chcl3 second highest, ch3cl third highest the last one has no dipole moment then?
agree ?	student	low	KCl will be in the middle ... agree?
ADJ CONJ	other	low	smaller or bigger?
$\mathbf{CL}_o\mathbf{QU}_o\mathbf{PR}_o$	other	low	i think the bp increases as we go down the table does all 3 increase down the table? i think the dipole moment is more important

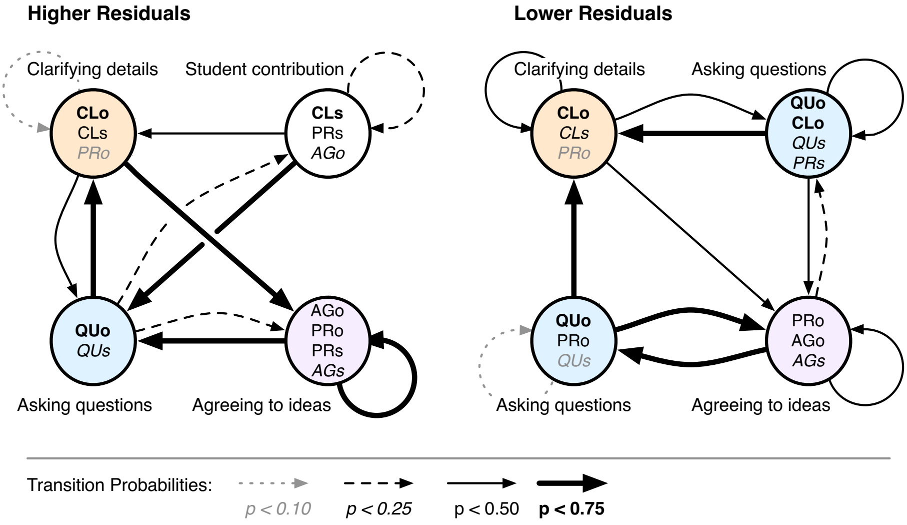

Fig. 1. Learned High and Low State Transitions

a strong flow between states that have high emission probabilities for questions and clarifying statements, and from clarification to agreement to proposals. In particular, the high-residual model favors transitions from questioning, to clarification, to agreement and new ideas, whereas there's a comparatively weak flow out of the clarification state in the low-residual model. The low-residual model also displays stronger tendencies toward loops in the clarification and questioning states. It may be that students who fit the lower-residual model find themselves in groups experiencing more confusion, but with less productive resolution. The low-residual model expects a lesser degree of student participation (as indicated by lower emission probabilities for student moves, versus moves by others). A hard-to-reach state focusing on student contributions is unique to the high-residual model, which favors reentry into the question-clarify-agree loop.

Table 4. Highly likely sequences, according to the HMMs for high and low residual scores (top and bottom). Note that comments are not included in the model.

Tag	Text
$\mathbf{PR}_o$	yea they are made up of the same molecules so i cant really tell yet
$\mathbf{QU}_s$	It's going to be in the middle right?
$\mathbf{CL}_o$	its going to be the smallest because the dipole moment is the smallest
$\mathbf{QU}_o$	so its actually smallest?
$\mathbf{CL}_s$	wait just kidding i read that wrong! Smallest.
$\mathbf{AG}_o$	ya smaller dipole=smaller boil pt

$\mathbf{PR_o}$	Polar molecules have a permanent dipole moment which is caused by differences in electronegativity between bonded atoms. One might have more electronegativity than the other causing a nonuniform electron distribution.
$\mathbf{CL_s}$	In my intro, it said dipole moments do not at all affect the boiling point
$\mathbf{PR_o}$	The table shows you it does though
$\mathbf{AG_s}$	yeah this one shows that it does
$\mathbf{CM_s}$	which is weird
$\mathbf{PR_o}$	They look like nonpolar molecules

Some highly probable sub-sequences according to each model are illustrated in Table 4, with examples from the corpus.

6 Conclusions and Future Work

The experiments presented in this paper identify successful methods for predicting learning outcomes from conversational transcripts. However, the small size of this dataset makes it difficult to draw robust conclusions of statistical significance. Future work will look to explore the predictive power of Active Learning sequences in larger-scale and more diverse collaborative learning contexts, and to pursue the potential in combining textual cues with conversational sequence information in a more sophisticated ways. Further, we hope to use such models as real-time formative assessments based on similar conversational cues to direct instruction and provide agile conversational support for collaborative learning.

Acknowledgements. This research was funded in part by NSF grants IIS-1320064 and OMA-0836012.

References

[1] Adamson, D., Dyke, G., Jang., H., Rosé, C.P.: Towards an agile approach to adapting dynamic collaboration support to student needs. International Journal of AI in Education (2013)

[2] Baker, R.S., Goldstein, A.B., Heffernan, N.T.: Detecting learning moment-by-moment. International Journal of Artificial Intelligence in Education 21(1), 5–25 (2011)

[3] Barros, B., Verdejo, M.: An approach to analyse collaboration when shared structured workspaces are used for carrying out group learning processes. In: International Conference on Artificial Intelligence in Education. Citeseer, Le Mans (1999)

[4] Barros, B., Verdejo, M.F.: Analysing student interaction processes in order to improve collaboration. The Degree Approach. International Journal of Artificial Intelligence in Education 11(3), 221–241 (2000)

[5] D'Mello, S.K., Craig, S.D., Witherspoon, A., Mcdaniel, B., Graesser, A.: Automatic detection of learner's affect from conversational cues. User Modeling and User-Adapted Interaction 18(1-2), 45–80 (2008)

[6] D'Mello, S.K., Graesser, A.: Modeling cognitive-affective dynamics with hidden markov models. In: Proceedings of the 32nd Annual Cognitive Science Society, pp. 2721–2726 (2010)

[7] Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)

[8] Gianfortoni, P., Adamson, D., Rosé, C.P.: Modeling of stylistic variation in social media with stretchy patterns. In: Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, pp. 49–59. Association for Computational Linguistics (2011)

[9] Gunawardena, C.N., Lowe, C.A., Anderson, T.: Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. Journal of Educational Computing Research 17(4), 397–431 (1997)

[10] Henri, F.: Computer conferencing and content analysis. Series F: Computer and Systems Sciences (1992)

[11] Howley, I., Mayfield, E., Carolyn, P.: Linguistic analysis methods for studying small groups. In: The International Handbook of Collaborative Learning, ch. 10, Routledge (2013)

[12] Kim, J., Li, J., Kim, T.: Towards identifying unresolved discussions in student online forums. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–91. Association for Computational Linguistics (2010)

[13] Kirschner, F., Paas, F., Kirschner, P.A.: A cognitive load approach to collaborative learning: United brains for complex tasks. Educational Psychology Review 21 (2009)

[14] Kumar, R.: Socially capable conversational agents for multi-party interactive situations. Ph.D. thesis, Carnegie Mellon University (2011)

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

[16] Leitão, S.: The potential of argument in knowledge building. Human Development 43(6), 332–360 (2000)

[17] Martin, J.R., Rose, D.: Working with discourse: Meaning beyond the clause. Continuum International Publishing Group (2003)

[18] Mayfield, E., Adamson, D., Rosé, C.P.: Hierarchical conversation structure prediction in multi-party chat. SIGDIAL 2012 (2012)

[19] Mayfield, E., Adamson, D., Rudnicky, A.I., Rosé, C.P.: Computational representational of discourse practices across populations in task-based dialogue. ICIC, Bangalore (2012)

[20] Rabiner, L., Juang, B.: An introduction to hidden markov models. IEEE ASSP Magazine 3(1), 4–16 (1986)

[21] Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Computers & Education 68, 458–472 (2013)

[22] Scardamalia, M., Bereiter, C.: Technologies for knowledge-building discourse. Communications of the ACM 36(5) (1993)

[23] Soller, A., Lesgold, A.: Analyzing Peer Dialogue from an Active Learning Perspective. In: Proceedings of the AI-ED 99 Workshop: Analysing Educational Dialogue Interaction: Towards Models that Support Learning, pp. 63–71 (1999)

[24] Soller, A., Lesgold, A.: A Computational Approach to Analyzing Online Knowledge Sharing Interaction. In: Proceedings of Artificial Intelligence in Education 2003, Sydney, Australia (2003)

[25] Soller, A., Wiebe, J., Lesgold, A.: A Machine Learning Approach to Assessing Knowledge Sharing During Collaborative Learning Activities. In: Proceedings of Computer-Support for Collaborative Learning 2002, Boulder, CO (2002)

[26] Twitchell, D.P., Jensen, M.L., Derrick, D.C., Burgoon, J.K., Nunamaker, J.F.: Negotiation outcome classification using language features. Group Decision and Negotiation 22(1), 135–151 (2013)

[27] Webb, N.M., Palinscar, A.S.: Group processes in the classroom. In: Berliner, D.C., Calfee, R.C. (eds.) Handbook of Educational Psychology, pp. 841–873. Prentice Hall, New York (1996)