Predicting Dialogue Acts for Intelligent Virtual Agents with Multimodal Student Interaction Data

Wookhee Min
North Carolina State University
Raleigh, NC 27695
wmin@ncsu.edu

Joseph B. Wiggins
North Carolina State University
Raleigh, NC 27695
jbwiggi3@ncsu.edu

Lydia G. Pezzullo
Tufts University
Medford, MA 02155
lydia@learndialogue.org

Alexandria K. Vail
North Carolina State University
Raleigh, NC 27695
akvail@ncsu.edu

Kristy Elizabeth Boyer
University of Florida
Gainesville, FL 32611
keboyer@ufl.edu

Bradford W. Mott
North Carolina State University
Raleigh, NC 27695
bwmott@ncsu.edu

Megan H. Frankosky
North Carolina State University
Raleigh, NC 27695
rmhardy@ncsu.edu

Eric N. Wiebe
North Carolina State University
Raleigh, NC 27695
wiebe@ncsu.edu

James C. Lester
North Carolina State University
Raleigh, NC 27695
lester@ncsu.edu

ABSTRACT

Recent years have seen a growing interest in intelligent game-based learning environments featuring virtual agents. A key challenge posed by incorporating virtual agents in game-based learning environments is dynamically determining the dialogue moves they should make in order to best support students' problem solving. This paper presents a data-driven modeling approach that uses a Wizard-of-Oz framework to predict human wizards' dialogue acts based on a sequence of multimodal data streams of student interactions with a game-based learning environment. To effectively deal with multiple, parallel sequential data streams, this paper investigates two sequence-labeling techniques: long short-term memory networks (LSTMs) and conditional random fields. We train predictive models utilizing data corpora collected from two Wizard-of-Oz experiments in which a human wizard played the role of the virtual agent unbeknownst to the student. Empirical results suggest that LSTMs that utilize game trace logs and facial action units achieve the highest predictive accuracy. This work can inform the design of intelligent virtual agents that leverage rich multimodal student interaction data in game-based learning environments.

Keywords

Game-Based Learning, Virtual Agents, Deep Learning, Multimodal.

1. INTRODUCTION

Recent years have witnessed a growing interest in intelligent game-based learning environments because of their potential to simultaneously promote student learning and create engaging learning experiences [23]. These environments incorporate personalized pedagogical functionalities delivered with adaptive learning techniques and the motivational affordances of digital games featuring believable characters and interactive story scenarios situated in meaningful contexts [13, 23]. A key feature of game-based learning environments is their ability to embed problem-solving challenges within interactive virtual environments, which can enhance students' engagement and facilitate learning through customized narratives, feedback, and problem-solving support [18, 25].

Game-based learning environments offer considerable opportunities for implementing virtual agents by delivering visually contextualized pedagogical strategies [14]. Intelligent virtual agents have been shown to deliver motivational benefits, promote problem-solving, and positively affect students' perception of learning experiences [14]. Virtual agents play a variety of roles in interactive learning environments including intelligent tutors, teachable agents, and learning companions [4].

A key challenge in developing intelligent virtual agents is devising accurate predictive models that dynamically attune pedagogical strategies to individual students using evidence from students' interactions with the learning environment. Previous research has focused on when to intervene [21] and what types of dialogue moves to make during students' problem-solving activities [3] to provide support in a timely, contextually relevant manner. Selecting appropriate pedagogical dialogue moves is critical [24] because failing to provide effective feedback may lead to decreased learning in a student experiencing boredom [1], lead a student who is confused to become disengaged [10], or negatively impact the outcome of dialogues [5].

Much of the previous work in this line of investigation has addressed this challenge through computationally modeling agents' *dialogue acts*, the underlying intention (e.g., greeting, question, suggestion) of the utterances, by utilizing sequences of actions within learning environments as evidence [2]. The current work builds on this by examining multimodal data streams, which

can provide rich evidence of students' cognitive and affective states, in addition to evidence captured from game trace logs. To effectively deal with the granular sequential data in parallel multimodal data streams, we investigate two sequence labeling techniques: a deep-learning technique, long short-term memory networks (LSTMs) [11]; and a competitive baseline approach, conditional random fields (CRFs) [26]. This work is inspired by the recent success of LSTMs in dealing with low-level data (e.g., speech signals), and particularly by their state-of-the-art performance in speech recognition tasks [16]. Additionally, hierarchical representation learning supported by deep learning provides advantages over other machine learning techniques by avoiding the need for labor-intensive feature engineering [16].

Our sequence labeling models are evaluated with 211 dialogue acts made by human wizards who interacted with 11 students playing CRYSTAL ISLAND, a game-based learning environment for middle school microbiology [23]. The interaction data include game trace logs, facial action units [17] processed from facial video recordings, and galvanic skin responses, all of which are utilized as input features for devising predictive models. Wizards used pre-designed utterances, which they selected from menus organized by dialogue act. Each selected utterance was then delivered to the student via speech synthesis. Wizards could observe the student's face, gaze, game screen, and voice while selecting dialogue moves, but facial action units, galvanic skin responses, and game trace logs were not directly accessible. We hypothesize that these unobserved multimodal data streams serve as proxies for the wizards' dialogue decisions and examine these as explanatory variables to predict the next dialogue act that a human wizard might choose.

LSTM and CRF models are devised utilizing subsets of the parallel multimodal data streams. Student-level cross-validation studies indicate that LSTMs utilizing game trace logs and facial action units outperform both CRFs and the majority class-based baseline with respect to predictive accuracy. Further, we find that the LSTM model effectively takes advantage of multimodal data streams, and it most effectively utilizes both game trace logs and facial action unit data. The results suggest that LSTM models can serve as the foundation for dialogue act modeling for intelligent virtual agents that dynamically adapts dialogues to individual students.

2. RELATED WORK

Recent work in game-based learning has explored a broad spectrum of subject matters ranging from computer science [18] and language to cultural learning [13]. Narrative-centered learning environments, which provide narrative adaptation for individual students in the context of intelligent game-based learning, have been found to deliver experiences in which learning and engagement are synergistic [13, 23]. Student interaction data from game-based learning activities has provided a rich source of information from which students' development of competencies [18, 25] and progress towards learning goals [19, 20] are diagnosed. Game-based learning environments can also be populated by virtual agents, whose design should consider students' cognitive and affective states [4, 14].

In parallel work on tutorial dialogue, it has been found that tutorial planning can take into account students' cognitive and affective states [7]. Planning dialogue moves and inducing turn-taking policies have been widely examined in supervised learning (e.g., hidden Markov models [2], directed graph representations [5]) and reinforcement learning [3, 21]. The approach described in this paper is the first to investigate dialogue move classification using LSTMs and CRFs that take as input sequential multimodal data streams, which can serve as the foundation for guiding the dialogue of intelligent virtual agents in game-based learning environments.

Figure 1. The CRYSTAL ISLAND game-based learning environment.

3. CRYSTAL ISLAND

Over the past several years, our lab has been developing CRYSTAL ISLAND (Figure 1), a game-based learning environment for middle school microbiology [23]. Designed as a supplement to classroom science instruction, CRYSTAL ISLAND's curricular focus has been expanded to include literacy education based on Common Core State Standards for reading informational texts. The narrative focuses on a mysterious illness afflicting a research team on a remote island. Students play the role of a visitor who is drawn into a mission to save the team from the outbreak. Students explore the research camp from a first-person viewpoint, gather information about patient symptoms and relevant diseases, form hypotheses about the infection and its transmission source, use virtual lab equipment and a diagnosis worksheet to record their findings, and report their conclusions to the camp's nurse.

Extending the previous edition of CRYSTAL ISLAND, we incorporated a prototype virtual agent into the game to investigate both affective and cognitive influences on students' learning processes. This virtual agent, a young female scientist named Layla (Figure 2), was designed as a near-peer mentor who supports the student through dialogue-based interactions.

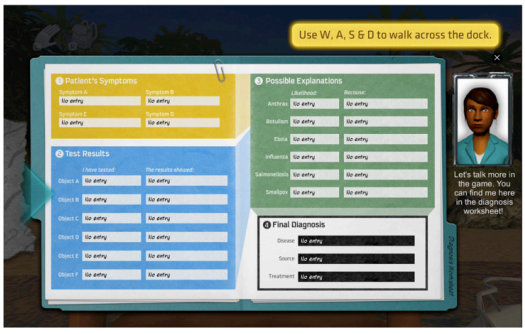

Figure 2. CRYSTAL ISLAND virtual agent.

In CRYSTAL ISLAND'S virtual world, students interact with learning resources such as books and posters, as well as with non-player characters through informative menu-based dialogue. As students progress through the game, they collect evidence and record their hypotheses in a "diagnosis worksheet." The student meets Layla when the diagnosis worksheet is opened (Figure 2).

With Layla's visual and speech synthesis prototypes in place, but no adaptive dialogue model implemented yet, a Wizard of Oz system was implemented to enable a human operator to provide the intelligence behind Layla's dialogue. When the human "wizard" decides to initiate a dialogue move, she chooses one of six dialogue acts (Table 1) from a menu interface, then selects a dialogue utterance from the act's set of pre-determined utterances. Layla then speaks the utterance through speech synthesis. The selection of dialogue moves was informed by the literature on dialogue systems for learning [8], as well as experience with a recent study conducted in the same middle school, in which pairs of middle school students interacted with CRYSTAL ISLAND together.

Three wizards controlled Layla's dialogue in the game from a room separated from the students, while observing the students through a live feed that included the student's facial video, the student's gaze superimposed in real time over a video capture of the game screen, and the student's voice as recorded through a headset microphone.

Data was collected in two studies implemented in the spring and summer of 2015 at a public middle school in Raleigh, North Carolina. In the spring study, participants were drawn from an after-school activity, and the summer study's participants were from classroom pull-outs. Of the 11 students who participated, 7 were female and 4 were male, with an average age of 12 (SD = 1.1). The data corpus contains 211 virtual agent dialogue acts across the students (average number of acts: 19.2, maximum number of acts: 41, and minimum number of acts: 3).

Table 1. Agent's dialogue acts and distributions of their use.

Dialogue Act	Distributions	Dialogue Act	Distributions
Greeting	58 (27.5%)	***Suggestion***	51 (24.2%)
Question	35 (16.6%)	***Feedback***	8 (3.8%)
Acknowledge-ment	43 (20.4%)	***Affective Statement***	16 (7.6%)

4. MULTIMODAL DATA

During the students' interactions with CRYSTAL ISLAND, both game actions and parallel sensor data were captured to collect both cognitive and affective features of students' experience. In the following subsections, we describe the three types of input data investigated in the present work.

4.1 Game Trace Logs

Students play CRYSTAL ISLAND using a keyboard and mouse. Student actions are logged for gameplay analysis and game telemetry [20]. In the present modeling work, seven key categories of actions are examined: moving around the camp, using the laboratory's equipment to test a hypothesis about the disease and its source, conversing with non-player characters, reading complex informational texts about microbiology concepts, taking embedded assessments associated with the informational texts, interacting with the diagnosis worksheet, and experiencing dialogue moves with the virtual agent. The total number of distinct actions is 143.

A total of 4,117 student actions were logged along with 211 dialogue acts by the virtual agent in the training data. Students took an average of 19.5 actions between two adjacent dialogue acts, where the minimum and maximum number of actions between any two adjacent dialogue acts are 1 and 217, respectively.

4.2 Galvanic Skin Response

Galvanic skin response (GSR) is a measurement of the level of conductance across the surface of the skin, which is driven by the activity of the sympathetic nervous system. GSR reflects a variety of cognitive and affective processes, including attention and engagement [6, 22]. In addition, the presence of significant spikes in students' GSR in response to certain events during a technology-supported learning activity has been found to be associated with learning-linked emotions and learning outcomes [12]. In this study, Empatica E4 bracelets on both wrists were used for GSR recording. These bracelets were chosen because, unlike palmar and fingertip GSR recording devices, they do not restrict the range of hand movement needed to play the game.

4.3 Facial Action Units

Facial expressions have been shown to have a relationship to self-reported and judged learning-centered affective states [1, 17]. Previous work has also found that facial expressions during learning can help predict a student's learning gains, frustration, and engagement [27]. Facial expressions can be examined non-invasively through video recordings taken during a student's interaction with a learning environment.

In this work, we observe facial expressions by analyzing a student's facial action units, which capture movement of the muscles in the face. Facial action units are grounded in the Facial Action Coding System, which was devised to make observations about facial movements [9]. In this study, facial videos were recorded via a webcam and analyzed using FACET, an automated system devised for tracking facial action units, because it allows for frame-by-frame tracking in the facial videos without the time intensive effort of human-tagging facial action units. FACET is the next generation of the Computer Expression Recognition Toolbox [17], which has been validated for both adults and children. In this study, we considered the subset of facial action units provided by FACET (Table 2). In the following section, we describe the deep learning-based dialogue act classifier that utilizes these three data sources.

Table 2. Facial action units examined.

Inner Brow Raiser (AU1)	Upper Lip Raiser (AU10)	Tightener (AU23)
Outer Brow Raiser (AU2)	Lip Corner Puller (AU12)	Lip Pressor (AU24)
Brow Lowerer (AU4)	Dimpler (AU14)	Lips Part (AU25)
Upper Lid Raiser (AU5)	Lip Corner Depressor (AU15)	Jaw Droop (AU26)
Cheek Raiser (AU6)	Chin Raiser (AU17)	Lip Suck (AU28)
Lid Tightener (AU7)	Puckerer (AU18)	
Nose Wrinkler (AU9)	Lip Stretcher (AU20)	

5. LSTM-BASED DIALOGUE MOVE DECISION MODEL

Long short-term memory networks (LSTMs) have demonstrated significant success in dealing with a series of raw signals, such as speech, yielding state-of-the-art performance in speech recognition tasks [16]. This inspires our work, which deals with low-level sensor data such as GSRs and facial AUs. In the following subsections, we present a high-level description of LSTMs [11], introduce how multimodal input data are synchronized and encoded into a trainable format, and describe how the LSTM-based dialogue move prediction models are configured.

5.1 LSTM Background

LSTMs are a type of gated recurrent neural network specifically designed for sequence labeling on temporal data. LSTMs, like standard recurrent neural networks, take the approach of sharing weights across layers at different time steps. LSTMs feature a sequence of memory blocks that include one or more self-connected memory cells along with three gating units [11]. In LSTMs, the input and output gates modulate the incoming and outgoing signals to the memory cell, and the forget gate controls whether the previous state of the memory cell is remembered or forgotten. This structure allows the model to preserve gradient information over longer periods of time [11].

In the implementation of LSTMs investigated here, the input gate (i_t), forget gate (f_t), and candidate memory cell state ($\widetilde{c_t}$) at time *t* are computed by Equations (1)–(3), respectively, in which *W* and *U* are weight matrices for the input (x_t) at time *t* and the cell output (h_{t-1}) at time *t*-1, *b* is the bias vector of each unit, and σ and *tanh* are the logistic sigmoid and hyperbolic tangent function, respectively.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$\widetilde{c_t} = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3)$$

Once these three vectors are computed, the current memory cell's state is updated to a new state (c_t) by modulating the current memory cell state candidate value ($\widetilde{c_t}$) via the input gate (i_t) and the previous memory cell state (c_{t-1}) via the forget gate (f_t). Through this process, a memory block decides whether to keep or forget the previous memory state and regulates the candidate of the current memory state via the input gate. This step is described in Equation (4), in which $\odot$ denotes element-wise multiplication:

$$c_t = i_t \odot \widetilde{c_t} + f_t \odot c_{t-1} \quad (4)$$

The output gate (o_t) calculated in Equation (5) is utilized to compute the memory cell output (h_t) of the LSTM memory block at time *t*, modulating the updated cell state (c_t) (Equation 6):

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \odot tanh(c_t) \quad (6)$$

Once the cell output (h_t) is calculated at time *t*, the next step is to use the computed cell output vectors to predict the label of the current training example. For the dialogue move decision model, we use the final cell output vector (h_t), assuming that h_t captures long-term dependencies from the previous time steps.

5.2 Data Encoding for Dialogue Move Decision Model

Each data stream from a suite of multimodal interaction data is of a sequential form. Because these data include fixed-rate recordings (e.g., facial action units and galvanic skin responses) with rates that differ between streams, as well as in-game action-driven recordings (e.g., game trace logs) with no set rate, the first step of data encoding is synchronizing input data across modalities.

We obtained from each student two series of galvanic skin responses (GSRs), one each for the left and right hand, as well as 19 facial action units (AUs). In the modeling work reported here, only the GSR information from the subject's dominant hand is utilized, so GSR is represented by a one-dimensional vector. AUs are represented by a 19-dimensional vector space per time stamp. GSR and AUs were logged with the frequencies of approximately 4Hz and 30 Hz, respectively. Game traces were recorded as events were triggered in the game, whenever the actions described in Section 4.1 were performed.

In contrast to GSR or AUs, which have continuous values, the game trace logs (GAME) consist of discrete indices for specific actions, indexed 1 to 143. To represent actions in a vector format, we employ the *one-hot-encoding* technique, in which a bit vector whose length is the total number of actions (143 in this work) is created while only the associated action bit is on (i.e., 1) while all other bits are off (i.e., 0). Once the vector representations for GAMEs are created, the next step is to synchronize the three data representations into an integrated representation.

To keep the length of data sequences manageable while preserving key game actions, we synchronize the multimodal data based on the game trace logs. All GSR and AU data collected between any two adjacent game actions are transformed into two vectors, using the following method:

- Vector 1: (75th percentile minus 50th percentile) per feature across all the data points between the two adjacent actions
- Vector 2: (50th percentile minus 25th percentile) per feature across all the data points between the two adjacent actions

We hypothesize that these two quartile-based vectors can capture variance of signals within an interval, while effectively avoiding outliers, smoothing out individual differences, and keeping the number of input features (183, or the sum of 143 for GAME, 38 for AU, and 2 for GSR) small enough to efficiently train LSTMs. Once these two vectors are created for the GSR stream and for each AU, the vectors are concatenated to the game trace log vector.

5.3 LSTM Model Configurations for Dialogue Move Decision

Prior to training LSTMs, the hyperparameters of the models must be determined. LSTM hyperparameters have often been explored using grid search or random search settings in the process of minimizing validation errors [20]. We adopt the grid search approach to empirically find an optimal configuration for a set of hyperparameters. In this work, we consider two hyperparameters: the number of hidden units for LSTMs among {32, 64} and the dropout rate [16], a model regularization technique, among {0.4, 0.7}. Both hyperparameters have significant influence on the performance of deep neural networks [11, 20].

In addition to LSTM-wide hyperparameters, this work also analyzes the isolated impacts of multimodal data sources. In order to perform this analysis, we examine all possible combinations of features, generating the following seven input feature sets: galvanic skin responses (GSRs), facial action units (AUs), game trace logs (GAMEs), GSRs and AUs, AUs and GAMEs, GSRs and GAMEs, and all three data sources. The dimension of a feature set is decided by summing up the dimensions of the features (see Section 5.2) that comprise the feature set.

In addition to the hyperparameters examined in the grid search, we apply a fixed value to the following hyperparameters for LSTMs: employing a softmax layer for classifying given sequences of interactions, adopting mini-batch gradient descent with a mini-batch size of 32, utilizing categorical cross entropy for the loss function, and employing a stochastic optimization method. The training process stops early if the validation score has not improved within the last 15 epochs. In this work, we evaluate our models using student-level leave-one-out cross validation, and so in each fold, 1 student's data is used for testing

(completely hidden) out of 11 students, while 8 students' and 2 students' data are utilized as the training and validation set, respectively. Finally, the maximum number of epochs is set to 100.

6. EVALUATION

To evaluate the proposed LSTM-based dialogue act classification (cast as six-class classification), we search for an optimal set of hyperparameters through cross-validation in the previously discussed grid search setting, and then perform feature-set level predictive performance analyses based on the chosen hyperparameters. Additionally, we compare each LSTM-based computational model to a competitive approach based on linear-chain conditional random fields (CRFs) [26] as well as a majority class baseline using the same cross-validation split for a pairwise comparison. CRFs are trained using the Block-Coordinate Frank-Wolfe optimization technique [15], and we adjust the regularization parameter for the optimization technique among {0.1, 0.5, 1.0} to find optimal CRFs as we do in LSTMs.

Table 3 presents feature-set-level cross-validation results. LSTMs with the hyperparameter configuration of 64 hidden units and 0.7 dropout rate achieve the highest predictive accuracy (34.1%), and CRFs trained with the regularization parameters of 0.5 achieved the second highest accuracy (32.2%). We use raw correct and incorrect prediction counts to calculate accuracy rates rather than reporting fold-based averaged accuracy rates, in an effort to avoid the potential for skew brought on by the wide variation in the number of data points per student (min: 3; max: 41).

Table 3. Student-level leave-one-out cross validation results across feature sets (64 hidden units and 0.7 dropout rate for LSTMs and 0.5 regularization parameter for CRFs).

	LSTMs	CRFs
GSRs	28.0%	19.9%
AUs	21.8%	25.6%
GAMEs	29.4%	**32.2%**
GSRs / AUs	26.1%	22.3%
AUs / GAMEs	**34.1%**	30.8%
GSRs / GAMEs	29.9%	29.4%
GSRs / AUs / GAMEs	31.3%	27.0%

In the evaluation, LSTMs that achieve the highest predictive accuracy utilize AUs and GAMEs ($LSTM_{AU/GAME}$), the accuracy of which constitutes a 43.9% marginal improvement over the baseline accuracy (23.7%). Note that the baseline accuracy is different from Table 1, because it is influenced by the random split made in cross validation. We conducted a Wilcoxon signed rank, a non-parametric statistical test for two related samples, to compare cross-validation results between the $LSTM_{AU/GAME}$ and the majority class baseline per fold. The test finds a statistically significant difference between $LSTM_{AU/GAME}$ and the baseline (Z=-2.25, p=0.024). The differences between $LSTM_{AU/GAME}$ and the best performing CRFs (p=0.67) and between the CRFs and the baseline (p=0.095) are not statistically significant.

It is noteworthy that AUs by themselves do not achieve a high predictive accuracy. This can be partially explained by noting that the facial action unit data stream was often temporarily lost (a vector filled with zeros is used in this case for the missing data), usually when the subject's face was not properly situated within the camera screen. It is surprising, however, to see that partially-missing AUs synchronized with GAMEs data helped improve the prediction of the next virtual agent dialogue act by outperforming GAMEs models (Z=-1.71, p=0.088) as well as AUs models (Z=-2.24, p=0.025).

The $LSTM_{AU/GAME}$'s outperformance might be explained by the information available to the human wizards as they chose dialogue acts: they were able to watch the subject's game play as well as facial expressions during the interaction with the game, which together potentially influenced the dialogue decisions. On the other hand, the AUs likely characterize aspects of the subject's affective states, and they can contribute to the improved predictive performance synergistically with GAMEs in LSTMs.

Overall, GAMEs serve as a strong predictor relative to other independent data sources: GAMEs models (29.4%) outperform the other two independent models induced utilizing GSRs (28.0%) or AUs (21.8%); in the meantime, each feature set that leverages GAMEs in addition to other data sources outperforms the corresponding feature set without the GAMEs (e.g., GSRs, AUs, and GAMEs (31.3%) vs. GSRs and AUs (26.1%)). Sequences of actions in the GAMEs may reflect students' underlying cognitive states such as plans, goals, and knowledge during problem-solving activities [19, 20], which wizards attempted to address through their dialogue act choices. It is expected that LSTMs' capacity for hierarchical feature abstraction enables them to recognize these high-level patterns from low-level action sequences.

It is interesting to observe that GSRs by themselves outperform the baseline but incorporating GSRs with AUs and GAMEs (31.3%) does not outperform $LSTM_{AU/GAME}$ (34.1%). Although much of the previous research has used GSR data streams as evidence for modeling humans' affective and cognitive states [22], the findings of the study presented here suggest that GSR collected using wrist sensors may not be the most informative data source for predicting a human-operated virtual agent's next dialogue act, particularly when other data sources are available.

7. CONCLUSION AND FUTURE WORK

Dialogue modeling is a critical functionality for pedagogically adaptive virtual agents. This paper has presented two sequence-modeling approaches to classifying human wizards' dialogue moves when utilizing multimodal observation sequences. Both conditional random fields (CRFs) and long short-term memory networks (LSTMs) have demonstrated significant promise as effective modeling techniques on the sequential, parallel, multimodal data from game trace logs, galvanic skin response, and facial action units. Both CRFs and LSTMs outperform the majority class-based baseline with respect to predictive accuracy, while LSTMs achieve the highest predictive accuracy. Feature-level analyses of LSTMs suggest that even incomplete facial action unit data can augment LSTMs' predictive performance along with game trace logs, while game trace logs serve as strong predictor in both computational approaches. Along with achieving a substantial improvement in the use of sequence labeling techniques, this work suggests a number of directions for future work.

First, it will be important to extend the current models to determine the timing of dialogue acts. Together with the current work, this will further enhance the potential capacity for intelligent virtual agents to provide adaptive pedagogical support. Second, it will be important to examine the relationships between students' cognition and affect as perceived by human wizards, and to investigate how they influence wizards' dialogue decision-making. Because multimodal interaction data may reflect students' affective and cognitive states, identifying the relationship between student models and dialogue acts can guide the design of advanced tutorial dialogue management capabilities for pedagogical agents.

8. ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation through Grant CHS-1409639. Any opinions, findings, conclusions, or recommendations expressed are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

9. REFERENCES

[1] Baker, R., D'Mello, S., Rodrigo, M.M. and Graesser, A. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human Computer Studies*. 68, 4, 223–241.

[2] Boyer, K., Phillips, R., Ha, E., Wallis, M., Vouk, M. and Lester, J. 2010. Leveraging Hidden Dialogue State to Select Tutorial Moves. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. 66–73.

[3] Chi, M., Vanlehn, K., Litman, D. and Jordan, P. 2011. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*. 21, 1-2, 83–113.

[4] Chou, C.Y., Chan, T.W. and Lin, C.J. 2003. Redefining the learning companion: The past, present, and future of educational agents. *Computers and Education*. 40, 3, 255–269.

[5] D'Mello, S.K., Olney, A. and Person, N.K. 2010. Mining Collaborative Patterns in Tutorial Dialogues. *Journal of Educational Data Mining*. 2, 1, 1–37.

[6] Dawson, M.E., Schell, A.M. and Filion, D.L. 2007. The Electrodermal System. *The Handbook of Psychophysiology*. 200–223.

[7] DeVault, D. et al. 2014. SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. 1061–1068.

[8] Dweck, C.S. 2002. The development of ability conceptions.

[9] Ekman, P. and Friesen, W. V 1977. Facial action coding system.

[10] Forbes-Riley, K. and Litman, D. 2012. Adapting to Multiple Affective States in Spoken Dialogue. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 217–226.

[11] Graves, A. 2012. *Supervised sequence labelling with recurrent neural networks*. Springer.

[12] Hardy, M., Wiebe, E., Grafsgaard, J., Boyer, K. and Lester, J. 2013. Physiological Responses to Events During Training: Use of Skin Conductance to Inform Future Adaptive Learning Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2101–2105.

[13] Johnson, W.L. 2010. Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education*. 20, 175–195.

[14] Johnson, W.L. and Lester, J.C. 2015. Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of Artificial Intelligence in Education*. 25, 25–36.

[15] Lacoste-Julien, S., Jaggi, M., Schmidt, M. and Pletscher, P. 2013. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. *Proceedings of the 30th International Conference on Machine Learning*. 28, 9.

[16] LeCun, Y., Bengio, Y. and Hinton, G. 2015. Deep Learning. *Nature*. 521, 7553, 436–444.

[17] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. and Bartlett, M. 2011. The Computer Expression Recognition Toolbox (CERT). *Automatic Face Gesture Recognition and Workshops (FG 2011)*. 298–305.

[18] Min, W., Frankosky, M., Mott, B., Rowe, J., Wiebe, E., Boyer, K. and Lester, J. 2015. DeepStealth: Leveraging Deep Learning Models for Stealth Assessment in Game-Based Learning Environments. *Proceedings of the 17th International Conference on Artificial Intelligence in Education*, 277–286.

[19] Min, W., Ha, E.Y., Rowe, J., Mott, B. and Lester, J. 2014. Deep Learning-Based Goal Recognition in Open-Ended Digital Games. *Proceedings of the 10th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 37–43.

[20] Min, W., Mott, B., Rowe, J., Liu, B. and Lester, J. 2016. Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks. *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. In Press.

[21] Mitchell, C., Boyer, K. and Lester, J. 2013. Evaluating State Representations for Reinforcement Learning of Turn-Taking Policies in Tutorial Dialogue. *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 339–343.

[22] Poh, M.Z., Swenson, N.C. and Picard, R.W. 2010. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*. 57, 5, 1243–1252.

[23] Rowe, J., Shores, L., Mott, B. and Lester, J. 2011. Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*. 21, 1-2, 115–133.

[24] Shute, V.J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., Ventura, M. and Almeda, V. 2015. Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*. 86, 224–235.

[25] Shute, V.J. and Ventura, M. 2013. *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

[26] Sutton, C. and McCallum, A. 2012. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*. 4, 4, 267–373.

[27] Vail, A., Grafsgaard, J., Wiggins, J., Lester, J. and Boyer, K. 2014. Predicting Learning and Engagement in Tutorial Dialogue: A Personality-Based Model. *Proceedings of the 16th ACM International Conference on Multimodal Interaction*. 255–262.