Zhao, R., Sinha, T., Black, A., Cassell, J. (2016) "Socially-Aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior." Proceedings of Intelligent Virtual Agents (IVA 2016).

Socially-Aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior

Ran Zhao, Tanmay Sinha, Alan W Black, Justine Cassell
Language Technologies Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213 USA
{rzhao1,tanmays,awb,justine}@cs.cmu.edu

Abstract. This work focuses on data-driven discovery of the temporally co-occurring and contingent behavioral patterns that signal high and low interpersonal rapport. We mined a reciprocal peer tutoring corpus reliably annotated for nonverbals like eye gaze and smiles, conversational strategies like self-disclosure and social norm violation, and for rapport (in 30 second thin slices). We then performed a fine-grained investigation of how the temporal profiles of sequences of interlocutor behaviors predict increases and decreases of rapport, and how this rapport management manifests differently in friends and strangers. We validated the discovered behavioral patterns by predicting rapport against our ground truth via a forecasting model involving two-step fusion of learned temporal associated rules. Our framework performs significantly better than a baseline linear regression method that does not encode temporal information among behavioral features. Implications for the understanding of human behavior and social agent design are discussed.

1 Introduction and Motivation

The year is 2025. Zack comes into math class with his personalized virtual peer agent Zoe projected on his glasses. Zoe smiles as she says to Zack, "You look tired today. I told you it was a bad idea to play "AR Starcraft" that late on weeknights!". Zack grimaces "OK, so I'm tired. But it was awesome! The whole math class was getting to know one another - that's work, right?" to which Zoe nods and responds by indexing their shared experience - "Perhaps, but last time you did this, I was too exhausted the next day to help you."

Zack and Zoe then work on the math task they are supposed to complete. Zoe starts off - "We need to solve this set of linear equations 5x*(3x-18)=10 first". Zack seems a bit confused "Well, I'm familiar with fractions, but I suck at linear equations." Zoe gazes at the work sheet, then back at Zack and finally provides motivational scaffolding in the form of negative self disclosure followed by praise, in order to boost their interpersonal bond, and Zack's confidence "Don't worry, I used to suck at linear equations too, but you're a rockstar at this stuff. You'll be fine. Besides which, we'll go through it together," following with a smile.

This vision illustrates several factors related to the important role that the relationship between learners, or learners and their tutors, can play in improving learning gains. While this phenomenon is described in the educational literature [20, 23], there has existed no rigorous models of the mechanism underlying the relationship between social and cognitive functioning in tasks such as these [24], nor do there exist computational models of interpersonal closeness that can drive the functioning of an intelligent tutor. There is therefore great opportunity to expand on the social capabilities of current educational technologies in order to create long-term interpersonal connectedness in the service of increased adaptivity in learning [30], and thereby increased learning gains. In this vein, here we investigate the dynamics of social interaction in longitudinal peer tutoring, as manifested in manifested in verbal and nonverbal behaviors. The aspect of social interaction that we focus on is rapport management, as rapport is argued to be one of the central constructs necessary to understanding successful helping relationships [4]), and rapport management is abundantly present in peer tutoring [32]

Let us at this point step back to describe what we mean by rapport. Rapport is often defined as ”a close and harmonious relationship in which the people concerned appear to understand each other's feelings or ideas and communicate well," however we feel it is best described by examples and so, below, are two examples from our corpus, of high and low rapport, respectively.

High Rapport:	Low Rapport:
P1:I suck at negative numbers;	***P2: [silent][long pause]***
P2: it's okay so do I;	***P1: shh;[long pause]***
***P1:*{*smile*}**	***P2: alright;***
P2:uh actually no I don't, negative numbers are easy	***P1: let me do my work;***

In our own prior work, we proposed a computational model of long-term interpersonal rapport to explain how humans in dyadic interactions build, maintain and destroy rapport through the use of specific conversational strategies [38]. Because these strategies function to fulfill specific social goals and are instantiated in particular verbal and nonverbal behaviors, studying the synergistic interaction of conversational strategies and nonverbal behaviors on rapport management is important. To do so, not only a qualitative examination of certain dyadic behavior patterns that benefit or hurt interpersonal rapport is essential, but it is also desirable to build automated frameworks to learn fine-grained behavioral interaction patterns that index such social phenomena. The latter has received less attention, in part due to the time-intensive nature of collecting and annotating behavioral data for different aspects of interpersonal connectedness, and the difficulty of developing and using machine learning algorithms that can take the time course of interaction among different modalities and between interlocutors into account. Learning fine-grained behavioral interaction patterns that index rapport is the focus of the current work. There are three key issues that we believe should be taken into consideration when performing such assessment.

(1) When the foundational work by [35] described the nature of rapport, three interrelating components were posited: positivity, mutual attentiveness and coordination. Their work demonstrated, that over the course of a relationship, positivity decreases and coordination increases. Factors such as these, then, depend on the stage of relationship between interlocutors [38], and therefore it is necessary to take into account the relationship status of a dyad when extracting dyadic patterns of rapport. (2) while our previous work [28] discovered some of the common behaviors exhibited by dyads in peer tutoring to build or maintain rapport; playful teasing, face-threatening comments, attention-getting, etc., tutors and tutees were looked at separately, and each of these behaviors was examined in isolation from one another. In the current work, our interest is in moving beyond individual behaviors to focus on temporal sequences of such

behaviors in the dyadic context. Likewise, our prior work did not distinguish between rapport management during task (tutoring) vs social activities. We believe that the interactions between verbal and nonverbal behaviors may manifest differently in social and tutoring periods, since the roles of a tutor and tutee are more evident in the tutoring compared to the social periods. (3) Most prior computational work examining rapport, such as [12, 13, 18], has used post-session questionnaires to asses rapport. However, to measure the effect of multimodal behavioral patterns on rapport and better reason about the dynamics of social interaction, a finer-grained ground truth for rapport is needed.

In this paper, then, we take a step towards addressing the above limitations. To create a longitudinal peer tutoring corpus, we compared friend to stranger dyads, bringing each dyad back for five face-to-face sessions over five weeks. In each session, two tutoring periods were interspersed with three social periods. The students switched roles so that each both tutored and was tutored. We employed thin-slice coding [2] to elicit ground truth for rapport, by asking naive raters to judge rapport for every 30 second slice of the hour long peer tutoring session, presented to raters in a randomized order. This, in turn allowed us to analyze fine-grained sequences of verbal and nonverbal behaviors that were associated with high or low rapport between the tutor and tutee.

As a side note, while the current paper addresses these phenomena in the context of peer tutors and intelligent tutoring agents, this work is part of a larger research program that targets more general models of how to predict rapport between interlocutors in real time, using as input the interaction among linguistic (verbal) and nonverbal (visual) behaviors. This basic science serves as input in some of our work into embodied conversational agents that can use the dyad's current rapport as part of a decision about what to say next to manage rapport with the user as, in turn, input into a decision about how best to help the user achieve his/her goals, goals that include, in some of our agents, peer tutoring.

2 Related Work

2.1 Individual-focused Temporal Relations

The study of temporal relationships between verbal and nonverbal behaviors has been of prime importance in understanding various social and cognitive phenomena. A lot of this work has focused on the observable phenomena of interaction (low level linguistic, prosodic or acoustic behaviors that can be automatically extracted) or has leveraged computational advances to extract head nods, gaze, facial action units, etc., as a step towards modeling co-occurring and contingent patterns inherent in an individual person's behavior. Since feature extraction approaches that aggregate information across time are not able to explicitly model temporal co-occurrence patterns, two popular technical approaches to investigate temporal patterns of verbal and nonverbal behaviors are histogram of co-occurrences [29] and motif discovery methods [27].

For instance, [21] presented a study of co-occurrence patterns of human nonverbal behaviors during intimate self-disclosure. However, contingent relations between different nonverbal behaviors was not considered, which could extensively contribute to the design of a social agent that interacts with a human over time. [36] learned behavioral indicators that were correlated to expert judgess opinions of each key performance aspect of public speaking. They fused the modalities by utilizing a least squared boosted regression ensemble tree and predicted speaker performance. However, this work also did not consider the effect of interactions among different modalities and their temporal relations. In similar vein, [6] introduced deep conditional neural fields to model the generation of gestures by integrating verbal and acoustic modalities, while using an undirected second-order linear chain to preserve temporal relations between gestures as well. However, this approach only modeled individual co-verbal gestures, without considering interaction between the speaker and the interlocutor.

In [17] temporal combinations of individual facial signals (such as nod, smiles etc.) were used to infer positive (agree, accept etc.) and negative (dislike, disbelief etc.) meanings via ratings by humans. An interesting take-away from this work was that a combination of signals could significantly alter the perceived meaning. For instance, facial tension alone and frown alone did not mean “dislike, but the combination frown and tension did. Tilt alone and gaze right down alone did not mean “not interested as significantly as the combination tilt and gaze. However, while a combination of these nonverbals signaled higher level constructs (that were in turn associated with some pragmatic meaning), the authors were more interested in how these combinations were perceived by humans, rather than necessarily in a predictive task or testing these combinations in a human-agent dialog.

2.2 Dyadic Temporal Relations

In a conversation, attending to the contribution of both interactants adds greater complexity in reasoning about the social aspects of the interaction.Listeners show their interest, attention and understanding in many ways during the speakers utterances. Such “listener responses” [10], which may be manifested through gaze direction and eye contact, facial expressions, use of short utterances like “yeah”, “okay”, and “hm-m” etc or even intonation, voice quality and content of the words, are carriers of subtle information. These cues may convey information regarding understanding (whether the listeners understand the utterance of the speaker), attentiveness (whether the listeners are attentive to the speech of the speaker), coordination, and so forth.

For instance, [14] looked at observable lexical, acoustic and prosodic cues produced by the speaker followed by back channeling from the listener. The authors found that the likelihood of occurrence of a backchannel from the interlocutor appeared to increase with simultaneous occurrence of one or more cues by the speaker, such as final rising intonation, higher intensity and pitch levels, longer inter-pausal units (maximal sequence of words surrounded by silence longer than 50 ms) etc. However, in this work, no attempt was made to use the temporal sequence or co-occurrence of observables preceding a backchannel to predict higher level social constructs such as positivity, coordination, attentiveness, or underlying psychological states such as rapport or trust.

[1] explored the interplay between head movements, facial movements like smile and eye brow raising, and verbal feedback in a range of conversational situations, including continued attentiveness, understanding, agreement, surprise, disappointment, acknowledgment and refusing information. As the situations became more negative (disappointment, refusing information), the accompanying nonverbals became more extensive in time - no longer just a head nod, but a series of movements. The authors claim that this series of movements functioned to add some extra information or to emphasize or contradict what had been said, but ground truth was not provided for these claims.

Finally in [7], the authors used sequence mining methods to automatically extract nonverbal behavior sequences of the recruiters that were representative of interpersonal attitudes. Then, Bayesian networks were deployed to build a generation model for computing a set of nonverbal sequence candidates, which were further ranked based on the previously extracted frequent sequences. Even though this work considered the effect of sequencing of nonverbal signals, their model could be improved by the addition of temporal information inside these sequences, the addition of verbal signals and modeling of listeners' behaviors as well.

3 Study Context

3.1 Data

Reciprocal peer tutoring data was collected from 12 American English-speaking dyads (6 dyads were friends and 6 strangers; 6 were boys and 6 girls), with a mean age of 13 years old ranging from 12 to 15, who interacted for 5 hourly sessions over as many weeks (a total of 60 sessions, and 5400 minutes of data), tutoring one another on procedural and conceptual aspects of linear equations [37]. All interactions were videotaped from three camera views (a frontal view of each participant and a side view of the two participants). Speech was recorded by lapel microphones in separate audio channels. Each session began with a period of getting to know one another, after which the first tutoring period started, followed by another small social interlude, a second tutoring period with role reversal between the tutor and tutee, and then the final social time. Prior work demonstrates that peer tutoring is an effective paradigm that results in student learning [31], making this an effective context to study dyadic interaction with a concrete task outcome. Our student-student data demonstrates that a tremendous amount of rapport-building takes place during the task of reciprocal tutoring [33]. In their recent review of the research on design spaces for computer supported reciprocal tutoring, [8] emphasize reciprocal tutoring to be a natural extension of one-on-one tutoring in today's networked world.

3.2 Annotations

We assessed rapport-building via thin slice annotation [2], or rapidly made judgments of interpersonal connectedness in the dyad, based on brief exposure to their verbal and nonverbal behavior. Naive raters were provided with a simple definition of rapport and three raters annotated every 30 second video segment of the peer tutoring sessions for rapport using a 7 point likert scale. Weighted majority rule was deployed to mitigate bias from the ratings of different annotators, account for label over-use and under-use and pick a single rapport rating for each 30 second video segment. The segments were presented to the annotators in random order so as to ensure that raters were not actually annotating the delta of rapport over the course of the session. Prior work has shown that such reliably annotated measures of interpersonal rapport are causally linked to behavioral convergence of low-level linguistic features (such as speech rate etc,) of the dyad [32, 33] and that greater likelihood of being in high rapport in the next 30 sec segment (improvement in rapport dynamics over the course of the interaction) is positively predictive of the dyad's problem-solving performance.

In addition, we also annotated the entire corpus for conversational strategies such as self-disclosure (Krippendorf's α= 0.753), reference to shared experience (α= 0.798), praise (α=1), social norm violation (α= 0.753) and backchannel (α= 0.72) in the first pass, and reciprocity in these strategies (using a time window of roughly 1 minute) in the second pass (α= 0.77). [34] has investigated the phenomenon of congruence or interpersonal synchrony in usage of such conversational strategies, in absolute number as well as the pattern of timings, and found positive relationships with rapport and problem-solving performance. In other work, we have also shown that these conversational strategies can be reliably detected from observable indicators of verbal, visual and acoustic cues an accuracy of over 80% and kappa ranging from 60-80% [39]. Finally, our temporal association rule framework comprised of nonverbal behaviors like eye gaze (Krippendorf's α= 0.893) and smiles (α= 0.746), which we have found to significantly co-occur with conversational strategies [39].

4 Method

The technical framework we employ in this work is essentially an approach for pattern recognition in multivariate symbolic time sequences, called the Temporal Interval Tree Association Rule Learning (Titarl) algorithm [15]. Since it is practically infeasible to predict exactly when certain behavioral events happen, it is suitable to use probabilistic approaches that can extract patterns with some degree of uncertainty in the temporal relation among different events. Temporal association rules, where each rule is composed of certain behavioral pre-conditions (input events) and behavioral post-conditions (output events), are one such powerful approach. In our case, input events are conversational strategies and nonverbal behaviors such as violation social norms, smile etc. The output event is the absolute value of thin-slice rapport. Because interpersonal rapport is a social construct that is defined at the dyadic level, the applied framework helps reveal interleaved behavioral patterns from both interlocutors. An example of a simple generic temporal rule is given below. It illustrates the rule's flexibility by succinctly describing not only the temporal inaccuracy of determining the temporal location of output event, but also its probability of being fired.

"If event A happens at time t, there is 50% chance of event B happening between time t+3 to t+5".

Intuitively, the Titarl algorithm is used to extract large number of temporal association rules (r) that predict future occurrences of specific events of interest. The dataset comprises both multivariate symbolic time sequences $E_{i=1\ldots n}$ and multivariate scalar time series $S_{i=1\ldots m}$, where $E_i = \{t_j^i \in \mathbb{R}\}$ is the set of times that event e_i happens and S_i is an injective mapping from every time point to a scalar value. Before the learning process, a parameter w or the window size is specified, which allows us at each time point t to compute the probability for the target event to exist in the time interval $[t, t + w]$.

The four main steps in the Titarl algorithm [15] are: (i) exhaustive creation of simple unit rules that are above the threshold value of confidence or support, (ii) addition of more input channels in order to maximize information gain, (iii) production of more temporally precise rules by decreasing the standard deviation of the rule's probability distribution, (iv) refinement of the condition and conclusion of the rules by application of Gaussian filter on temporal distribution. Confidence, support and precision of the rule are three characteristics to validate its interest and generalizability. For a simple unit rule r: $e_1 \xrightarrow{[t,t+w]} e_2$ (confidence: x%, support:y%), confidence refers to the probability of a prediction of the rule to be true, support refers to the percentage of events explained by the rule and precision is an estimation of the temporal accuracy of the predictions.

$$confidence_r = P((t \in E_1)|(t' \in E_2), t' - t \leq w) \tag{1}$$

$$support_r = \frac{\{\#e_2|\text{r is active}\}}{\#e_2} \tag{2}$$

$$precision_r = \frac{1}{\text{standard deviation}_r} \tag{3}$$

5 Experimental Results

We first separated out friend and stranger dyads to learn rules from their behaviors separately. We also tagged the data as occurring during a social or tutoring period, and as being generated by a tutor or a tutee. We then randomly divided the friend and stranger groups into a training set (4 dyads) and test set (2 dyads). In the first experiment, we extracted a potentially large number of temporal association rules affiliated with each individual rapport state (from 1 to 7). In this experiment, for each event, we looked back 60 seconds to find behavioral patterns associated with it. A representative example is shown in figure 1, and descriptions of some of the rules in the test set whose confidence are above 50% and for whom the number of cases the rule applies to are more than 20 times are described below, divided into friends (F) and strangers (S) and into high rapport (H), defined as thin-slice rapport states 5, 6, and 7 and low rapport (L), defined as states 1, 2, and 3.

5.1 Behavioral Rules for Friends

There are 14,458 total rules for friends with confidence higher than 50%, 14,345 of which apply to friends in high rapport states. Overall, engaging in reference to shared experience, smiling while violating a social norm and overlapping speech are associated with high rapport. Examples are:

FH 1 *One of the student smiles while the other violates a social norm (Social period)*
FH 2 *One of the students refers to shared experience (Social period)*
FH 3 *One student smiles and violates a social norm, and the second smiles and gazes at the partner within the next minute (Social period)*
FH 4 *The two conversational partners overlap speech while one is smiling, following which the second starts smiling within the next minute (Social period)*
FH 5 *The tutee reciprocates a social norm violation while overlapping speech with the tutor, following which the tutor smiles and violates a social norm (Task period)* **[shown in Figure 1]**

In contrast to the high number of rules with confidence higher than 50% for friends in high rapport, there are only 113 rules that satisfy these criteria for friends in low rapport. Some examples are:

FL 1 *The tutor finishes violating a social norm while gazing at the tutee's work sheet, and within the next minute the tutee follows up with a social norm violation, but gazing at his/her own work sheet (Task period)*

FL 2 *The tutor reciprocates a social norm violation without a smile and neither the tutee nor the tutor gaze at one another. Meanwhile, the tutee begins violating another social norm within the next minute (Task period)*

FL 3 *The tutor backchannels while gazing at his/her own work sheet and does not smile. Moreover, the tutor also overlaps with the tutee in the next minute (Task period)*

5.2 Behavioral Rules for Strangers

There are 761 total rules for strangers, of which 130 are rules that apply to strangers in high rapport. In general, smiling and overlapping speech while using particular conversational strategies are associated with high rapport. Some examples are:

SH 1 *One of the interlocutors smiles while the other gazes at him/her and begins self-disclosing, and they overlap speech within the next minute (Social period)*

SH 2 *One of the interlocutors smiles and backchannels in the next minute (Social period)*

SH 3 *The interlocutors' speech overlaps and the tutee smiles within the next minute (Task period)*

631 rules, then, explain strangers in low rapport. Interestingly, rules that explain low rapport among strangers most often come from task periods. In general, overlapping speech after a social norm violation leads to low rapport in strangers. Some examples are:

SL 1 *The tutor smiles and gazes at the worksheet of the tutee while the tutee does not smile (Task period)*

SL 2 *The tutor violates social norms while being gazed at by the tutee, and their speech overlaps within the next minute (Task period)*

SL 3 *The tutor smiles and the tutee violates a social norm within the next 30 seconds, before their speech overlaps within the next 30 seconds (Task period)* **[shown in Figure 2]**

6 Validation and Discussion

In order to demonstrate that the extracted temporal association rules can be reliably used for forecasting of interpersonal human behavior, we first applied machine learning to perform an empirical validation, which we describe in the next subsection. The motivation behind constructing this forecasting model was to prove the automatically learned temporal association rules are good indicators of the dyadic rapport state. In the subsequent subsections of the discussion, we will discuss implications of our work for the understanding of human behavior and the design of "socially-skilled" agents, linking prior strands of research.

6.1 Estimation of Interpersonal Rapport

In addition to its applicability to sparse data, one of the prime benefits of the temporal association rule framework to predict a high-level construct such as rapport lies in its flexibility in modeling presence/absence of human behaviors and also the inherent uncertainty of such behaviors, via a probability distribution representation in time. In summary, the estimation of rapport comprises two steps: in the first step, the intuition is

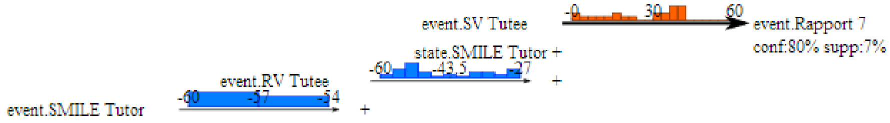

Fig. 1: **Friends in high rapport - The tutee reciprocates a social norm violation while overlapping speech with the tutor, following which the tutor smiles while the tutee violates a social norm.**

An example from the corpus is shown below:

Tutor: Sweeney you can’t do that, that’s the whole point{smile}; **[Violation of Social Norm]**
Tutee: I hate you.I’ll probably never never do that; **[Reciprocate Social Norm Violation]**
Tutor: Sweeney that’s why I’m tutoring you{smile};
Tutee: You’re so oh my gosh{smile}.We never did that ever; **[Violation of Social Norm]**
Tutor: {smile}What’d you say?
Tutee: Said to skip it{smile};
Tutor: I can just teach you how to do it;

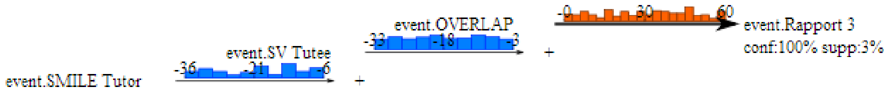

Fig. 2: **Strangers in low rapport - The tutor smiles and the tutee violates a social norm within the next 30 seconds, before their speech overlaps within the next 30 seconds.**

An example from the corpus is shown below:

Tutee: divide oh this is so hard let me guess;eleven;
Tutor: you know;
Tutee: six;
Tutor: next problem is is exactly the samesmile, over eleven equals, eleven x over eleven;
Tutee: I don’t need your help; **[Violation of Social Norm]**
Tutor: {Overlap}That is seriously like exactly the same.

to learn the weighted contribution (vote) of each temporal association rule in predicting the presence/absence of a certain rapport state (via seven random-forest classifiers); in the second step, the intuition is to learn the weight of each binary classifier for each rapport state, to predict the absolute continuous value of rapport (via linear regression). For clarity, we will use the following three mathematical subscripts to represent different types of index. i: index of output events, k: index of time-stamps,j: index of temporal association rules.

Each individual rapport state is treated as a discrete output event e_i, where $i = 1, 2, 3, 4, 5, 6, 7$. We learn the set of temporal association rules $R_i = \{r^i_j\}$ for each output event e_i. In the first step, a matrix M_i is constructed with $|T_i|$ rows and $1 + |R_i|$ columns, where $T_i = \{t^i_k \in \mathbb{R}\}$ denotes the set of time-stamps at which at least one of the rules in set R_i is activated. $M_i(k, j) \in [0, 1]$ denotes confidence of the rule r^i_j at the particular time point t^i_k. The extra column represents the indicator function of rapport

Relationship Status	**t-test value**	**Mean value (Mean Square Error)**	**Effect Size**
Friends	t(1,14)=-6.41***	Titarl=1.257, Linear Regression=2.120	-0.42
Strangers	t(1,14)=-8.78***	Titarl=0.837, Linear Regression =1.653	-0.62

Table 1: Statistical analysis comparing mean square regression of Titarl-based regression and a simple linear regression, for all possible combination of training and test sets in the corpus. Effect size assessed via Cohen's d. Significance: ***:$p < 0.001$, **:$p < 0.01$, *:$p < 0.05$

state: $M_i(k, |R_i| + 1) = \{1, \text{if } t_k^i \in E_i; 0 \text{ otherwise}\}$. Seven random-forest classifiers ($f_i(t)$ and $t \in T_i$)) are then trained on each corresponding matrix M_i using the last column (binary) as the output label and all other columns as input features [16]. In the second step, another matrix G with $|T|$ rows and 1+$|C|$ columns is formalized, where $|C|$ is the number of random-forest classifiers, $G(k, i) = f_i(t_k)$ and $T = \{t_k | t_k \in T_i, i = 1...7\}$. The last column is the absolute number of rapport state gathered by ground truth. This matrix is used to train a linear regression model.

For our corpus, as part of the Titarl-based regression approach, we first extracted the top 6000 rules for friend dyads and 6000 rules for stranger dyads from the training dataset, with the following parameter settings: minimum support: 5%, minimum confidence: 5%, maximum umber of conditions: 5, minimum use: 10. Second, we fused those rules based on algorithm discussed above and applied them on test set, performing a 10-fold cross validation. In order to test the robustness of the results, we repeated the experiment for all possible random combinations of training (4 dyads) and test (2 dyads) sets for friends and strangers, and performed a correlated samples t-test to test whether our approach results in lower mean squared error compared to a simple linear regression model that treats each of the verbal and nonverbal modalities as independent features to predict the absolute value of rapport. Evaluation for performance metrics in this basic linear regression approach was done using the supplied test set of randomly chosen 2 dyads for each experimental run. In addition, we also calculated effect size via Cohen'sd d ($2t/\sqrt{df}$), where t is the value from the t-test and df refers to the degrees of freedom. Results in Table 1 suggest that the Titarl-based regression method has a significantly lower mean square error than the naive baseline linear regression method.The high effect size in both strangers (d=-0.62) and friends (d=-0.42) further prove the substantial improvement on accuracy of assessing rapport by Titarl-based regression comparing to simple linear regression.

These results have been integrated into a real-time end-to-end socially aware dialog system (SARA),[1] described in [26]. SARA is capable of automatically detecting conversational strategies based on verbal, nonverbal, and acoustic features in the user's input [39], relying on the conversational strategies detected in order to accurately estimate rapport between the interlocutors, reasoning about what conversational strategy to respond with as the next turn, and generating those appropriate responses in the service of more effectively carrying out her task duties. To our knowledge, SARA is the first socially-aware dialog system that relies on visual, verbal, and vocal cues to detect user social and task intent, and generates behaviors in those same channels to achieve her social and task goals.

[1] sociallyawarerobotassistant.net

6.2 Implications for Understanding Human Behavior

One of the important behavior patterns that plays out differently across friends and strangers, and whose interactions can lead to either high or low rapport, is smiling in combination with social norm violations and speech overlap. A violation of social norms without a smile is always followed by low rapport. On the other hand, a social norm violation accompanied by a smile is followed by high rapport when followed by overlap and performed among friends. Meanwhile, violating social norms while smiling leads to low rapport when followed by overlap if performed among strangers [See FH1, FH3, FH5, FL1, FL2, SL3]. What we may be seeing here is what [11] described as embarrassment following violations of "ceremonial rules" (social norms or conventional behavior), which is less often seen among family and friends than among strangers and new acquaintances. Similarly, [22] emphasized that the smile is a kind of hedge, signaling awareness of a social norm being violated and serving to provoke forgiveness from the interlocutor. Overlap in this context may be an index of the high coordination that characterizes conversation among friends whereby simultaneous speech indicates comfort, or that same overlap may indicate the lack of coordination that characterizes strangers who have not yet entrained to one another's speech patterns [5]. Our findings provide further empirical support for this body of prior work.

Another important contingent pattern of behaviors discussed here is the interaction between smile and backchannels [See SH2, FL3]. In general a backchannel + smile was indicative of high rapport, perhaps because the smile + backchannel indicated that the listener was inviting a continuation of the speaker's turn, but also indicating his/her appreciation of the interlocutor's speech [3].

We also discover the interaction between smile, the conversational strategy of self-disclosure and overlaps [See SH1]. Smiles invite self-disclosure, after which an overlap demonstrates responsiveness of the interlocutor. [25] have shown that partner responsiveness is a significant component of the intimacy process that benefits rapport. Finally we described how the presence of overlaps with a nonverbal behavior or conversational strategy often signals high rapport in friends but low rapport in strangers [See SH3, FL3, SL2, SL3]. Prior work has found that friends are more likely to interrupt than strangers, and the interruptions are less likely to be seen as disruptive or conflictual [5].

6.3 Implications for Social Agent Design

Rules such as those presented above can play a fundamental role in building socially-aware agents that adapt to the rapport level felt by their users in ways that previous work has not addressed. For example, [12] extracted a set of hand-crafted rules based on social science literature to build a rapport agent. Such rules not only need expert knowledge to craft, but may also be hard to scale up and to transfer to different domains. In our current work, we alleviate this problem by automatically extracting behavioral rules that signal high or low rapport, learning on verbal and nonverbal annotations of a particular corpus, but employing only the annotations of conversational strategies that did not concern the content domain of the corpus. This also represents an advance on work by [19] that improved rapport through nonverbal and para-verbal channels, but did not take linguistic information or temporal co-occurrence across modalities into account. We included linguistic information in our rules and In other work we have shown that the linguisic information (conversational strategies) that formed an essential part of the temporal rules presented here can be automatically recognized [39]. Similarly, [9]'s gaze-reactive pedagogical agent diagnoses disengagement or boredom by the use of eye trackers. However, only taking eye gaze into account forfeits the potential synergistic effect of interaction across modalities.

As noted above, while our current work focused on developing an interpretable and explanatory model of temporal behaviors to serve as a building block for our rapport-aligned peer-tutoring system (RAPT), the framework can be applied for prediction of other social phenomena of interest in virtual agent systems (such as trust and intimacy), in domains as diverse as survey interviewing, sales, and health.

7 Conclusion

In this work, we utilized a temporal association rule framework for automatic discovery of co-occurring and contingent behavior patterns that precede high and low interpersonal rapport in dyads of friends and strangers. Our work provides insights for better understanding of dyadic multimodal behavior sequences and their relationship with rapport which, in turn, moves us forward towards the implementation of socially-aware agents of all kinds - including "socially-skilled" virtual peer tutors that can assess the state of a relationship with a student, sigh in frustrated solidarity about a learning task at hand, and know how to respond to maximize learning in the peer tutoring context.

Among the patterns our rules discovered were the interaction of smiles and backchannels in signaling mutual attention and appreciation, and the pattern of self-disclosure, followed or preceded by smiles and speech overlap, as an indicator of high rapport. We found smiles to be one way in which interlocutors appear to mitigate the face-threat of social norm violations such as insults. However, our experiments discovered that while the presence of speech overlaps with smiles and social norm violations in friends signals high rapport, the presence of speech overlaps with social norm violations in strangers signals low rapport. In addition, for prediction of rapport, we observed the benefits (significantly lower mean square prediction error) of constructing predictor variables that work on fine-grained representation of social behaviors, explicitly model the temporal relations among them and encode ordering as well as timing, over using simple aggregated behavioral descriptors in a baseline linear regression model that are crudely informative.

Limitations of the current work include our focus on rapport states; in future work we will also want to find the temporal association rules that lead to a delta in rapport. In addition, while the current work discovers those behaviors that directly precede a rapport state, we have not yet verified that the link is causal. In service to that goal, our current work has implemented the temporal association rules as a real-time module, and has integrated them into a working virtual agent system. Our future work will use this system to evaluate the causal nature of these rules, and their effect on human - virtual agent interaction.

References

1. Jens Allwood and Loredana Cerrato. A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, pages 7–22. Copenhagen, 2003.
2. Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

3. Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. A listening agent exhibiting variable behaviour. pages 262–269, 2008.
4. Joseph N Cappella. On defining conversational coordination and rapport. *Psychological Inquiry*, 1(4):303–305, 1990.
5. Justine Cassell, Alastair J Gill, and Paul A Tepper. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*, pages 41–50. ACL, 2007.
6. Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. pages 152–166, 2015.
7. Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *International Conference on Intelligent Virtual Agents*, pages 120–133. Springer, 2014.
8. Chih-Yueh Chou and Tak-Wai Chan. Reciprocal tutoring: Design with cognitive load sharing. *International Journal of Artificial Intelligence in Education*, pages 1–24, 2015.
9. Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398, 2012.
10. Donna T Fujimoto. Listener responses in interaction: A case for abandoning the term, backchannel. 2009.
11. Erving Goffman. *Interaction ritual: Essays in face to face behavior*. AldineTransaction, 2005.
12. Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. Virtual rapport. pages 14–27, 2006.
13. Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer, 2007.
14. Agustín Gravano and Julia Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *INTERSPEECH*, pages 1019–1022, 2009.
15. Mathieu Guillame-Bert and James L. Crowley. Learning temporal association rules on symbolic time sequences. pages 159–174, 2012.
16. Mathieu Guillame-Bert and Artur Dubrawski. Learning temporal rules to forecast events in multivariate time sequences.
17. Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. Searching for prototypical facial feedback signals. In *Intelligent Virtual Agents*, pages 147–153. Springer, 2007.
18. Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *Intelligent Virtual Agents*, pages 68–79. Springer, 2011.
19. Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. pages 68–79, 2011.
20. David W Johnson. Student-student interaction: The neglected variable in education. *Educational researcher*, 10(1):5–10, 1981.
21. Sin-Hwa Kang, Jonathan Gratch, Candy Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th ICAAMS-Volume 1*, pages 63–70, 2012.
22. Dacher Keltner and Brenda N Buswell. Embarrassment: its distinct form and appeasement functions. *Psychological bulletin*, 122(3):250, 1997.
23. Adena M Klem and James P Connell. Relationships matter: Linking teacher support to student engagement and achievement. *Journal of school health*, 74(7):262–273, 2004.

24. Karel Kreijns, Paul A Kirschner, and Wim Jochems. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in human behavior*, 19(3):335–353, 2003.
25. Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R Pietromonaco. Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of personality and social psychology*, 74(5):1238, 1998.
26. Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar J. Romero, Sushma Akoju, and Justine Cassell. Socially-aware animated intelligent personal assistant agent. In *17th Annual SIGdial Meeting on Discourse and Dialogue*, 2016.
27. Yukiko I Nakano, Sakiko Nihonyanagi, Yutaka Takase, Yuki Hayashi, and Shogo Okada. Predicting participation styles using co-occurrence patterns of nonverbal behaviors in collaborative learning. pages 91–98, 2015.
28. Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. Rudeness and rapport: Insults and learning gains in peer tutoring. In *Intelligent Tutoring Systems*, pages 11–21. Springer, 2012.
29. Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. pages 23–30, 2015.
30. Nikol Rummel, Erin Walker, and Vincent Aleven. Different futures of adaptive collaborative learning support. *International Journal of Artificial Intelligence in Education*, 26(2):784–795, 2016.
31. Anna M Sharpley, James W Irvine, and Christopher F Sharpley. An examination of the effectiveness of a cross-age tutoring program in mathematics for elementary school children. *American Educational Research Journal*, 20(1):103–111, 1983.
32. Tanmay Sinha and Justine Cassell. Fine-grained analyses of interpersonal processes and their effect on learning. In *Artificial Intelligence in Education*, pages 781–785. Springer, 2015.
33. Tanmay Sinha and Justine Cassell. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 2015 Workshop on Modeling Interpersonal Synchrony, 17th ACM International Conference on Multimodal Interaction*. ACM, 2015.
34. Tanmay Sinha, Ran Zhao, and Justine Cassell. Exploring socio-cognitive effects of conversational strategy congruence in peer tutoring. In *Proceedings of the 2015 Workshop on Modeling Interpersonal Synchrony, 17th ACM International Conference on Multimodal Interaction. ACM*, 2015.
35. Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.
36. Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. Multimodal public speaking performance assessment. pages 43–50, 2015.
37. Zhou Yu, David Gerritsen, Amy Ogan, Alan W Black, and Justine Cassell. Automatic prediction of friendship via multi-model dyadic features. In *14th Annual SIGdial Meeting on Discourse and Dialogue, Metz, France*, 2013.
38. Ran Zhao, Alexandros Papangelis, and Justine Cassell. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*, pages 514–527. Springer, 2014.
39. Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *17th Annual SIGDIAL Meeting on Discourse and Dialogue*, 2016.