Socially-Aware Dialog System

Ran Zhao

May 2018

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Alexander I.Rudnicky, CMU (Chair)
Alan W.Black, CMU (Co-Chair)
Louis-Philippe Morency, CMU
Amanda Stent, Bloomberg

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosoph
In Language and Information Technologies

Keywords: Socially-Aware, multimodal machine learning, temporal sequence learning, cognitive architecture, neural dialog model, socio-pychology theoretical framework, rapport,conversational strategy, spoken dialog systems,discourse analysis, social reasoning,natural language generation

Abstract

In the past two decades, spoken dialog systems, such as those commonly found in cellphones and other interactive devices, have emerged as a key factor in human-computer interaction. For instance, Apple's Siri, Microsoft's Cortana, and Amazon's Alexa help human users complete tasks more efficiently. However, research in this area has yet to produce dialog systems that build interpersonal closeness over the course of a conversation along with carrying out the task. This project attempts to address that shortcoming. Specifically, research in computational linguistics (Bickmore and Cassell, 1999) has shown that people pursue multiple conversational goals in dialog, which include those that fulfill propositional functions to contribute information to the dialog; those that fulfill interactional functions to manage conversational turn-taking; and those that fulfill interpersonal functions to manage the relationship between interlocutors. Although spoken dialog systems have greatly advanced in modeling the propositional and, to a lesser extent, interactional functions of human communication, these systems fall short in replicating the interpersonal functions of conversation. We propose that this interpersonal deficiency is due to a lack of models of interpersonal goals and strategies in human communication.

As dialog systems become more common and are used more frequently as interfaces to search and other computing tasks, propositional content and interactional content will not suffice. In this thesis, therefore, we address these challenges by proposing a socially-aware intelligent framework that exploits a path to systematically generate dialogs that fulfill interpersonal functions.

In (Zhao et al., 2014a), we clarify that a socially-aware intelligent framework can explain how humans in dyadic interactions build, maintain, and tear down social bonds through specific conversational strategies that fulfill specific social goals and that are instantiated in particular verbal and nonverbal behaviors. In order to operationalize this framework, we argue that four capabilities are needed to achieve a socially-aware intelligent system. The system must (1) automatically infer human users' social intention by recognizing their social conversational strategies, (2) accurately estimate social dynamics by observing dyadic interactions, (3) reason through appropriate conversational strategies while accounting for both the task goal and social goal, and (4) realize surface-level utterances that blend task and social conversation. Our socially-aware dialog system focuses on blended conversations that mix a goal-oriented task with social chat. As a proof of concept, we have induced a modular-based socially-aware personal assistant for a conference.

Finally, we propose to apply our socially-aware intelligent framework in negotiation dialog. We formulate a two-phase method to blend negotiation utterance with social conversation. Our pilot study shows that the system can facilitate negotiation by building a social bond with a human user.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Spoken dialogue systems have been widely deployed as an interactive interface in various devices. Three prevailing types of spoken dialog systems/bots have been developed: a task completion bot that helps users book movie tickets, make restaurant reservations, and so on; an information retrieval bot that supports an interactive Q&A system over knowledge base; and a social chatbot that aims to engage users in open-domain dialog. Our goal is to develop a dialogue system that combines the functions of the task completion bot and social chatbot and that models intrinsically-interdependent social phenomena with humans in ways that improve task performance.

Our primary interest in the social phenomena is interpersonal rapport, since rapport has been identified as an important function of human interaction. But, to our knowledge, no model exists of building and maintaining rapport between humans and conversational agents over the course of a conversation that operates at the level of the dyad. **In this thesis work, we mainly focus on designing a socially-aware dialog system that builds interpersonal closeness (rapport) over the course of a conversation while completing the task by understanding human behavior and generating an appropriate response.** In chapter 2, we leverage existing literature and a corpus of peer tutoring data (Yu et al., 2013b) to develop a computational model of rapport, which serves as the foundation to our socially-aware theoretical framework (Zhao et al., 2014a). To operationalize this framework, in chapter 3,we review the architecture of our modular-based socially-aware dialog system, a personal assistant that helps conference attendees achieve their goals, including introducing them to other attendees and informing them about sessions that fit their interests (Matsuyama et al., 2016). This chapter also provides an overview of our proposed modules and demonstrates the effectiveness of rapport-building conversational strategies on improving task performance. Before describing each module we developed, in chapter 4.1, we conduct a follow-up experimental study on the proposed theoretical framework to quantitatively investigate the pattern of conversational strategy usage in human dialog as well as initially validate its effectiveness on both social (rapport) and task (learning) (Sinha et al., 2015). Toward this end, we examine similarity in use and timing of relationship-oriented communicative strategies such as self-disclosure, reference to shared experience, and praise during a reciprocal peer tutoring interaction. Then, we introduce each of our designed modules for a socially-aware dialog system. In terms of detection, in chapter 4.2, we explore the content level of the utterance and leverage quantitative methods to automatically recognize different conversational strategies (Zhao et al., 2016a). Since our system can automatically recognize different conversational strategies, in chapter 4.3, we step toward data-driven discovery of the temporally co-occurring and contingent behavioral patterns that signal high and low interpersonal rapport (Zhao et al., 2016b). We

validate the discovered behavioral patterns by predicting rapport against our ground truth (30-second thin slice rapport) via a forecasting model involving two-step fusion of learned temporal associated rules. In terms of reasoning, in chapter 5, our system first carries out the classic AI task: reasoning to determine how best to fulfill the user's goals (Romero et al., 2017a). Then, it carries out a new type of reasoning what we call social reasoning to determine how to converse with the user, including spoken language and body language, to best accomplish both the task (e.g., information-seeking, teaching, and calendar management) and social goals (e.g,. managing rapport).

Finally, we explore how our socially-aware intelligent framework performs in negotiation dialog, where interpersonal dynamics are surprisingly significant. In chapter 6, we propose a two-phase method that leverages the power of a neural dialog model and reinforcement learning algorithm to blend social and task conversation. This method aims to optimize the negotiation outcome and the social outcome. Our pilot study demonstrates promise in facilitating interpersonal rapport while improving negotiation performance.

1.1 Thesis Statement

In this thesis work, we mainly focus on designing a socially-aware dialog system that builds interpersonal closeness (rapport) over the course of a conversation along with carrying out the task through understanding human behaviors and generating appropriate response.

1.2 Thesis Contributions

- Computational model of rapport:The computational model is the first to explain how humans in dyadic interactions build, maintain, and destroy rapport through the use of specific conversational strategies that function to fulfill specific social goals, and that are instantiated in particular verbal and nonverbal behaviors (Zhao et al., 2014a; Sinha et al., 2015).
- Techniques for automatic recognizing Conversational Strategy: We have implemented a conversational strategy classifier to automatically recognize the user's conversational strategies particular ways of talking, that contribute to building, maintaining or sometimes destroying a budding relationship. These include self-disclosure (SD), elicit self-disclosure (QE), reference to shared experience (RSD), praise (PR), violation of social norms (VSN), and back-channel (BC). By including rich contextual features drawn from verbal, visual and vocal modalities of the speaker and interlocutor in the current and previous turns, we can successfully recognize these dialog phenomena with an accuracy of over 80% and with a kappa of over 60% (Zhao et al., 2016a,d).
- Techniques for automatic estimating rapport level: We use the framework of temporal association rule learning to perform a fine-grained investigation into how sequences of interlocutor behaviors signal high and low interpersonal rapport. The behaviors analyzed include visual behaviors such as eye gaze and smiles, and verbal conversational strategies, such as self-disclosure, shared experience, social norm violation, praise and back-channels. We developed a forecasting model involving two-step fusion of learned temporal associated rules. The estimation of rapport comprises two steps: in the first step, the intuition is 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 corresponding to each of the binary classifiers for the rapport states, in order to predict the absolute continuous value of rapport (via linear regression) model. Ground truth for the rapport state was obtained by having naive annotators rate the rapport between two interactants in the teen peer-tutoring corpus for every 30 second slice of an hour long interaction.Our framework performs significantly better than a baseline linear regression method that does not encode temporal information among behavioral features (Zhao et al., 2016b).

- Techniques for enabling a dialog system to plan conversational style and strategy towards achieving both social and task goal: I was collaboration with Oscar J. Romero to design a social reasoner that decides the appropriate conversational style and strategy with which the dialogue system describes the information the user desires so as to boost the strength of the relationship between the user and system (rapport). I contribute to the idea of using spreading activation model and the design of pre-conditions and post-conditiosn of each conversational strategy. Oscar J. Romero implemented the spreading activation model- a behavior network consisting of activation rules that govern which conversation strategy the system should adopt next. We conduct several experiments to validate the effectiveness of social reasoner. Our Social Reasoner is inspired both by analysis of empirical data of friends and stranger dyads engaged in a task, and by prior literature in fields as diverse as reasoning processes in cognitive and social psychology, decision-making, sociolinguistics and conversational analysis. Our experiments demonstrated that, when using the Social Reasoner in a Dialogue System, the rapport level between the user and system increases in more than 35% in comparison with those cases where no Social Reasoner is used (Romero et al., 2017a).
- Techniques for blending task conversation with conversational strategies: we propose a two phase method that interweaves task utterance with conversational strategies to engage human users in negotiation. Our algorithm operates sequentially in a reasoning-and-generation loop: In the task phase, we leverage off-the-shelf end-to-end dialogue models for negotiation to build a dialogue manager which decides the next system's task intention. Then, during the social phase, we employ a theory-driven, template-based natural language generator to realize the task intention as a genre of social conversational strategy. Subsequently, we propose to leverage the power of reinforcement learning to find out the best conversational strategy policy to optimize the social outcome in negotiation.

Chapter 2

Theoretical Framework[1]

2.1 Introduction

Human are deeply interdependent with each other in society.Modeling such social process between human and system is complicated. As the first step, we start to learn empirical findings of human-to-human interaction through reviewing previous socio-pyschological literatures about intrinsically interdependent social phenomena, such as rapport, trust or interpersonal closeness. From there, we develop the first dyadic computational model of rapport that able to explain how humans in dyadic interactions build, maintain, and destroy rapport through the use of specific conversational strategies that function to fulfill specific social goals, and that are instantiated in particular verbal and nonverbal behaviors. This theoretical framework will serve as the guidelines for us to further design and implement our socially-aware dialog system.

Rapport, a feeling of connection and closeness with another, feels good, but it also has powerful effects on performance in variety of domains, including negotiation (Drolet and Morris, 2000), child care (Burns, 1984), counselling (Kang et al., 2012) and education (Bernieri and Rosenthal, 1991).As agents increasingly take cover tasks such as those described above, we maintain that it is important to evoke a feeling of rapport in people interacting with those agents so as to improve their task collaboration - and recognize rapport in people interacting with agents so as to know when the system has been successful, it turns out, however, that what constitutes rapport-evoking and rapport signaling behavior varies widely. While prior work [e.g. (Karacora et al., 2012)] has confirmed that some rapport-signaling behavior such as attentiveness is capable of enhancing task performance, there has existed no rigorous models of the mechanism underlying the relationship between social and cognitive functioning in tasks such as these (Kreijns et al., 2003), nor do there exist computational models of interpersonal closeness that can tell us how rapport signaling behavior should change over the course of a long-term collaboration between a human and an agent. In our paper (Zhao et al., 2014a), we claimed that one obstacle to models of this sort is the fact that, as (Bernieri and Gillis, 2001) has written,"rapport is a social construct that must be defined at the level of a dyad or larger group." Dyadic processes of this sort have traditionally posed challenges to modeling since, as (Bickmore et al., 2005) have described, a change in the state of one partner will produce a change in the state of one partner will produce a change in the

[1]This section incorporates text from (Zhao et al., 2014a) which describes a collaboration between Ran Zhao, Alexandros Papangelis and Justine Cassell. My contribution to this work were designing the theoretical computational model of rapport management based on social science literatures. Also, I involved in writing and modifying the publication version of the paper.

state of the other. We believe that prior attempts have not sufficiently distinguished between the social functions that lead to rapport, the conversational and behavioral strategies that play a role in those social functions, and the observable phenomena that make up those strategies. Rapport is sometimes experienced on a first meeting but most often it must be built and maintained - or it will destroyed. Drawing these distinctions has also allowed us to move toward an implementable computational architecture,described in the following sub-section of this chapter, that takes into account both participants' cognition,intentions, action and beliefs, and their interplay, within one person and across the dyad.

In this work, we rely on a rich background of literature across the social sciences, as well as on data (Yu et al., 2013a) from our own research into peer tutoring between dyads of friends and of strangers across several months. These data have been annotated for verbal and nonverbal behaviors, as well as for relevant conversational strategies.

2.2 Related Work

A number of prior papers have addressed the issue of rapport, or related notions such as trust, friendship, and intimacy, between people and agents. An early paper (Cassell et al., 1999) used prior work in sociolinguistics and social psychology to develop a computational model of trust, and a computational architecture to establish trust between a person and virtual agent. The system, however, did no assessment of the user's level of trust, and only built trust through verbal behavior - primarily small talk. While successful in building trust - particularly with extroverts - a subsequent paper (Bickmore and Cassell, 2005) demonstrated the need for incorporating nonverbal behavior into the model. Since then, Bickmore and his colleagues have gone on to develop a model that describes strategies for an agent to build a relationship with a user over time.

Until recently, much like the early work described above, these systems have primarily engaged in a set of predetermined conversational strategies without associated updates in underlying goals or representations of the user or the user-system dyad (Vardoulakis et al., 2012). While not always successful at promoting rapport, these strategies have had positive effects on the non-dyadic construct of engagement (Bickmore et al., 2011). (Bickmore and Schulman, 2012) has relied on accommodation theory to design conversational strategies intended to generate discourse that matches a user's level of intimacy, and to increase intimacy. The prior goal was met but not the latter, perhaps because, as the authors themselves indicate, the model of intimacy was quite simplistic, without the kinds of goals, subgoals, and conversational strategies laid out here. On the other hand, accommodation theory provided a successful means for assessing the user's level of intimacy, which bears keeping in mind for future work. Following on from this work, (Sidner, 2012) developed a planning algorithm that keeps track of the intimacy level of the user, and produces session plans that target both relational and task goals. The activity planning approach seems promising, however the session plans appear to be made up of activities that are appropriate at a particular level of closeness rather than activities that have been shown specifically to *increase* closeness. Our approach, whereby conversational strategies target sub-goals that specifically manage rapport, might be more successful at moving the system and user further along on the relational continuum.

An alternative approach is represented by the work of Gratch and colleagues (Gratch et al., 2006; Huang et al., 2011), who target immediate rapport in the service of implementing a sensitive listener. In this work, the level of goals and conversational strategies are avoided, and

instead the agent attemps to elicit the experience of rapport by working at the level of observable phenomena - coordinating its nonverbal behavior to the human user. Rather than treating rapport as a dyadic or interpersonal construct, they address it similarly to other display functions and perhaps not surprisingly, as with other engaging displays, they have found increased user engagement. Most recently they have extended this approach to the analysis of the nonverbal behaviors that accompany intimate self-disclosure (Kang et al., 2012). However, by not taking into account the relative roles of the two interlocutors, and the nature of their relationship, they have ignored the significant difference in conversational strategies between interlocutors with different levels of power in the relationship.

In contrast to the prior work described here, our work distinguishes between the dyad's goals (overarching goals such as "create rapport" or sub-goals such as "index commonality"), their conversational strategies (such as "violate sociocultural norms through rude talk" or "initiate self-disclosure") and the observable verbal and nonverbal phenomena that instantiate those phenomena (such as mutual eye gaze, embarassed laughter, or insults). This tri-partite distinction allows us to generate the same behaviors (insults, for example) in different contexts (early or late in the relationship) to achieve different goals (destroy rapport or enhance it). The unit of analysis of the computational model we present is the dyad, with system state updates impacting the model of the user, and of the user's model of the system, and particular weight placed on intrinsically dyadic constructs such as reciprocity.

2.3 Social Science Literatures for Rapport Management

(Tickle-Degnen and Rosenthal, 1990)'s work on the changing nonverbal expression of rapport over the course of a relationship has had significant impact on the development of virtual agents. They provide an actionable starting point by outlining the experience of rapport as a dynamic structure of three interrelating behavioral components: positivity, mutual attentiveness and coordination. Behavioral positivity generates a feeling of friendliness between interactants; mutual attentiveness leads to an experience of connectedness; and behavioral coordination evokes a sense of "being in synch". The work posits that the relative weights of those components change over the course of a relationship; the importance of mutual attentiveness remains constant, while the importance of positivity decreases and that of coordination increases.

While (Tickle-Degnen and Rosenthal, 1990)'s work is predicated on a dual level of analysis - what they call "molecular" and "molar," researchers in virtual agents have relied more on the molecular level, meaning that they have translated (Tickle-Degnen and Rosenthal, 1990)'s components directly into observable behavioral expression or action. (Tickle-Degnen and Rosenthal, 1990), however, propose that it is the molar level that is more predictive - that is, that theory should attend to the conversational strategies and goals of communication that interactants use to be positive, be attentive and to coordinate. In fact, they suggest that "initial encounters are rigidly circumscribed by culturally acceptable and stereotypical behavior" while, after some time, "rather than following more culturally-defined communication conventions, they would develop their own conventions and show more diversity in the ways they communicate thoughts to one another." This aspect of their work has largely been ignored in subsequent computational approaches to rapport. In the development of agent models and an architecture to realize them, however, this leaves us less than well-informed about what the agents should do. How do we determine what is meant by "stereo-

typical behavior" or "more diversity in the ways they communicate"? How should we represent the goals of two interactants and conversational strategies to fulfill the goals? In the current work, then, we discuss a broad range of literature that allows us to understand the kinds of strategies that interactants use in rapport management, and the kinds of goals and functionality those interactants intend. As we do so, we pay particular attention to the dyadic nature of these constructs, and how they change over the course of a relationship. Our review focuses on 3 top-level goals that make up rapport - **face management**, **mutual attentiveness**, and **coordination** - and some of the subgoals that achieve those top-level goals - such as *becoming predictable*, *appreciating the other's true self*, and *enhancing the other's face*. We also describe many of the conversational strategies that achieve those goals - `initiating mutual self-disclosure`, `adhering to behavioral expectations or norms`, and so forth.

(Spencer-Oatey, 2005) offers an alternative approach to (Tickle-Degnen and Rosenthal, 1990)'s to conceptualizing the strategies and behaviors that contribute to rapport, and we find it more complete and more convincing for our purposes. She points out that rapport management comprises the task of increasing rapport, but also maintaining, and destroying it. In her perspective, each of these tasks requires management of face which, in turn, relies on behavioral expectations, and interactional goals. Our data support the tremendous importance of face, as the teens alternately praise and insult one another, all the while hedging their own positive performance on the algebra task in order to highlight the performance of the other. The data in (Yu et al., 2013b) also contain numerous examples of mutual attentiveness and coordination as putative input into rapport management, but we found it difficult to code positivity independently of its role in face. Our formulation below, therefore, posits a tripartite approach to rapport management, comprising mutual attentitiveness, coordination, and face management.

Face management: (Brown and Levinson, 1978) define positive face as, roughly, a desire by each of us to be approved of. They posit that politeness functions to avoid challenging that desire, as well as to boost the other's sense of being approved, while *face-threatening acts* (FTA) challenge face. (Spencer-Oatey, 2008a), however, points out that this definition ignores the interpersonal nature of face, and she defines "identity face" as the desire to be recognized for one's positive social identity, as well as one's individual positive traits. In this context, FTAs can challenge one's sense of self or one's identity in the social world. On the flip-side, *face-boosting acts* can create increased self-esteem in the individual, and increased interpersonal cohesiveness - or rapport - in the dyad. Of course (Spencer-Oatey, 2005) points out that what constitutes politeness, other face-boosting acts, and FTAs, is not fixed, and is largely a subjective judgement about the social appropriateness of verbal and non-verbal behaviors. She attributes these judgments about social appropriateness to our "sociality rights and obligations" - how we feel entitled to be treated based on the behaviors we expect from others – which in turn derive from sociocultural norms, including the relative power and status of the two members of the dyad, and interactional principles. Fulfilling these rights and obligations induces a feeling of being approved and, in turn, increases rapport.

What, however, are these sociocultural norms and interactional principles? A key aspect of the theory laid out here is that *behavioral expectations* (the instantiation of "sociality rights and obligations") are allied with sociocultural norms early in a relationship, and become more interpersonally determined as the relationship proceeds. Thus, the stranger dyads in our data spend a fair amount of time agreeing with one another when they first meet, in ways that fit upper

middle class politeness norms (when asked what he wants to be when he grows up, one teen responds "I kind of want to be a chef" to which the other politely responds "I'd think about that too"). Friends, on the other hand, are less likely to demonstrate polite responses (one teen asks the other "wait why do you have to keep your hat on" to which the other responds "it's [his neck] not supposed to be in the sun" and receives in reply "yeah it's really swollen and ugly"). In both cases while the behavioral expectations have changed (politeness has been replaced by teasing), the fact of meeting them continues to be rapport-increasing.

How does one learn enough about the other to adapt behavioral expectations? **Mutual attentiveness** is an important part of the answer, as (Tickle-Degnen and Rosenthal, 1990) have described. Mutual attentiveness may be fulfilled by providing information about oneself through small talk (Cassell and Bickmore, 2003) and self-disclosure (Moon, 2000). Social penetration theory (Taylor and Altman, 1987) describes the ways in which, as a relationship deepens, the breadth and depth of the topics disclosed become wider and deeper, helping the interlocutor to gain common ground as a basis for an interpersonally-specific set of behavioral expectations. Self-disclosure, however, plays another role in rapport-building, as when successful it is reciprocal (Derlega et al., 1993) – self-disclosure in our data is most often met with reciprocal self-disclosure at a similar level of intimacy. This kind of mutual responsiveness signals receptivity and appreciation of another's self-disclosure (Derlega et al., 1993) and the very process enhances **coordination** among the participants (much as we argued is the case for small talk (Cassell and Bickmore, 2003)), likewise increasing a sense of rapport. The goal of coordination as a path to rapport is also met by verbal and nonverbal synchrony (Zanna, 1999), and this is common in our own data.

In addition, while self-disclosure is not always negative, it may be, and this is a way to challenge one's own face, and thereby boost the face of the other. For that reason it is common in rapport management. In our own data, for example, strangers quickly began to share superficial negative facts about themselves, such as their presumed poor performance on the algebra pre-test at the beginning of the session. When met with a self-disclosing utterance at the same level of intimacy and with the same negative valence ("oh my gosh I could not answer like half of those"), the interlocutors increased mutual gaze and smiling, and proceeded to more intimate topics, such as their poor performance at keeping their pets alive. In fact, (Bronstein et al., 2012a) found that in a negotiation setting not reciprocating negative self-disclosure led to decreased feelings of rapport. (Treger et al., 2013) point out the role of humor in rapport; it is a particularly interesting rapport management strategy as it too follows behavior expectations, whereby generally-accepted humor is successful early in the relationship, and humor that violates sociocultural norms may be successful as a strategy to increase liking and rapport only later in the relationship. In our data from teenagers, this rule is only sometimes observed, and the effect of humor that violates behavior expectations is swift and negative.

Self-disclosure, then, serves multiple goals in rapport management. Yet another is to reveal aspects of one's "true self" as a way of indicating one's openness to being truly seen by the other, and hence one's availability for rapport. According to (Rogers, 1966), the "true-self" is composed of important aspects of one's identity that are not always validated in one's daily life. People are highly motivated to make these important aspects of identity a 'social reality' - to have these attributes acknowledged by others so that they become authentic features of their "self-concept" (Bargh et al., 2002). This explains *why* interlocutors engage in self-disclosure -

perhaps even why rapport is sought in interactions with strangers.

Based on the literature surveyed above, it is clear that mutual attentiveness to, and learning about and adhering to, the behavioral expectations of one's interlocutor is helpful in building rapport. Initially, when interactants are strangers, without any knowledge of their interlocutor's behavioral expectations, they adhere to a socioculturally-ratified model (general expectations established as appropriate in their cultural and social milieu). This may include behaving politely and in accordance with their relative social roles. As the relationship proceeds, interlocutors increasingly rely on knowledge of one another's expectations, thereby adhering to a shared and increasingly interpersonally-specific set of sociality rights and obligations, where more general norms may be purposely violated in order to accommodate each other's behavioral expectations.

Why, however, might two interactants violate sociocultural norms when others around them are adhering to those norms? (Baumeister and Leary, 1995) suggests that people have an unconscious motivation to affiliate themselves to a group, which drives them to participate in social activities and search for long-term relationships. The fact of violating sociocultural norms may in fact reinforce the sense that the two belong in the same social group and this may enhance their unified self-image (Tajfel and Turner, 1979) through reinforcing the sense of in-group connectedness through a comparison with other individuals who don't know these specific rules of behavior. This is supported by our own findings on peer tutoring, whereby rudeness predicts learning gain (Ogan et al., 2012a). We know that rapport between teacher and student increases learning. When tutor and tutee are strangers, their behavior complies with sociocultural norms. Impoliteness may reduce the learning gain in strangers by challenging rapport through violating those sociocultural behavioral expectations. When tutor and tutee are friends, however, they have knowledge of one another's behavioral expectations and are thus able to follow interpersonal norms and sacrifice sociocultural norms. Rudeness, may be a part of the interpersonal norms. It may also be a way to cement the sense that the two are part of a unified group, and different from those around them. The topics they are rude about may also serve to index commonalities between the two, as referring to shared experience also differentiates in-group from out-group individuals.

2.4 Study Context

Figure 2.1: Camera View 1 and Camera View 2

Reciprocal peer tutoring data was collected from 12 American English-speaking dyads (6 friends and 6 strangers; 6 boys and 6 girls), with a mean age of 13 years, who interacted for 5 hourly sessions over as many weeks (a total of 60 sessions, and 5400 minutes of data), tutoring one another in algebra (Yu et al., 2013b). 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. The setting is shown in Figure 2.1 We chose peer tutoring as it is a domain in which rapport has been shown to have a positive effect on student learning see (Ogan et al., 2012a).

Prior work demonstrates that peer tutoring is an effective paradigm that results in student learning (Sharpley et al., 1983), making this an effective context to study dyadic interaction with a concrete task outcome. Our student-student data, in addition, demonstrates that a tremendous amount of rapport-building takes place during the task of reciprocal tutoring (Sinha and Cassell, 2015b).

2.5 Theoretical Computational Model of Rapport Management

The literature review above, while not allowing each component sub-goal or strategy the space it deserves, provides a sense of the complexity, but also of the mundane nature of rapport management between people. We wish to be seen and known the way we truly are, and we want the way we are to be approved; we desire affiliation with a social group; we are more comfortable when the behavior of our interlocutors matches our expectations; we wish for the success of our interpersonal and our task goals. These common sense and everyday goals work together to lead us to desire rapport, and to build it, even with strangers, and to put effort into maintaining it with friends and acquaintances. In order to represent these goals and desires in a computational

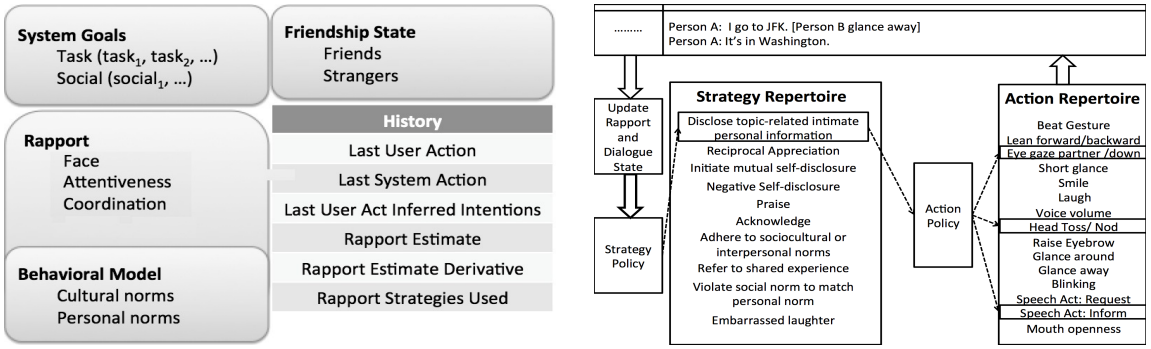


Figure 2.2: Dyadic state (left) and Strategy/Action repertoire (right)

model, (Zhao et al., 2014a) emphasize the fact that while rapport is dyadic, it nevertheless depends on the cognition, actions, beliefs and intentions of each interlocutor, and on the perception by each interlocutor of these aspects of the mind of the other interlocutor. In the computational model, therefore, we represent the state of each participant, and of that participant's perception of the state of the interlocutor, which enables us to reason about the cognition and rapport orientation (enhancement, maintenance, destruction) of the dyad, based on observable behaviors. More specifically, Figure 2.2(left) presents the *dyadic state*, which may be updated after each user's turn or incrementally. Figure 2.2(right) displays how a user and system state leads to a choice of *Strategy* and then of *Action* (although the latter is beyond the scope of the current chapter). Of course, in order to allow rapport state monitoring and management, we need to detect the goals and conversational strategies of the interlocutors on the basis of the behaviors we observe them engaging in, and we need to assess their contribution to each rapport orientation. Below, for rapport enhancement, maintenance and destruction we list, from the perspective of the agent

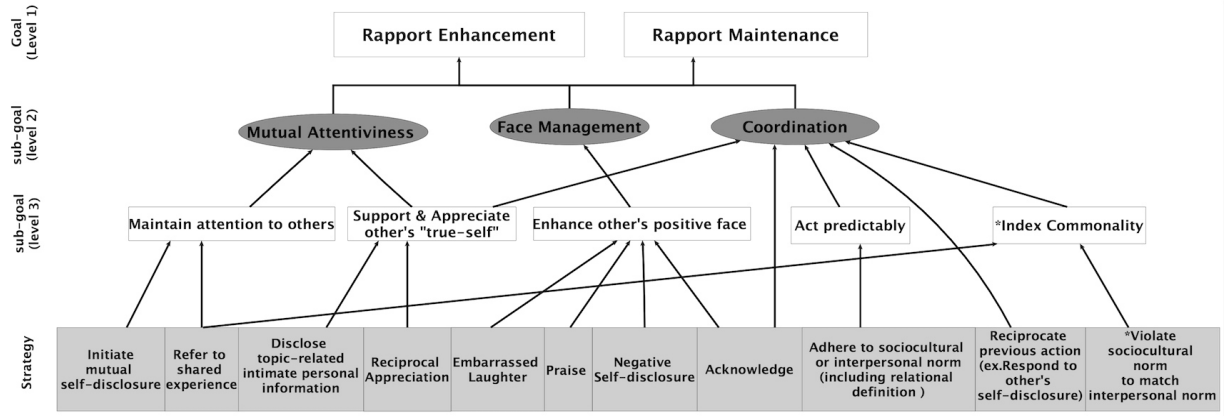


Figure 2.3: Social Functions and Conversational Strategies for Rapport Enhancement and Maintenance

trying to achieve those goals, the strategies and their contribution to the series of sub-goals and interrelating behavioral components of rapport we laid out above - face, mutual attentiveness, coordination. The conversational strategies enumerated here are no doubt not exhaustive. However they include all phenomena found in the literature that were also represented in our data.

In the **rapport-enhancement** orientation (Figure 2.3), people are assumed to begin at state T_1 (stranger) and to have a desire to build rapport with each other, for the reasons laid out above. If we regard rapport-enhancement as a shared task of the dyad, there are different paths to achieve it. In terms of face, people might establish the sub-goal of boosting the interlocutor's face in order to achieve the goal of increasing rapport. Some conversational strategies to accomplish this are to self-disclose negative information, to praise or acknowledge the other's social value, or embarrassed laughter. Social comparison theory (Festinger, 1954) describes how individuals are able to realize and claim more positive social value for themselves through comparison with the other's weaknesses. Our peer tutors illustrate this when they engage in embarassed laughter around their weaknesses in algebra, giving an opportunity for their partner to feel more competent.

As described above, predictability is a core part of coordination. In order to achieve this sub-goal, interactants adhere to behavior expectations. At the initial state T_1, the expectations are guided by sociocultural norms which include the obligation to engage in social validation of the interlocutor's self-disclosures, and to reciprocate with similarly intimate self-disclosure. This also functions to signal attentiveness to the interlocutor. In fact, initiating mutual self-disclosure is a compelling strategy for learning about an individual at the initial stage of the relationship as well as for signaling attentiveness. In our data we also observed that peers often demonstrate mutual attentiveness by referring to past shared experience. As well as increasing common ground, acknowledging and reciprocating reference to previous experience function to increase coordination (Zhao et al., 2014a).

In the **rapport-maintenance** orientation (Figure 2.3), people are assumed to begin at state T_2 (Acquaintance) and have a desire to maintain the current harmonious relationship. Those marked with (*) refer to rapport maintenance only. Typically, friends have some knowledge of each other's behavioral expectations and in order to maintain high rapport, dyads mark their affiliation with one another, and their shared membership in a social identity group. Indexing

commonality strengthens connectedness between in-group members. Compared to stranger peers, friend peers refer to more intimate shared experiences. Moreover, contrary to the sociocultural norms that govern behavior during rapport enhancement, friends may violate sociocultural norms to match their interlocutor's behavioral expectations for example, through rudeness to one another or swearing, both of which were common among friends in our corpus.

In the two orientations just described, we presented strategies for building and maintaining rapport with our interlocutor. However, the **rapport-destruction** orientation is useful in the sense that detecting it will help us choose appropriate rapport "recovery" strategies. (Zhao et al., 2014a) provides more details about rapport-destruction strategies.

2.6 Conclusion

Leveraging a broad base of existing literature and a corpus of data of friends and strangers engaging in peer tutoring, we have made steps towards a unified theoretical framework explaining the process of enhancing, maintaining and destroying rapport in human to human interaction. It should be noted that in this chapter we have traced the relationship between rapport management goals and sub-goals and their associated conversational strategies. We have occasionally described how a conversational strategy is instantiated by a set of observable verbal and nonverbal actions, which will provide quantitative analysis in the following chapters.

Chapter 3

Computational Architecture for Socially-Aware Framework[1]

3.1 Introduction

In the previous chapter, we introduce our theoretical framework of rapport. Now, we induce a modular-based computational architecture to operationalize this framework, which will serve as a personal assistant in a conference and is functioned to facilitate rapport with human user along with carrying out task of recommending session to attend and people to meet in a conference. The goal of the system is to leverage rapport to elicit personal information from the user that can be used to improve the helpfulness and personalization of system responses. We will review the function of each module in this architecture and the details of our developed modules will be discussed in the next following chapters.

3.2 Computational Architecture

Figure 3.1 shows the overview of the architecture, which is from our work (Matsuyama et al., 2016). Our developed modules in this thesis are **Conversational Strategy Classifier** (Chapter 4.2),**Rapport Estimator** (Chapter 4.3) and **Social Reasoner** (Chapter5).We integrate them together with other common components of a dialog system such as ASR, NLU and TTS.

All modules of the system are built on top of the Virtual Human Toolkit (Hartholt et al., 2013).During the recognition and understanding procedures, Microsoft's Cognitive Services API converts speech to text, which is then fed to Microsoft's LUIS (Language Understanding Intelligent Service) to identify user intents. OpenSmile (Eyben et al., 2010) extracts acoustic features from the audio signal, including fundamental frequency (F0), loudness (SMA), jitter and shimmer, which then serve as input to the rapport estimator and the conversational strategy

[1]This section incorporates text from (Matsuyama et al., 2016) which describes a collaboration between Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao,Oscar J. Romero, Sushma Anand Akoju and Justine Cassell. My contribution to this work were integrating Conversational Strategy Classifier, Rapport Estimator, BEAT, Smart Body modules to the whole architecture. Additionally, I contributed to the design of the social reasoner and the message passing infrastructure. I was also involved in writing and modifying the publication version of the paper.

classifier modules.Conversational strategy classifier will automatically recognize particular styles and strategies of talking that contribute to building, maintaining or sometimes destroying a budding relationship. OpenFace (Baltrušaitis et al., 2016)) detects 3D facial landmarks, head pose, gaze and Action Units, and these also serve as input to the rapport estimator.Rapport estimator provides the score of current relational state between the user and system. In decision-make step,task reasoner that focuses on obtaining information to fulfill the user's goals, and a social reasoner that chooses ways of talking that are intended to build rapport in the service of better achieving the user's goals. Finally, on the basis of the output of the dialog manager (which includes the current conversational phase, system intent, and desired conversational strategy) sentence and behavior plans are generated.The generated sentence plan is sent to BEAT, a non-verbal behavior generator(Cassell et al., 2004), which tailors a behavior plan (including relevant hand gestures, eye gaze, head nods, etc.) and outputs the plan as BML (Behavior Markup Language), which is a part of the Virtual Human Toolkit (Hartholt et al., 2013).This plan is then sent to SmartBody, which renders the required non-verbal behaviours.

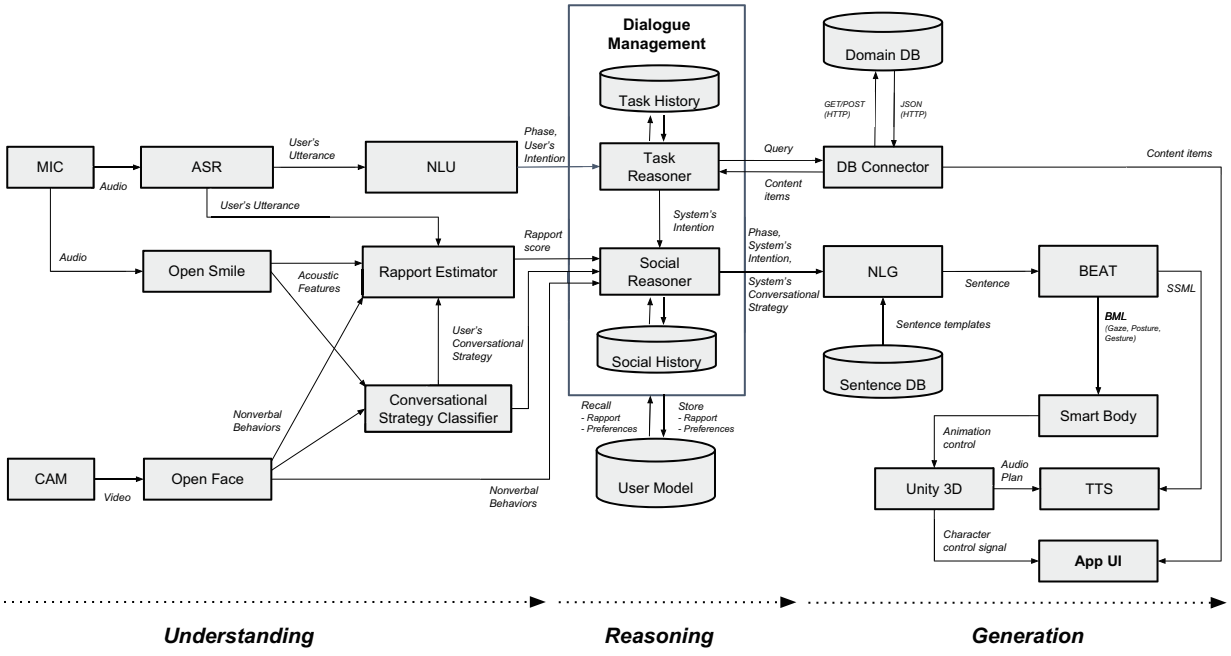


Figure 3.1: Computational Architecture of dyadic rapport (Matsuyama et al., 2016)

3.3 Dialog Examples

Our integrated system was demoed at the SIGDIAL conference (Matsuyama et al., 2016). Sample dialogue in Appendix is extracted from an actual interaction with the system, annotated with the outputs of the different modules as the system works to meet social and task goals.

Chapter 4

Multimodal Behavior Understanding in the Social Context

4.1 Socio-Cognitive Effects of Conversational Strategy Congruence[1]

4.1.1 Introduction

In chapter 2, we introduce our computational model of rapport. As conversational strategies contribute to building, maintaining (or, sometimes, destroying) a budding relationship, it is important to understand what conversational strategies people use, whether they use some more than others, and whether particular ones are more useful than others in particular learning or social contexts (perhaps some work better with boys than girls, or some with old friends rather than new, or some might work better in social situations while others in classroom work etc). In addition, it is important to understand the time-bound patterns of reciprocity, synchrony, or the lack thereof in the use of these conversational strategies to see if not just the type of strategy but also the mutuality of its use, and the dynamics of its use over time plays a role in the impact of its use.

Therefore, as a follow-up experimental study, we leverage structure from the patterns of conversational strategy usage (both cumulative and temporal) by students engaged in reciprocal peer tutoring, and examining their impact on learning and rapport. Specifically, we looked at the conversational strategies of self-disclosure, reference to shared experience and praise and operationalized two measures to quantify the similarity in their usage by two partners : a)first, absolute difference in the number of strategies used, b)second, a dynamic time warping based distance to capture alignment in the timings of strategy usage. Furthermore, since researchers have long posited that friends learn better together than do strangers, we investigated how friends and strangers differ in their conversational strategy exchange patterns. In addition, as prior work has found differences between boys and girls in the use of relationship-building talk, we looked at

[1]This section incorporates text from (Sinha et al., 2015) which describes a collaboration between Tanmay Sinha, Ran Zhao and Justine Cassell. My contribution to this work were writing the coding mannual for conversational strategy annotation and conducting statistical analysis. I was also involved in writing and modifying the publication version of the paper.

gender differences. Finally, we examined how interlocutors differ while working in task vs. social conversations, as well as in their interaction with each other over a period of weeks.

4.1.2 Related Work

The social exchange theory, which defines social behaviors as an exchange (Homans, 1958; Hill and Stull, 1982) motivates our analysis. During the development of relationships, social exchange is regulated by a series of obligations (Emerson, 1976) - how we feel entitled to respond based on the behaviors we expect from others (for e.g, desire to be approved of). Some prior work has proposed that reciprocity is a very important social norm in the early stages of a relationship (Altman, 1973). Conversational strategies can be perceived as social bids. The degree to which interlocutors are successful determines how these exchange patterns will vary over time. For instance, there is always risk and benefit involved during self-disclosures (Petronio, 2012). The costs of disclosing are increased vulnerability and less privacy. The benefits are increased trust, rapport and reciprocation, which could outweigh the costs (Ioinson and Paine, 2007). Therefore, we are interested in investigating how peer-directed social bids, facilitated by the conversational strategies of self-disclosure, reference to shared experience and praise, affect learning and rapport outcomes.

4.1.3 Experiment Setup

In this study, we conduct our experiment on CMU reciprocal peer tutoring dataset (Yu et al., 2013b), which has been explained in chapter 2.Since our goal is to explore socio-cognitive effects of conversational strategies congruence, we introduce both learning outcome and rapport outcome in the following.

Learning Outcome

For every reciprocal peer tutoring session, the tutee was provided with a working sheet comprising of ≈ 10 questions on linear equations, which were to be solved and briefly explained step-by-step. In addition, the tutor was given a correctly solved version of the working sheet that he/she used to guide the tutee in the tutoring period.

To assess learning outcomes during and after the process of tutoring, we computed the following two measures reflecting problem correctness: a)***L1 attempted***: Total percentage of problems correctly solved by the tutee in each of the peer-tutoring sub-sessions in a session, out of total problems attempted in the working sheet. A question in the working sheet was marked as attempted, if at least one step was partially or fully solved, b)***L2 solved***: Total percentage of problems correctly solved by the tutee in each of the 2 peer-tutoring sub-sessions in a session, out of total problems present in the working sheet.

Rapport Outcome

After each session, both participants in the dyad completed 7 point likert scale (1 = Disagree Strongly; 7 = Agree Strongly) questionnaires, reflecting the dimensions of **Attentiveness** (3-item scale indexing interest, attention and respectfulness of the partner towards the speaker, Cronbach α =0.42), **Positivity** (2-item scale indexing friendliness and warmth towards the partner, α =0.72), **Coordination** (3-item scale indexing whether partners felt in sync, could say everything that they

wanted to say and that the interaction was not frustrating, α =0.64), and **Long Term Rapport** (3-item scale indexing whether the partners felt that they knew each other, were more comfortable and had greater liking compared to the previous interaction session, α =0.78). In addition, the questionnaire asked about **Self-efficacy** (7-item scale indexing whether the partners thought they were good tutors, learned a lot from tutoring and were concerned about tutoring quality, motivation and impact on the tutee, α =0.5).

In order to compute a dyadic measure from individual questionnaire ratings, we computed the following two measures: a)***R1 total***: Total score for each questionnaire dimension, calculated by addition of individual questionnaire scores, b)***R2 mean&sd***: Mean of the score for each questionnaire dimension for the dyad, subtracted by the standard deviation. Intuitively, this metric will be higher if average questionnaire scores are higher for the dyad, as well as individual variability from the mean is lower, and vice versa.

4.1.4 Methodology

Coding of Conversational Strategies

In order to construct a reliably annotated corpus, we employed 3-5 human annotators to code conversational strategies that prior work has shown to contribute to rapport. In this section, we explored conversational strategies of self-disclosure, reference to shared experience and praise. Annotators were provided with explicit definitions and examples to use in making their judgment(see Appendix). Inter-rater reliability of conversational strategy annotations, computed via Krippendorff's alpha was 0.753 for self-disclosure, 0.798 for reference to shared experience and 1.0 for praise. After achieving high enough inter-rater reliability, most of the sessions were coded by independently by the annotators. Below, we briefly describe the rationale behind our coding manuals, along with example utterances from our dataset depicting these three conversational strategies.

Operationalizing Similarity Constructs

We utilized dynamic time warping (DTW) to obtain a global distance that can characterize how conversational strategy usage for each partner in the dyad is aligned in time. Originally presented by (Kruskal and Liberman, 1983) for speech recognition purposes, this technique allows two time series that are similar but locally out of phase to align in a nonlinear manner. In our case, each element in the two time series refers to time from the start of a peer tutoring session (in seconds) at which each individual in the dyad used certain conversational strategy. Concretely, given these two time series, say $A = [a_1, ...a_n] \in \mathbb{R}^{1Xn}$ and $B = [b_1, ...b_n] \in \mathbb{R}^{1Xn}$, DTW is a technique to align A and B such that the sum of the Euclidean distances between the aligned samples is minimized. In order to perform this alignment, DTW can distort (or, as (Kruskal and Liberman, 1983) call it, "warp") the time axis - compressing it at some places and expanding it at others.

Thus, by viewing the time axis as a stretchable one, DTW is able to match (via construction of a warping path) a point of time series A even with surrounding points of time series B. Minimum global dissimilarity, or DTW distance can be assumed as the stretch-insensitive measure of the inherent difference between two given time series. Furthermore, the shape of the warping curve itself provides information about which point matches which, i.e., the pair wise correspondences of time points can be easily inspected.

This warping path is a central part of comparing the two time series, since it determines which points match and are to be used for calculating the distance between the two time series. One simplistic way, for instance, is linear matching, that aligns the ith point of the first curve with the ith point of the second curve. Because this matching is very sensitive to small distortions in the time axis, a more computationally expensive way is to perform a complete matching. This technique calculates the distance between every point of the first curve and every point of the second curve. For every point, the smallest distance to the other curve is decided. These distances are summed and divided by the number of points. Each point can match with no more than one point of the other curve.

In contrast, a non-linear (elastic) alignment provided by DTW produces an intuitive similarity measure, allowing similar shapes to match even if they are locally out of phase on the time axis, by allowing elastic shifting in order to detect similar shapes with different phases. (Rabiner and Juang, 1993) provide a more comprehensive technical description of the DTW algorithm.

To investigate similarity in the pattern (timing) of conversational strategy usage, we employed the dtw package in R (Giorgino, 2009), and selected the following parameters: a)First, we utilized the $symmetric2$ step-pattern (Sakoe and Chiba, 1978) that lists transitions allowed while searching for the minimum distance path between the two time series, with the constraint that one diagonal step costs as much as the two equivalent steps along the sides. Intuitively, step-patterns limit the maximum amount of time stretch and compression allowed at any point of the alignment, b)Second, we performed an open-ended alignment, meaning that we freed the endpoint of time series B in order to allow for a partial match. Intuitively, relaxing the end-point constraint results in computing the alignment which best matches all of time series B with a leading part of time series A, c)Third, we normalized the DTW distance by the length of the two input time series, in order to accommodate time-series of varying lengths (certain dyads, for instance, do lot more self-disclosure relative to other dyads).

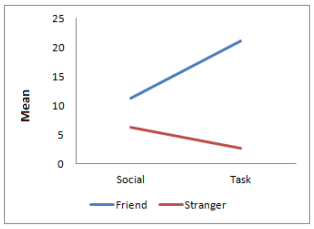
Finally, since conversational strategy usage by the partners in the dyad not only varies in terms of expressive frequency, but also in terms of time progression, we computed two measures reflective of similarity between partners in the dyad: a)**Diff**: This metric measures the absolute difference between the number of a specific conversational strategy usage by both students in a dyad. Thus, a lower difference would imply that the count of a specific conversational strategy is very similar for the dyad, b)**Diff**+**Time**: Since **Diff** metric cannot take into account the temporal distribution of conversational strategies, we utilized the normalized DTW distance obtained from the time warping algorithm as a second measure reflective of the similarity in the pattern of timings of conversational strategy usage.

4.1.5 Explanatory Analysis

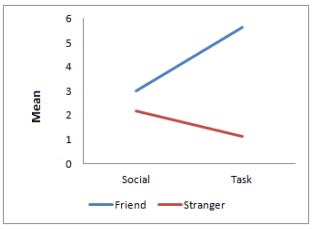
We conducted a four-way $2X2X2X5$ repeated measures ANOVA to investigate the effect of gender (male, female), relationship status (friends, strangers), period (social, task) and session (1, 2, 3, 4, 5) on the a)total conversational strategy usage by partners in a dyad, b)absolute & DTW difference in the number of usages by partners in a dyad. Period and session were used as within-subject repeated measures.

For all statistically significant effects ($p < 0.05$) reported, we also looked at effect size (Cohen's d) in order to assess the practical significance of these results. In essence, effect size is the difference between two means (e.g., friends minus strangers) divided by the pooled standard

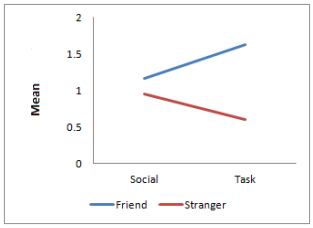
deviation (adjusted with weights for the sample sizes) of the two conditions (Ellis, 2010).Figure 4.1 demonstrates most of our findings. Next, we computed correlations (Pearson r, Spearman Rank ρ) to find relationships among the outcome variables of learning, rapport and operationalized measures of conversational strategy similarity. Significance of the correlation was assessed via two tailed t-test. Since the total conversational strategy usage hides information about the exchange patterns of the individuals involved in peer tutoring, we were interested in testing the impact of only conversational strategy similarity on socio-cognitive outcomes.All details of the analysis are provided in (Sinha et al., 2015).



a. Self-Disclosure (Total)



b. Self-Disclosure (Diff)



c. Shared Experience (Diff)

Figure 4.1: Plots depicting significant interaction effects from the repeated measures ANOVA

4.1.6 Conclusion

To summarize, in this work, we explored how the similarity in aggregated count of conversational strategies by partners, as well its temporal variations, relate to rapport and learning. We discovered significant effects that explain importance of gender and relationship status in deciphering conversational strategy exchange patterns, as well as a different set of rules governing social exchanges within these settings.

4.2 Predictive Model for Conversational Strategies Recognition[2]

4.2.1 Introduction

In this section, we propose a technique to automatically recognize conversational strategies. We demonstrate that these conversational strategies are most effectively recognized when verbal

[2]This section incorporates text from (Zhao et al., 2016a) which describes a collaboration between Ran Zhao,Tanmay Sinha and Justine Cassell. My contribution to this work were proposing the idea, extracting multimodal features, running some parts of statistical test, training and evaluating the machine learning model.Also, I involved in writing and modifying the publication version of the paper.

(linguistic), visual (nonverbal) and vocal (acoustic) features are all taken into account.

4.2.2 Related Work

(Wang et al., 2016) developed a model to measure self-disclosure in social networking sites by deploying emotional valence, social distance between the poster and other people and linguistic features such as those identified by the Linguistic Inquiry and Word Count program (LIWC) etc. While the features used here are quite interesting, this study relied only on the verbal aspects of talk, while we also include vocal and visual features.

Interesting prior work on quantifying social norm violation has taken a heavily data-driven focus (Danescu-Niculescu-Mizil et al., 2013b; Wang et al., 2016). For instance, (Danescu-Niculescu-Mizil et al., 2013b) trained a series of bigram language models to quantify the violation of social norms in users' posts on an online community by leveraging cross-entropy value, or the deviation of word sequences predicted by the language model and their usage by the user. Another kind of social norm violation was examined by (Riloff et al., 2013), who developed a classifier to identify a specific type of sarcasm in tweets. They utilized a bootstrapping algorithm to automatically extract lists of positive sentiment phrases and negative situation phrases from given sarcastic tweets, which were in turn leveraged to recognize sarcasm in an SVM classifier. Experimental results showed the adequacy of their approach.

(Wang et al., 2012) investigated the different social functions of language as used by friends or strangers in teen peer-tutoring dialogs. This work was able to successfully predict impoliteness and positivity in the next turn of the dialog. Their success with both annotated and automatically extracted features suggests that a dialog system will be able to employ similar analyses to signal relationships with users. Other work, such as (Danescu-Niculescu-Mizil et al., 2013a) has developed computational frameworks to automatically classify requests along a scale of politeness. Politeness strategies such as requests, gratitude and greetings, as well as their specialized lexicons, were used as features to train a classifier.Additional past work are discussed in (Zhao et al., 2016a).

However, a common limitation of the above work is its focus on only the verbal modality, while studies have shown conversational strategies to be associated with specific kinds of nonverbal behaviors. For instance, (Kang et al., 2012) discovered that head tilts and pauses were the strongest nonverbal cues to interpersonal intimacy. Unfortunately, here too only one modality was examined. While nonverbal behavioral correlates to intimacy in self-disclosure were modeled, the verbal and vocal modalities of the conversation was ignored. Computational work has also modeled rapport using only nonverbal information (Huang et al., 2011). In what follows we describe our approach to modeling social conversational phenomena, which relies on verbal, visual and vocal content to automatically recognize conversational strategies. Our models are trained on a peer tutoring corpus, which gives us the opportunity to look at conversational strategies as they are used in both a task and social context.

4.2.3 Ground Truth

We assessed our automatic recognition of conversational strategies against this corpus annotated for those strategies (as well as other educational tutoring phenomena not discussed here). Inter-rater reliability (IRR) for the conversational strategy annotations, computed via Krippendorff's

alpha, was 0.75 for self-disclosure, 0.79 for reference to shared experience, 1.0 for praise and 0.75 for social norm violation. IRR for visual behavior was 0.89 for eye gaze, 0.75 for smile count (how many smiles occur), 0.64 for smile duration and 0.99 for head nod. Below we discuss the definitions of each conversational strategy and nonverbal behavior that was annotated.

4.2.4 Understanding Conversational Strategies

Our first objective was to understand the nature of different conversational strategies. Towards this end, we first under-sampled the non-annotated examples of self disclosure, shared experience, praise and social norm violation in order to create a balanced dataset of utterances. The utterances chosen to reflect the non-annotated cases were randomly selected. We made sure to have a similar average utterance length for all annotated and non-annotated cases, to prevent conflation of results due to lower or higher opportunities for detection of multimodal features. The final corpus (selected from 60 interaction sessions) comprised of 1014 self disclosure and 1014 non-self disclosure, 184 shared experience and 184 non-shared experience, 167 praise and 167 non-praise, 7470 social norm violation and 7470 non-social norm violation.

Second, we explored observable verbal and vocal behaviors of interest that could potentially be associated with different conversational strategies, assessing whether the mean value of these features were significantly higher in utterances with a particular conversational strategy label than in ones with no label (two-tailed correlated samples t-test). Bonferroni correction was used to correct the p-values with respect to the number of features, because of multiple comparisons involved. Finally, for all significant results (p <0.05), we also calculated effect size via Cohen's d to test for generalizability of results.

Third, for visual behaviors like smile, eye gaze, head nod, we binarized these features by denoting their presence (1) or absence (0) in one clause. If an individual shifts gaze during a particular spoke conversational strategy, we might have multiple types of eye gaze represented. We performed χ^2 test to see whether the appearance of visual annotations were independent of whether the utterance belonged to a particular conversational strategy or not. For all significant χ^2 test statistics, odds ratio (o) was computed to explore co-occurrence likelihood. Majority of the features discussed in the subsequent sub-sections were drawn from qualitative observations and note-taking, during and after the formulation of our coding manuals.

Verbal

We used Linguistic Inquiry and Word Count (LIWC 2015) (Pennebaker et al., 2015) to quantify verbal cues of interest that were semantically associated with a broad range of psychological constructs and could be useful in distinguishing conversational strategies. The input to LIWC were conversational transcripts that had been transcribed and segmented into syntactic clauses.

Self-disclosure: We observed personal concerns of students (sum of words identified as belonging to categories of work, leisure, home, money, religion and death etc) to be significantly higher, than in non self-disclosure utterances with a moderate effect size (d=0.44), signaling that students referred significantly more to their personal concerns during self-disclosure. Next, due to the fact that self-disclosures are often likely to comprise of emotional expressions when revealing one's likes and dislikes (Sparrevohn and Rapee, 2009), we used the LIWC dictionary to capture

words representative of negative emotions (d=0.32) and positive emotion words (d=0.18). Also, to formalize the intuition that when people reveal themselves in an authentic or honest way, they are more personal, humble, and vulnerable, the standardized LIWC summary variable of Authenticity (d=1.16) was taken into account. Finally, as expected, we found self-disclosure utterances had significantly higher usage of first person singular pronouns (d=1.62).

Reference to shared experience: We looked at three LIWC categories: (1) Affiliation drive, which comprises words signaling a need to affiliate such as ally, friend, social etc (d=0.92), (2) Time Orientation words, which capture past (mostly in ROE) , present (mostly in RIE) and future focus and comprises words such as ago, did, talked, today, is, now, may, will, soon etc (d=0.95). Such words are not only used by interlocutors to index commonality within a time frame (Enfield, 2013), but also to signal an increased need for affiliation with the conversational partner, perhaps to indicate common ground (Clark, 1996), (3) First person plural such as we, us, our etc. In line with expectations, this feature had high effect size (d=0.93), since interlocutors focused on both themselves and the conversational partner.

Praise: We looked at positive emotions (d=2.55), since praise is one form of verbal persuasion that increases the interlocutor's confidence and boosts self-efficacy (Zimmerman, 2000). Most of the praise utterances in our dataset were not very specific or directed at the tutee's performance or effort. Also, the LIWC standardized summary variable of Emotional Tone from LIWC was considered for the sake of completeness, which puts positive emotion and negative emotion dimensions into a single summary variable, such that the higher the number, the more positive the tone (d=3.56).

Social norm violation: We looked at different categories of off-task talk from LIWC, such as social processes comprising words related to friends, family, male and female references (d=0.78), biological processes comprising words belonging to the categories of body, health etc (d=0.30) and personal concerns (d=0.24). The effect sizes across these categories ranged from moderate to low. Next, we looked at usage of swearing words like fuck, damn, shit etc and found low effect size (d=0.13) for this category in utterances of social norm violation. For the LIWC category of anger (words such as hate, annoyed etc), the effect size was moderate (d=0.27).

In our qualitative analysis of social norm violation utterances, we had discovered interactions of students to be reflective of need for power, meaning attention to or awareness of relative status in a social setting (perhaps this could be a result of putting one student in the tutor role). We formalized this intuition from the LIWC category of power drive that comprises words such as superior etc (d=0.18). Finally, based on prior work (Kacewicz et al., 2009) that found increased use of first-person plural to be a good predictor of higher status, and increased use of first-person singular to be a good predictor of lower status, we posited that when students violated social norms, they were more likely to freely make statements that involved others. However, the effect size for first-person plural usage in utterances of social norm violation was negligible (d=0.07). Table 4.1 provides complete set of results.

Vocal

In our qualitative observations, we noticed the variations of both pitch and loudness when interlocutors used different conversational strategies. We were thus motivated to explore the mean difference of those low-level vocal descriptors as differentiators among the different conversational

Conversational Strategy	Verbal/Vocal(Speaker)	t-test value	Mean value	Effect Size
1. Self-Disclosure	LIWC Personal Concerns	t(1013)=7.06***	SD=4.13, NSD=1.58	d=0.44
	LIWC Positive Emotion	t(1013)=2.98**	SD=7.61, NSD=5.50	d=0.18
	LIWC Negative Emotion	t(1013)=5.51***	SD=5.62, NSD=2.22	d=0.32
	LIWC First Person Singular	t(1013)=25.87***	SD=20.12, NSD=7.77	d=1.62
	LIWC Authenticity	t(1013)=18.59***	SD=66.71, NSD=34.07	d=1.16
	pcm-loudness-sma-amean	t(1013)=4.11***	SD=0.64, NSD=0.59	d=0.26
2. Shared Experience	LIWC Affiliation Drive	t(183)=6.22***	SE=4.64, NSE=0.77	d=0.92
	LIWC Time Orientation	t(183)=6.47***	SE=24.89, NSE=15.02	d=0.95
	LIWC First Person Plural	t(183)=6.29***	SE=3.99, NSE=0.48	d=0.93
	shimmerLocal-sma-amean	t(183)=-2.21*	SE=0.18, NSE=0.194	d=0.32
3. Praise	LIWC Positive Emotion	t(166)=16.48***	PR=55.63, NPR=4.56	d=3.56
	LIWC Emotional Tone	t(166)=22.96***	PR=91.1, NPR=33.5	d=2.55
	pcm-loudness-sma-amean	t(166)=-3.33***	PR=0.5, NPR=0.6	d=-0.51
	jitterLocal-sma-amean	t(166)=2.93*	PR=0.1, NPR=0.07	d=0.45
	shimmerLocal-sma-amean	t(166)=2.56*	PR=0.2, NPR=0.18	d=0.39
4. Social Norm Violation	LIWC Social Processes	t(7469)=33.98***	VSN=17.35, NVSN=6.45	d=0.78
	LIWC Biological Processes	t(7469)=12.95***	VSN=4.21, NVSN=1.38	d=0.30
	LIWC Personal Concerns	t(7469)=10.61***	VSN=2.61, NVSN=1.33	d=0.24
	LIWC Swearing	t(7469)=5.85***	VSN=0.49, NVSN=0.11	d=0.13
	LIWC Anger	t(7469)=11.64***	VSN=1.19, NVSN=0.20	d=0.27
	LIWC Power Drive	t(7469)=7.83***	VSN=1.99, NVSN=1.14	d=0.18
	LIWC First Person Plural	t(7469)=3.23**	VSN=0.85, NVSN=0.64	d=0.07
	pcm-loudness-sma-amean	t(7469)=31.24***	VSN=0.69, NVSN=0.56	d=0.72
	F0final-sma-amean	t(7469)=26.6***	VSN=231.09, NVSN=206.99	d=0.61
	jitterLocal-sma-amean	t(7469)=-4.09***	VSN=0.083, NVSN=0.087	d=-0.09
	shimmerLocal-sma-amean	t(7469)=-7.02***	VSN=0.1818, NVSN=0.1897	d=-0.16

Table 4.1: Complete Statistics for presence of numeric verbal and vocal features in Self-Disclosure (SD)/Non-Self Disclosure (NSD), Shared Experience (SE)/Non-Reference to Shared Experience (NSE), Praise (PR)/Non-Praise (NPR) and Violation of Social Norms (VSN)/Non-Violation of Social Norms (NVSN). Effect size assessed via Cohen's d. Significance: ***:$p < 0.001$, **:$p < 0.01$, *:$p < 0.05$

strategies. By using Open Smile (Eyben et al., 2010), we extracted two sets of basic features - for loudness features, pcm-loudness and its delta coefficient were tested; for pitch-based features, jitterLocal, jitterDDP, shimmerLocal, F0final and also their delta coefficients were tested. pcm-loudness represents the loudness as the normalised intensity raised to a power of 0.3. F0final is the smoothed fundamental frequency contour. JitterLocal is the frame-to-frame pitch period length deviations. JitterDDP is the differential frame-to-frame jitter. ShimmerLocal is the frame-to-frame amplitude deviations between pitch periods.

Self-disclosure: We found a moderate effect size for pcm-loudness-sma-amean (d=0.26). Despite often becoming excited when disclosing things that they loved or liked, sometimes students also seemed to hesitate and spoke at a lower pitch when they revealed a transgressive act. However, the effect size for pitch was negligible. One potential reason for our results not aligning with hypothesis could be consideration of utterances with annotations of enduring states as well

as transgressive acts together.

Reference to shared experience: We found a moderate negative effect size for the shimmerLocal-sma-amean (d=-0.32).

Praise: We found negative effect size for loudness (d=-0.51), meaning the speakers spoke in a lower voice when praising the interlocutor (mostly the tutee). We also found positive and moderate effect sizes for jitterLocal-sma-amean (d=0.45) and shimmerLocal-sma-amean (d=0.39).

Social norm violation: We found high effect sizes for pcm-loudness-sma-amean (d=0.72) and F0final-sma-amean (d=0.61) and interestingly, negative effect sizes for jitter (d=-0.09) and shimmer (d=-0.16). One potential reason could be that when student violate social norms, their behaviors are likely to become outliers compared to their normative behaviors. In fact, we noticed usage of "joking" tone of voice (Norrick, 2003) and pitch different than usual, to signal a social norm violation. When the content of the utterance was unaccepted by the social norms, students also tried to lower down their voice, which could be a way of hedging these violations. Table 4.1 provides complete set of results.

Visual

Computing the odds ratio o involved comparing the odds of occurrence of a non-verbal behavior for a pair of categories of a second variable (whether an utterance was a specific conversational strategy or not). Overall, we found that that smile and gaze were significantly more likely to occur in utterances of self-disclosure (o(Smile)=1.67, o(gP)=2.39, o(gN)=0.498, o(gO)=0.29, o(gE)=2.8) compared to a non self-disclosure utterance. A similar trend was observed for reference to shared experience (o(Smile)=1.75, o(gP)=3.02, o(gN)=0.58, o(gO)=0.31, o(gE)=4.19) and social norm violation (o(Smile)=3.35, o(gP)=2.75, o(gN)=0.8, o(gO)=0.47, o(gE)=1.67) utterances, compared to utterances that did not belong to these categories.

The high odds ratio for gP in these results suggests that an interlocutor was likely to gaze at their partner when using specific conversational strategies, signaling attention towards the interlocutor. The extremely high odds ratio for smiling behaviors during a social norm violation is also interesting. However, for praise utterances, we did not find all kinds of gaze and smile to be more likely to occur than non-praise utterances. Only gazing at partner (o(gP)=0.44) or their worksheet (o(gN)=4.29) or gazing elsewhere (o(gE)=0.30) were among the non-verbals that were significantly greatly present in praise utterances. Table 4.2 provides complete set of results for the speaker (as discussed above) and also for the listener.

4.2.5 Machine Learning Modeling

In this part, our objective was to build a computational model for conversational strategy recognition. Towards this end, we first took each clause, or the smallest units that can express a complete proposition, as the prediction unit. Next, three sets of features were used as input. The first set f_1 comprised verbal (LIWC), vocal and visual features of the speaker, informed from the qualitative and quantitative analysis as discussed above. While LIWC features helped in categorization of words used during usage of a particular conversational strategy, they did not capture contextual usage of words within the utterance. Thus, we also added bigrams, part of speech bigrams and word-part of speech pairs from the speaker's utterance.

Conversational Strategy	Visual (Speaker) - χ^2 test value - Odds Ratio	Visual (Listener) - χ^2 test value - Odds Ratio
1. Self-Disclosure	Smile - χ^2(1,1013)=20.67*** - o=1.67 Gaze (gP) - χ^2(1,1013)=93.04*** - o=2.39 Gaze (gN) - χ^2(1,1013)=35.1*** - o=0.49 Gaze (gO) - χ^2(1,1013)=173.88*** - o=0.29 Gaze (gE) - χ^2(1,1013)=120.77*** - o=1.8	Smile - χ^2(1,1013)=18.63*** - o=1.63 Gaze (gP) - χ^2(1,1013)=131.34*** - o=2.84 Gaze (gN) - χ^2(1,1013)=73.23*** - o=0.38 Gaze (gO) - χ^2(1,1013)=152.12*** - o=0.31 Gaze (gE) - χ^2(1,1013)=78.92*** - o=2.37
2. Shared Experience	Smile - χ^2(1,183)=4.73* - o=1.75 Gaze (gP) - χ^2(1,183)=25.37*** - o=3.02 Gaze (gN) - χ^2(1,183)=3.73* - o=0.58 Gaze (gO) - χ^2(1,183)=27.87*** - o=0.31 Gaze (gE) - χ^2(1,183)=38.13*** - o=4.19	Smile - χ^2(1,183)=7.53** - o=2.07 Gaze (gP) - χ^2(1,183)=33.36*** - o=3.59 Gaze (gN) - χ^2(1,183)=17.68*** - o=0.32 Gaze (gO) - χ^2(1,183)=16.55*** - o=0.41 Gaze (gE) - χ^2(1,183)=32.45*** - o=3.92
3. Praise	Gaze (gP) - χ^2(1,166)=9.94*** - o=0.44 Gaze (gN) - χ^2(1,166)=37.52*** - o=4.29 Gaze (gO) - N.S Gaze (gE) - χ^2(1,166)=14.44*** - o=0.30	Gaze (gP) - χ^2(1,166)=14.22*** - o=0.39 Gaze (gN) - χ^2(1,166)=15.19*** - o=0.33 Gaze (gO) - χ^2(1,166)=24.23*** - o=3.30 Gaze (gE) - χ^2(1,166)=9.77** - o=0.39
4. Social Norm Violation	Smile - χ^2(1,7469)=871.73*** - o=3.35 Gaze (gP) - χ^2(1,7469)=911.89*** - o=2.75 Gaze (gN) - χ^2(1,7469)=34.82*** - o=0.8 Gaze (gO) - χ^2(1,7469)=515.26*** - o=0.47 Gaze (gE) - χ^2(1,7469)=195.17*** - o=1.67 Head Nod - χ^2(1,7469)=8.06** - o=0.77	Smile - χ^2(1,7469)=869.29*** - o=3.37 Gaze (gP) - χ^2(1,7469)=609.06*** - o=2.27 Gaze (gN) - χ^2(1,7469)=239.22*** - o=0.55 Gaze (gO) - χ^2(1,7469)=110.48*** - o=0.70 Gaze (gE) - χ^2(1,7469)=12.38** - o=1.14 Head Nod - χ^2(1,7469)=44.51*** - o=0.56

Table 4.2: Complete Statistics for presence of binary non-verbal features in Self-Disclosure (SD), Shared Experience (SE), Praise (PR) and Violation of Social Norms (VSN). Odds ratio signals how much more likely is a non-verbal behavior likely to occur in conversational strategy utterances compared to non-conversational strategy utterances. Significance: ***:$p < 0.001$, **:$p < 0.01$, *:$p < 0.05$

.

In addition to the speaker's behavior, we also added two sets of interlocutor behavior to capture the context around usage of a conversational strategy. The feature set f_2 comprised visual behaviors of the interlocutor (listener) in the current turn. The feature set f_3 comprised verbal (bigrams, part of speech bigrams and word-part of speech pairs), vocal and visual features of the interlocutor in the previous turn.

Finally, early fusion was applied on these multimodal features (by concatenation) and L2 regularized logistic regression with 10-fold cross validation was used as the machine learning algorithm, with rare threshold for feature extraction being set to 10 and performance evaluated using accuracy and kappa[3] measures. The table 4.3 shows our comparison with other standard machine learning algorithms such as Support Vector Machine (SVM) and Naive Bayes (NB), where we found Logistic Regression (LR) to perform better in recognition of the four conversational strategies. In next sub-section, we therefore denote the feature weights derived from logistic regression in brackets to offer interpretability of results.

[3]The discriminative ability over chance of a predictive model, for the target annotation, or the accuracy adjusted for chance

Conversational Strategy	LR	SVM	NB
Self-disclosure	Acc=0.85 κ= 0.7	Acc=0.84 κ= 0.68	Acc=0.83 κ= 0.65
Shared Experience	Acc=0.84 κ= 0.67	Acc=0.82 κ= 0.64	Acc=0.79 κ= 0.59
Praise	Acc=0.91 κ= 0.81	Acc=0.90 κ= 0.80	Acc=0.88 κ= 0.76
Social Norm Violation	Acc=0.80 κ= 0.61	Acc=0.78 κ= 0.55	Acc=0.73 κ= 0.47

Table 4.3: Comparative Performance Evaluation using Accuracy (Acc) and Kappa (κ) for Logistic Regression (LR), Support Vector Machine (SVM) and Naive Bayes (NB)

4.2.6 Results and Discussion

Self-Disclosure: We could successfully identify self-disclosure from non self-disclosure utterances with an accuracy of **85%** and a kappa of **70%**. The top features from feature set f_1 predictive of speakers disclosing themselves included gazing at partner (0.44), head nodding (0.24) and not gazing at their own worksheet (-0.60) or the interlocutor's worksheet (-0.21). Head nod is a way to emphasize what one is saying (Poggi et al., 2010), while gazing at the partner signals one's attention. Higher usage of first person singular by the speaker (0.04) was also positively predictive of self-disclosure in the utterance. The top features from feature set f_2 predictive of speakers disclosing included listener behaviors such as head nodding (0.3) to communicate their attention (Schegloff, 1982), gazing elsewhere (0.12) or at the speaker (0.09) instead of gazing at their own worksheet (-0.89) or the speaker's worksheet (-0.27). The top features from feature set f_3 predictive of speakers disclosing included no smiling (-0.30),no head nodding (-0.15) and lower loudness in voice (-0.11) from the interlocutor in the last turn.

Reference to shared experience: We achieved an accuracy of **84%** and kappa of **67%** for prediction. The top features from feature set f_1 predictive of speakers referring to shared experience included not gazing at own worksheet (-0.66), partner's worksheet (-0.40) or at the partner (-0.22), no smiling (-0.18) and having lower shimmer in voice (-0.26). Instead, words signaling affiliation drive (0.07) and time orientation (0.06) from the speaker were deployed to index shared experience. The top features from feature set f_2 predictive of speakers using shared experience included listener behaviors such as smiling (0.53) perhaps to indicate appreciation towards the content of the talk, or encourage the speaker to go on (Niewiadomski et al., 2010). Besides, the listener gazing elsewhere (0.50) or at the speaker (0.47), and neither gazing at own worksheet (-0.45) nor head nodding (-0.28) had strong predictive power. The top features from feature set f_3 predictive of speakers using shared experience included lower loudness in voice (-0.58), smiling (0.47), gazing elsewhere (0.59), at own worksheet (0.27) or at the partner (0.22) but not at partner's worksheet (-0.40) from the interlocutor in the last turn.

Praise: For praise, our computational model achieved an accuracy of **91%** and kappa of **81%**. The top features from feature set f_1 predictive of speakers using praise included gazing at partner's worksheet (0.68) indicative of directing attention to the partner's (perhaps the tutee's) work, smiling (0.51), perhaps to mitigate the potential embarassment of praise (Niewiadomski

et al., 2010) and head nodding (0.35) with a positive tone of voice (0.04), perhaps to emphasize the praise. The top features from feature set f_2 predictive of speakers using praise included listener behaviors such as head nodding (0.45) for backchanneling and acknowledgement and not gazing at partner's worksheet (-1.06), elsewhere (-0.5) or at the partner (-0.49). The top features from feature set f_3 predictive of speakers using praise included smiling (0.51), lower loudness in voice (-0.91) and overlap (-0.66) from the interlocutor in the last turn.

Violation of Social Norm: We achieved an accuracy of **80%** and kappa of **61%** for prediction. The top features from feature set f_1 predictive of speakers violating social norms included smiling (0.40), gazing at partner (0.45) but not head nodding (-0.389). (Keltner and Buswell, 1997) introduced a remedial account of embarrassment, emphasizing that smiles signal awareness of a social norm being violated and serve to provoke forgiveness from the interlocutor, in addition to being a hedging indicator. (Kraut and Johnston, 1979) posited that smiling evolved from primate appeasement displays and is likely to occur when a person has violated a social norm. The top features from feature set f_2 predictive of speakers violating social norms included listener behaviors such as smiling (0.54), gazing at own worksheet (0.32) or at the partner's (0.14). The top features from feature set f_3 predictive of speakers violating social norms included high loudness (0.86) and jitter in voice (0.50), lower shimmer in voice (-0.53), gazing at own worksheet (0.49) and no head nodding (-0.31) from the interlocutor in the last turn.

4.2.7 Post-experiment[4]

The accuracy of detection violation of social norm is comparatively lower than other conversational strategies.Thus, we conduct a post-experiment analysis and propose a more advanced model based on our previous studies.Social norms are shared rules that govern and facilitate social interaction. Violating such social norms via teasing and insults may serve to upend power imbalances or, on the contrary reinforce solidarity and rapport in conversation, rapport which is highly situated and context-dependent (Ogan et al., 2012a). In such a sway, we hypothesize that the performance of detect social norm violation should attribute to the fact that logistic regression fails to model the dialog context during its prediction. we extend our previous work by leveraging the power of recurrent neural networks and multimodal information present in the interaction, and propose a predictive model to recognize social norm violation. Using long-term temporal and contextual information, our model achieves an F1 score of 0.705.

Model

We treated a dialog D as a sequence of clauses $c_0, ...c_T$, where T was the number of clauses in the D. Each clause c_i was a tuple $([w_0^i, ...w_m^i], e_i)$, where $[w_0^i, ...w_m^i]$ was the m words in the clause c_i, and e_i was the corresponding meta information such as the relationship of the dyad and nonverbal behavior during the generation of the clause. The handcrafted feature of size 3782 was

[4]This section incorporates text from (Zhao et al., 2016d) which describes a collaboration between Tiancheng Zhao, Ran Zhao,and Justine Cassell. My contribution to this work were proposing the idea, extracting multimodal features, implementing and training logistic regression and local-context RNN model, writing and modifying the publication version of the paper. My co-author, Tiancheng Zhao, contributed to implementing and training global-context RNN model, writing and modifying the publication version of the chapter.

denoted as f_i, and could be viewed as a mapping function $F : c_i \rightarrow f_i$. Meanwhile, each clause was associated with a binary label $y_i \in \{0, 1\}$ that indicates the ground truth of whether c_i is a violation of social norm. Eventually, the goal was to model $p(y_t|c_{0:t})$, the conditional distribution over whether the latest clause was a violation of social norm, given the entire history of the dialog.

- Logistic Regression Model
  We first trained a L2 regularized logistic regression model using the proposed verbal and visual features f_i as inputs (leftmost in Figure 4.2). This model serves as our baseline.

- Local/Global-Context RNN Model
  Past empirical results suggest two possible hypotheses of improving the model performance:

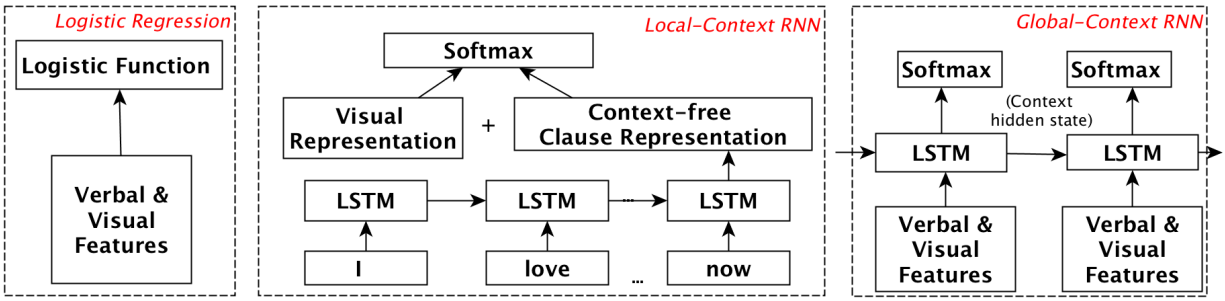


Figure 4.2: Three proposed computational models.

1. improvement in clause level representation 2. inclusion of contextual information for prediction. Therefore, we designed Local/Global-Context models to test these hypotheses.

The Local-Context recurrent neural network (RNN) models the context inside a clause at the word-level by encoding word embeddings of size 300 in a clause c_i sequentially using a Long-short Term Memory (LSTM) cell of size 300. The mechanism of LSTM is defined as:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ j_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W[h_{t-1}, x_t]$$

$$c_t = f_t \odot c_{t-1} + i_t \odot j_t$$

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

We treated last hidden LSTM output h_m^i as the clause embedding and concatenated that with the corresponding meta information vector e_i. The combined vector was linearly transformed and then fed into a softmax function.

Next our Global-Context RNN investigated the influence of clause-level context in detecting social norm violation, by using the LSTM cells to model the long-term temporal dependencies. For a fair comparison, we used the same hand-crafted feature f_i used in the logistic regression model as the representation of clause c_i. As shown in Figure 4.2, we first obtained a linear embedding of size 150 $emb_i = W_e f_i + b_i$ of f_i. Then emb_i was used

as the inputs to LSTM of size 600. The hidden output h_i at each time step was fed into a multilayer perceptron (MLP) with 1 hidden layer of size 100. We applied 50% dropout regularization (Zaremba et al., 2014) at the input/output of LSTM and MLP hidden layer for better generalization. Finally the model was optimized w.r.t to the cross entropy loss. A further challenge was the length of dialog. The average number of clauses in training dialog was 817.8, which made it computationally intractable to backpropagate through the entire sequence. Therefore, truncated backpropagation through time (TBPTT) (Sutskever, 2013) was used by unrolling the network for 20 steps. The final state of LSTM of each batch was fetched into the next batch as the initial state.

Experiment Result

We observed that Global-Context RNN with 2 LSTM layers outperformed other models as showed in Table 4.4. First, by comparing logistic regression model with our best model, the result indicates the strong predictive power of long-term temporal contextual information on the task of detecting social norm violation in dialog. On the other hand, Local-Context RNN model did not achieve significant improvement on overall performance regarding to logistic regression, which means that our learned clause representation through training process has less competence compared to hand-crafted features inspired from linguistic knowledge. One potential reason for such a result could be insufficient amount of training set in order to learn a generic clause representation.

Table 4.4: Performance comparsion for the 3 evaluated models

	Precision	Recall	F-measure
Logistic Regression	0.573	0.583	0.578
Local-Context RNN	0.478	0.747	0.583
Global-Context RNN (1-layer)	0.689	0.696	0.693
Global-Context RNN (2-layer)	**0.690**	**0.720**	**0.705**

4.2.8 Conclusion

In this work, by performing quantitative analysis of our peer tutoring corpus followed by machine learning modeling, we learnt the discriminative power and generalizability of verbal, vocal and visual behaviors from both the speaker and listener, in distinguishing conversational strategy usage.

We found that interlocutors usually accompany the disclosure of personal information with head nods and mutual gaze. When faced with such self-disclosure listeners, on the other hand, often nod and avert their gaze . When the conversational strategy of reference to shared experience is used, speakers are less likely to smile, and more likely to avert their gaze (Cassell et al., 2007). Meanwhile, listeners smile to signal their coordination. When speakers praise their partner, they direct their gaze to the interlocutor's worksheet, smile and nod with a positive tone of voice. Meanwhile, listeners simply smile, perhaps to mitigate the embarrassment of having been praised.

Finally, speakers tend to gaze at their partner and smile when they violate a social norm, without nodding. The listener, faced with a social norm violation, is likely to smile extensively

(once again, most likely to mitigate the face threat of social norm violations such as teasing or insults). Overall, these results present an interesting interplay of multimodal behaviors at work when speakers use conversational strategies to fulfil interpersonal goals in a dialog.

4.3 Predictive Model for Rapport Assessment[5]

4.3.1 Introduction and Motivation

Conversational strategies in our computational model of rapport function to fulfill specific social goals and are instantiated in particular verbal and nonverbal behaviors.Thus, studying the synergistic interaction of conversational strategies and nonverbal behaviors on rapport management is important. In the first section of this chapter we qualitatively examine certain dyadic behavior patterns that benefit or hurt interpersonal rapport. Now, we move forward 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. There are three key issues that we believe should be taken into consideration when performing such assessment.

(1) When the foundational work by (Tickle-Degnen and Rosenthal, 1990) 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, and therefore it is necessary to take into account the relationship status of a dyad when extracting dyadic patterns of rapport. (2) while (Ogan et al., 2012a) 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, (Ogan et al., 2012a) 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 (Gratch et al., 2006, 2007; Huang et al., 2011), 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 section, we take a step towards addressing the above limitations. We employed thin-

[5]This section incorporates text from (Zhao et al., 2016b) which describes a collaboration between Ran Zhao, Tanmay Sinha, and Justine Cassell. My contribution to this work were proposing the idea, implementing the temporal association rule used the toolkit developed by Mathieu Guillame-Bert, extracting temporal association rules, validating temporal association rules by training predictive model for rapport estimation. Also, I was involved in writing and modifying the publication version of the paper.

slice coding (Ambady and Rosenthal, 1992) 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 section addresses these phenomena in the context of peer tutors and intelligent tutoring agents, this work analyzes rapport in the conversational strategy level, which is domain-independent. Thus, our predictive model of rapport could be easily and generally applied to other domains of dyadic interaction.

4.3.2 Related Work

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 (Ramanarayanan et al., 2015) and motif discovery methods (Nakano et al., 2015).

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" (Fujimoto, 2009), 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. Several interesting past work are discussed in (Zhao et al., 2016b).

4.3.3 Study Context

In this study, we conduct our experiment on CMU reciprocal peer tutoring dataset(Yu et al., 2013b), which has been explained in chapter 2.

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). Finally, our temporal association rule framework comprised of nonverbal behaviors like eye gaze (Krippendorf's α= 0.893) and smiles (α= 0.746).

Rapport Annotations

We assessed rapport-building via thin slice annotation (Ambady and Rosenthal, 1992), 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 (Sinha and Cassell, 2015a,b) 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.

4.3.4 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 (Guillame-Bert and Crowley, 2012). 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 $E_{i=1...n}$ and multivariate scalar time series $S_{i=1...m}$, 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 (Guillame-Bert and Crowley, 2012) 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.

$$confidence_r = P((t \in E_1)|(t' \in E_2), t' - t \leq w) \tag{4.1}$$

$$support_r = \frac{\{\#e_2|\text{r is active}\}}{\#e_2} \tag{4.2}$$

$$precision_r = \frac{1}{\text{standard deviation}_r} \tag{4.3}$$

4.3.5 Experimental Results

We first separated out friend and stranger dyads to learn rules from their behaviors. 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.

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 4.3]**

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)*

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 4.4]**

4.3.6 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.

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 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 $R_i = \{r_j^i\}$ 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 $T_i = \{t_k^i \in \mathbb{R}\}$ 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_j^i at the particular time point t_k^i. The extra column represents the indicator function of rapport 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 (Guillame-Bert and Dubrawski, 2014). 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 4.5 suggest that the Titarl-based regression method has a significantly lower mean square

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 4.5: 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: ***:$p < 0.001$, **:$p < 0.01$, *:$p < 0.05$

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.

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 (Goffman, 2005) 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, (Keltner and Buswell, 1997) 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 (Cassell et al., 2007).

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 (Bevacqua et al., 2008).

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. (Laurenceau et al., 1998) 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 (Cassell et al., 2007).

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, (Gratch et al., 2006) 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.

4.3.7 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.

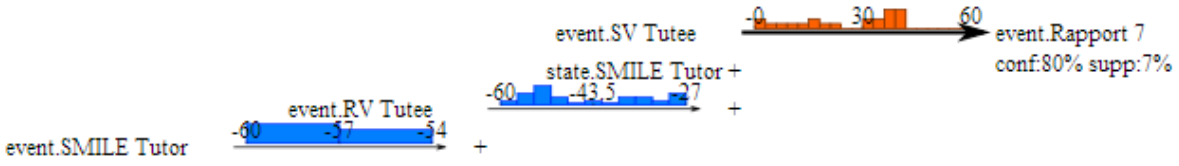


Figure 4.3: Friends in high rapport - The tutee reciprocates a social norm violation while overlapping speech with the tutor, following which the tutor smiles while the tutee violates a social norm.

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;

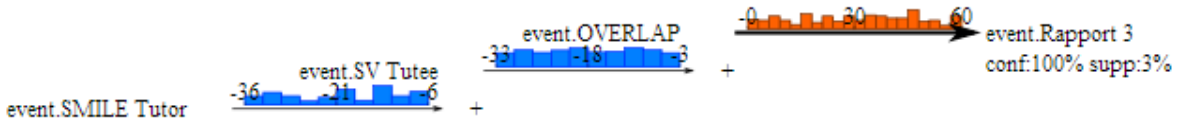


Figure 4.4: Strangers in low rapport - The tutor smiles and the tutee violates a social norm within the next 30 seconds, before their speech overlaps within the next 30 seconds.

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.

Chapter 5

Discourse Planning for Social Dialog [1]

5.1 Introduction and Motivation

Dialog manager is in charge of picking up the next system action respect to the task. Similar, we develop a module named social reasoner that focuses on managing relational bonds with human user through reasoning the usage of the conversational strategy for system. Specifically here we are interested in seven common conversational strategies shown to positively impact rapport (Tajfel and Turner, 1979; Spencer-Oatey, 2008b): Self-Disclosure (**SD**), revealing personal information, to decrease social distance; Question Elicitation of Self-Disclosure (**QESD**), which is used to encourage the other interlocutor to self-disclose; Reference to Shared Experiences (**RSE**), that indexes common history; Praise (**PR**), that serves to increase self-esteem in the listner and therefore interpersonal cohesiveness; Adhere to Social Norm (**ASN**), that increases coordination by adhering to behavior expectations guided by sociocultural norms; Violation of Social Norm (**VSN**), where general norms are purposely violated to accommodate the others behavioral expectations; and Acknowledgement (**ACK**), a way to show that the interlocutor is listening.

Given that rapport-management is a dyadic process, intrinsically involving two individuals, our system must fulfill two critical prerequisites: understanding the *user's* conversational strategy in real-time, and estimating the level of rapport, or relationship strength, at any given moment. The first prerequisite was fulfilled by our trained multimodal *Conversational Strategy Classifier* introduced on last chapter,which has been integrated into our decision-making system. The second prerequisite was fulfilled by our temporal association rule-based *Rapport Estimator*, has been shown to have strong predictive power on rapport estimation in real time, which is also integrated into the social reasoner.

Social Reasoner module is capable of taking input from both the Rapport Estimator and User's Conversational Strategy Classifier described and functioned to reason about how to respond to the social intentions underlying those particular behaviors (such as to raise rapport), and generating appropriate social conversational responses with the system's goal of always keeping rapport high

[1]This section incorporates text from (Romero et al., 2017a) which describes a collaboration between Oscar J. Romero, Ran Zhao and Justine Cassell. My contribution to this work were proposing the idea of using spreading activation model, designing pre-conditions and post-conditions of conversational strategy and conducting statistical test experiments. Also, I was involved in writing and modifying the draft version of the paper.

in order to increase trust and long-term engagement. While there are several potential approaches, most of them are not suitable for our purposes: since the large and increasing number of inputs that the Social Reasoner must process continuously, selecting a proper conversational strategy becomes a combinatorial explosion problem that results almost intractable to solve with a pure symbolic approach such as production rule systems or classic planners. On the other hand, (Romero et al., 2017a) argued that pure sub-symbolic or connectionist approaches fail to semantically express the relationships between inputs, outputs, and negative and positive consequences of triggering a particular conversational strategy. Therefore, we employ a hybrid approach that takes advantage of the features of a classic planner governed by spreading activation dynamics. In fact, the hybrid model proposed by (Maes, 1989) and extended by (Romero, 2011), so-called Behavior Networks, perfectly fits our needs.

5.2 Related Work

Below we will describe related work that focuses on computational modeling decision-making processing in agent to build long-term relationship with human.

(Bickmore and Schulman, 2012) proposed a computational model of user-agent relationship which was inspired from accommodation theory. They defined a set of activities that user is willing to perform with agent. Those activities were described as dialog acts. Their reactive algorithm selected the most appropriate dialog act in order to advance user-agent intimacy. However, the study indicated that their algorithm successfully adapted to user's desired intimacy level but failed to increase intimacy along with the user-agent interaction. As a side note, their system understood user-agent relationship through questionnaire instead of automatically reasoning the real-time closeness level, which was harmful to their decision-making process.

Similarly, (Coon et al., 2013) targeted on developing closeness in human-agent interactions through implementing an algorithm to plan appropriate joint activities. The algorithm modeled the difference between relationship stages from stranger to companion. The decision-making process of this activities planner was based on the required closeness level of each activity while the algorithm optimized its performed activities to achieve user-agent intimacy over time. However, since (Coon et al., 2013) handcrafted specifics activities for each stage, it is a challenge to scale up their algorithm.

Actually, we are not the first ones to propose using a behavior network to model social dialog in human communication. In the past, (Cassell and Bickmore, 2003) constructed a discourse planner that could interleave small talk and task talk during the real estate buyer interview. The conversational moves such as introducing new topic in dialog were planned in order to maximize trust building while pursing task goal of selling real estate. Their implementation utilized activation network to simply adjust agent linguistic behavior - more or less polite, more or less task-oriented, or more or less deliberative, but not for deciding which conversational strategy fitted better during each state of the conversation.

5.3 System Architecture

Using a Global Workspace approach and a spreading activation model, we endow our social reasoner with both short-term and long-term decision-making skills that allow it to reactively

select a proper conversational strategy while deliberatively tailoring a plan (sequence of conversational strategies) in the background. Our purpose here is to motivate and then evaluate the use of this kind of Social Reasoner, which has some specific properties due to its hybrid nature, specifically to a) efficiently make both short-term decisions (real-time or reactive reasoning) and long-term decisions (deliberative reasoning and planning); b) the knowledge is encoded by using both symbolic structures (i.e., semantic-labeled nodes and links) and sub-symbolic operations (i.e., spreading activation dynamics); and c) its network's operation is grounded on cognitive psychological phenomena such as subliminal priming, automaticity with practice, and selective attention, whereas the design of its network's structure relies on observations extracted from data-driven models.

5.3.1 Modules Description

The Social Reasoner's architecture is depicted in figure 6.4. They are described as follows:

1) Working Memory (WM): short-term memory that stores chunks of environmental information (percepts) that are then processed by the Social Reasoner's decision module; 2) Goals: a hierarchy of both task (e.g., generate a recommendation) and social goals (e.g., build rapport); 3) Social Reasoner History (SRH): records of all past decisions (i.e., system conversational strategies); 4)Selective Attention (SA): the most relevant, important, urgent, and insistent information at the moment, which will be selected to be processed by the decision module based on the Global Workspace Theory (Baars, 2003); 5) Action Selection (AS): this module is in charge of choosing a conversational strategy as a consequence of the decision-making dynamics. This module is implemented as a Behavior Network (originally proposed by (Maes, 1989) and extended by (Romero, 2011)). 6) Learning Processing (LP): this module is responsible of adapting the system parameters in real-time. However, this is part of our future work so we will not go into further details; 7) Other Modules: Social Reasoner interfaces with other modules that are commonly used in dialog Systems and conversational agents, such as ASR, NLU, NLG, etc.

Social Reasoner's Decision module is crafted as a network of interacting nodes where decision-making emerges from the dynamics of relationships among those nodes.

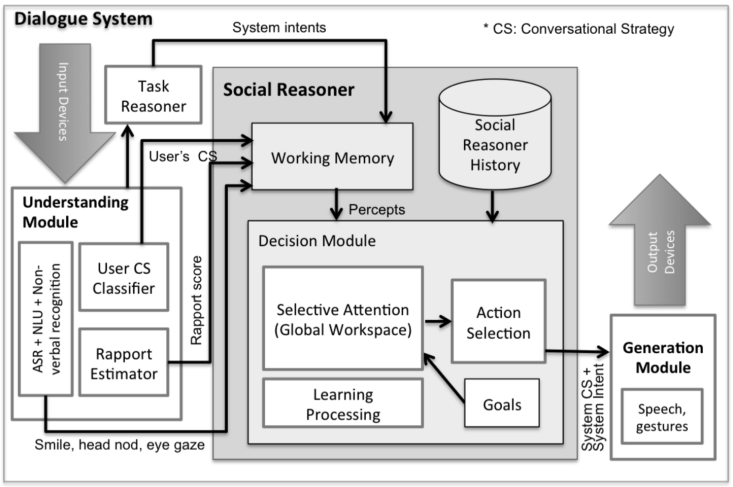


Figure 5.1: System Architecture

5.4 Computational Model

In the following, we will provide details of our Behavior Network formalism.

A Behavior Network (BN) is a spreading activation model proposed by (Maes, 1989) as a collection of competence modules which works in a continuous domains. Behavior selection is modeled as an emergent property of activation/inhibition dynamics among all behaviors. A behavior i can be described by a tuple $< c_i, a_i, d_i, \alpha_i >$ where c_i is a list of pre-conditions which have to be fulfilled before the behavior can become active, a_i and d_i represent the expected (positive and negative) effects of the behavior's action in terms of an add list and a delete list. Additionally, each behavior has a level of activation α_i. If the proposition p about environment is true and p is in the pre-condition list of the behavior i, there is an active link from the state p (proposition about environment) to the behavior i. If the goal g has an activation greater than zero and g is in the add list of the behavior i, there is an active link from the goal g to the behavior i. Internal links include predecessor links, successor links, and conflicter links. There is a successor link from behavior i to behavior j for every proposition p that is member of the add list of i and also member of the pre-condition list of j A predecessor link from behavior j to behavior i exists for every successor link from i to j. There is a conflicter link from behavior i to behavior j for every proposition p that is a member of the delete list of j and a member of the pre-condition list of i. The following is the procedure for decision-making:

1. Calculate the excitation coming in from environment.
2. Spread excitation along the predecessor, successor, and conflicter links, and normalize the behavior activations so that the average activation becomes equal to π.
3. Check any executable behaviors, choose the one with the highest activation, and execute it. A behavior is executable if all the pre-conditions are true and if its activation is greater than the global threshold. If no behavior is executable, reduce the threshold and repeat the cycle.

Additionally, the model defines five global parameters that can be used to tune the spreading activation dynamics: π is the mean level of activation, θ is the threshold for becoming active which is lowered each time none of the modules could be selected and reset to its initial value otherwise, ϕ is the amount of activation energy a proposition that is observed to be true injects into the network, γ is the amount of energy a goal injects into the network, and δ is the amount of activation energy a protected goal takes away from the network.

One important contribution made to the original Behavior Networks model is that we use a "partial matching" approach rather than a strict "full matching" approach; that is, the original model states that a behavior is activated only when all its pre-conditions are true, which works well when using discrete variables, however, we deal with continuous variables in a frequently-changing environment, so behaviors are almost never activated under these conditions. We propose the definition of categories to group sets of well-defined pre-conditions with something in common. An inclusive OR operator is used to evaluate intra-category pre-conditions and an AND to evaluate inter-category pre-conditions, that is, there must be at least one pre-condition per category that is true. This scheme is much more flexible and allows more combinations of pre-conditions that can trigger a particular behavior.

In our model, each behavior corresponds to a specific conversational strategy (e.g., SD, PR, VSN, etc.) where pre-conditions are divided into categories, as shown in table 5.1, and post-conditions are defined in terms of what the expected states are after performing the current

Table 5.1: Pre-condition and Post-condition Categories

Category	**Pre-conditions and Post-conditions**
Rapport level	low, medium, and high
Rapport delta	decreased, maintained and increased
System and User conv. strategies	asn, vsn, sd, qesd, se, ack, pr, not–asn, not–vsn, not–sd, not–qesd, not–se, etc.
User non-verbals	gaze–elsewhere, gaze–partner, head–nod, smile, etc.
dialog history	number–of–turns, sd–user–history, pr–system–history, qesd–user–history, etc.
System intent	greeting, do–goal–elicitation, start–interest–elicitation, start–recommendation, do–recommendation, end–recommendation, farewell, etc.

conversational strategy (e.g., rapport score increases, user smiles, etc.). This kind of chaining reasoning based on linked pre-conditions and post-conditions endows the system with planning ahead capabilities. Intuitively, the Social Reasoner can tailor a deliberative plan as the aggregation of nodes connected through both predecessor and successor links, for instance, when a conversation starts the most likely sequence of nodes could be: <ASN, ASN, SD, PR, SD ...VSN ...>, that is, initially the system establishes a cordial and respectful communication with user (ASN), then it uses SD as an icebreaking strategy(Altman and Taylor, 1973), followed by PR to encourage the user to also perform SD. After some interaction, if the rapport level is high, a VSN is performed. Coalitions are created between nodes, so ASN would spread forward some energy to SD, and SD would spread backward some energy to ASN, and the same between SD and PR, and between PR and SD, etc. Inhibitory links avoid wrong conversational strategies to be triggered. The Social Reasoner is adaptive enough to respond to unexpected users actions by executing a reactive plan that emerges from forward and backward spreading activation dynamics as well as from the network's parameters configuration that determines the global system's behavior, for instance, it can make the system more goal-oriented vs. situation-oriented, more adaptive vs. biased to ongoing plans, more thoughtful vs. faster.

5.5 Design of the Decision-Making module

5.5.1 Sources of Information

As is clear from the description above, the nature of the pre-condition and post-conditions is key to the functioning of the systems. We extracted information for these conditions from two sources: theoretical and empirical data.

Theoretical Sources

Rapport Theory: Based on our proposed computational model of rapport in Chapter 2, at the beginning of the interaction, one tends to be tentative and polite, adhering to social norms. Initiating a self-disclosure at this stage will both signal attention and elicit self-disclosure from

the interlocutor which, in turn, enables both parties to gradually learn each other's behavioral expectations. During this stage of interaction, praise can boost self-esteem and motivate the interlocutor to diminish social distance. Thus, adhering to social norms, self-disclosure and praise are three trending conversational strategies in the early stage of communication. As the interaction proceeds, interlocutors have more interpersonal knowledge to guide their behavior. They refer to shared experience to index commonality and purposely violate social norm in order to accommodate each other's behavioral expectations, and signal that they are now outside the phase of pure politeness.

Norm of Reciprocity: Reciprocity of behavior (Burger et al., 2009) plays an important role in increasing coordination between interlocutors. Our annotations of conversations revealed that most of the conversational strategies described here are used reciprocally (referring to shared experience evokes the same behavior from one's conversation partner). Thus, one pre-condition for praise is that the user hasn't praised in the previous turn.

Data-driven Sources

Data-driven discovery by temporal association rule: (Zhao et al., 2016b) applied a data mining algorithm to separately learn behavioral rules for friends and strangers. In our Social Reasoner, we input *phase* of interaction (early, middle, late) as a variable. Early stages of the interaction were determined by rules learned from the stranger data, later stages by friend rules. For instance, a rule that strangers followed was: *one of the interlocutors smiles while the other gazes at the partner and begins self-disclosing*, so we defined smile as one of a set of optional pre-conditions for self-disclosure.

Data from Wizard-of-Oz study: We collected data from 228 English-speakers interacting with a virtual assistant acting as a guide that recommends sessions to attend and people to meet at the conference. In each session, a dyad consisting of a user and the virtual assistant (using a Wizard of Oz protocol) interacted through a dialog system interface for around 8-10 minutes. During conversation, the agent elicited the users interests and preferences and used these to improve its recommendations. The user's verbal and non-verbal behaviors were recorded by the system while the woz-er picked the next utterance for the agent depending on the user's utterance, the current task and goal, as well as the WoZer's assessment of most appropriate conversational strategy to build rapport. After conducting the study, only those decisions made by the woz-er that had a significant impact on building rapport (i.e., increasing rapport) and raising engagement (defined here as increase conversation length) were taken into account.

5.5.2 Encoding of Pre-conditions & Post-conditions

(Romero et al., 2017a) modeled a Behavior Network with seven behaviors (one for each conversational strategy). Their pre-conditions and post-conditions were designed by following a two-way tuning process: initially, for each behavior, we identified a sub-set of precondtions and post-conditions (from table 5.1) based on the theoretical foundations provided in section 5.5.1; then we validated the previous model through the empirical analysis of data obtained from the Wizard-of Oz study. For the latter process, we ran a feature selection statistic analysis, more specifically, a bidirectional elimination stepwise regression that allowed us, through a series of

partial F-test, to include or drop candidate variables from each behavior. This process helped us to discover which sub-set of variables and features should be considered as pre-conditions and post-conditions for each behavior because of their impact and significance. For instance, the theoretical foundation guided us to identify a sub-set of pre-conditions for PR as follows: `<low-rapport, not-pr-user, not-pr-history-user, ...>` however, the stepwise regression analysis told us that we need to include at least three more pre-conditions: `<high-rapport>` (F: 95.7, p-value: 0.00), `<gaze-elsewhere>` (F: 56.8, p-value: 0.00002) and `<rapport-increased>` (F: 17.6, p-value: 0.00073); and remove pre-condition `<not-pr-history-user>`) (F: 3.4, p-value: 0.005) to improve the accuracy on conversational strategy prediction. An excerpt of the final tuned behaviors' pre-conditions and post-conditions is shown in appendix.

5.5.3 Spreading Activation Parameters:

Following the guidelines proposed by (Romero, 2011; Romero and de Antonio, 2012) and through empirical analysis, we determined that the best configuration of the spreading activation parameters is as follows:

1. To keep the balance between deliberation and reactivity, $\phi > \gamma$, so $\phi = 68$ and $\gamma = 42$.
2. To keep the balance between bias towards ongoing plan vs. adaptivity, $\pi > \gamma \wedge \pi < \phi$, so $\phi = 50$.
3. To preserve sensitivity to goal conflict, $\delta > \gamma$, so $\delta = 75$.

5.6 Experimentation and Results

Our experiments focused on evaluating three aspects of our work: 1) determining whether social reasoning can increase rapport and raise engagement; 2) evaluating the degree of effectiveness and accuracy of the Social Reasoner after the data-driven tuning process; and 3) evaluating the performance of the Social Reasoner during interaction with users.

5.6.1 Experiment 1: Social Reasoning validity

H_0 : *Social Reasoning doesn't contribute significantly to build rapport and increase conversational engagement in comparison with traditional dialog systems.*

For this experiment we divided the WOZ study dataset of 228 sessions (section 5.5.1) into two groups: dialog turns that used conversational strategies and dialog turns that did not use any conversational strategy (plain behavior). Then, we observed the rapport score (1-7), our variable of interest. We ran an one-way ANOVA analysis in order to find out whether there is a statistically significant difference between the groups at $p < .05$. The ANOVA is shown in table 5.2.

Table 5.2: ANOVA for Experiment 1.

Sc. of Variation	**df**	**SS**	**MS**	**F**	**p**
Between groups	2	1012398	687297.4	4.52	0.007%
Withing groups	154	1672037	293898.8		
Total	156	2684435			

Since p is less than .05 we can conclude that there is a statistically significant difference between the two groups. A Tukey post-hoc test revealed that rapport scores of the group that uses social reasoning was higher ($5.65 \pm 0.4, p = .032$) in comparison with the group that uses a traditional approach – no social reasoning – ($3.17 \pm 0.6, p = .028$) and therefore we can reject the null hypothesis H_0 that social reasoning doesn't contribute significantly to build rapport. Likewise, we conclude that using social reasoning may improve social bonds (rapport) on a 35.4% during a conversation.

5.6.2 Experiment 2: Social Reasoner's accuracy

H_0 : *Data-driven tuning process doesn't improves Social Reasoner's accuracy*

For this experiment we used the WOZ study dataset as a ground truth. Then we ran a simulation for all 228 sessions, where system inputs were signals from the understanding module, the task reasoner, and the history databases; and outputs were the conversational strategies picked by the woz-er. Then, we compared each woz-er output with the social reasoner's output for two different scenarios: before tuning the decision-making module (i.e., using only a theoretical-driven design) and after tuning (i.e., using both a theoretical and data-driven design). We ran an one-way ANOVA analysis and results are shown in table 5.3.

Since p is considerably lower than α, we can conclude that there is a statistically significant difference between the two groups. A Tukey post-hoc test revealed that rapport scores of the group that received a data-driven tuning was higher ($4.83 \pm 0.5, p = .027$) in comparison with the group that only used a theoretical-based design ($3.05 \pm 0.4, p = .033$) and therefore we can reject the null hypothesis that data-driven tuning doesn't improve the Social Reasoner's accuracy. Also, we conclude that using a data-driven tuning process along with a theoretical-driven design may improve the accuracy of Social Reasoner up to a 25.4%.

Table 5.3: ANOVA for Experiment 2.

Sc. of Variation	**df**	**SS**	**MS**	**F**	**p**
Between groups	4	2984714	873394.3	5.34	0.005%
Withing groups	173	3439465	363797.8		
Total	175	6424179			

5.6.3 Experiment 3: Social Reasoner's performance

For this experiment we chose four well-characterized conversational sessions from the dataset log files in the post-experimental evaluation to test the system's performance. Below is the description of each one:

Flat User Scenario (FU): user's verbal and non-verbal behaviors remain the same during conversation, e.g., rapport level is medium all the time, no smile, and user's conversational strategy is ASN most of the time.

Incremental Engagement Scenario (IE): user is getting more engaged in conversation over time, e.g., rapport level increases gradually, user smiles more often, and user's conversational strategy is mostly SD and VSN.

Low Rapport Scenario (LR): during most of the conversation user keeps a low rapport level, no smiles and barely makes eye contact.

Losing Interest Scenario (LI): initially, user is very engaged during conversation (i.e., high rapport, a lot of smiles and eye contact, user's conversational strategies are SD and VSN, etc.) but gradually he is losing interest.

Table 5.4: Social Reasoner's performance. MSE: Mean Square Error, MSE Rate: [1 - ($MSE_{SR} \div MSE_{TD}$)]

Scenario	**Std Dev**	**MSE**$_{TD}$	**MSE**$_{SR}$	**MSE Rate**
FU	0.83	1.31	0.86	34.35%
IE	0.73	2.12	1.68	20.75%
LR	0.52	0.96	0.68	29.16%
LI	0.93	1.54	1.05	31.81%

Table 5.4 shows the statistical data for experiment 3. The MSE for each scenario is the mean square error of 20 turns, where an error is considered as a drop on the rapport score as consequence of activating the wrong conversational strategy. The MSE rate presents the performance relationship between MSE_{TD} (a traditional dialog system that doesn't use conversational strategies) and MSE_{SR} (a dialog system that uses our Social Reasoner). It is important to notice that, for the experiments executed, the proposed Social Reasoner model improves the performance results obtained by a traditional dialog system a rate between 20% and 34%.

It is worth mentioning that having the highest activation level is not the only criteria to chose a particular conversational strategy (CS), but it must be also executable and its activation level must be over the threshold, otherwise, the next CS which meets those conditions will be chosen.

Intuitively, one can deduce that the Social Reasoner emergently tailors a plan as the combination of SD, PR and QESD strategies when detecting the user is not engaged during interaction as expected (e.g., in LR and LI scenarios). Conversely, VSN is avoided when trying to recover both user's attention and interest, and also his rapport level is low (as at the end of LR, and in FU). On the other hand, reactive decisions such as using VSN or RSE are made when the system detects the user is more receptive to this kind of strategies, even if they are not the ones with the highest activation level. ACK is more likely to appear when there is evidence of progressive raising of user's engagement, since conversational strategy such as ASN, SD and RSE spread more activation energy forward and backward to it. Also, it is interesting to see how ASN is activated at an early stage of the conversation (e.g., IE scenario) but remains accumulating energy during the whole interaction so it can be easily triggered if the system realizes that a previous action (as consequence of using a particular CS) causes a diminishing on the rapport level. Finally, PR is continually used when the Social Reasoner detects no significant changes on user's verbal and non-verbal behaviors that can raise rapport, specially when other conversational strategy such as SD and QESD have been used without success.

5.7 Conclusion

We proposed a hybrid adaptive Social Reasoner component that determines which conversational strategy should be used in order to build and maintain rapport with a user. The Social Reasoner

interacts with several modules that can be connected and disconnected while its behavior remains robust. A spreading activation approach was merged with classic planner features and extended to allow the system to partially match pre-conditions by using an OR operator rather than the conventional AND operator, and as a consequence expanding the number of possible combinations between matched pre-conditions and triggered conversational strategies.

Chapter 6

Proposed Work: A Social Intelligent Negotiation Dialog System (SOGO)

6.1 Introduction and Motivation

Whether we are deciding between a salad or fast food for lunch or asking a coworker to complete a project ahead of schedule in exchange for help at a later date, people negotiate every day. When we hold conflicting interests, we must negotiate to pursue our ultimate goals. Thus, negotiation an act wherein participants with unique motives cooperate and compete to maximize their own benefits. Virtual agents are a powerful tool for teaching negotiation skills and modeling negotiation in an agent shows great promise in a variety of domains (Guttman and Maes, 1998). Indeed, many current empirical studies are making progress in this area (Mell and Gratch, 2017; Konovalov et al., 2016; Faratin et al., 1998; Hindriks et al., 2009).

Drawing on this work, we recognize that negotiation is both a challenging reasoning problem as well as linguistic problem. Although people are adept at navigating the trade-off between cooperation and competition, algorithms have yet to develop such reasoning and linguistic fluency. Therefore adversarial agents, such as AlphaGo, which aim to beat interlocutors in a zero-sum game, cannot work because they privilege competition, not cooperation. Similarly, a personal assistant like Alexa or Siri, which privileges cooperation, overlooks competition. Futhere, most studies regard negotiation as a sole reasoning or planning problem, like searching for optimized outcomes and thus aim to sharpen the agent's tactics. However, negotiation also relies on appropriate language to maintain relationships and optimize a plan. Prior research in human-human negotiation (Nadler, 2003; Kong et al., 2014) has shown that social factors such as trust and rapport underlie both challenges. In this study, we leverage different linguistic devices to build rapport between a human and a dialog system, which thus fosters integrative agreement during negotiation.

Negotiators are encouraged to share crucial information and cooperate to reduce the risk of impasse because of a sense of *rapport*, a feeling of connection and closeness with another (Bronstein et al., 2012b; Nadler, 2003). Also, (Curhan et al., 2006) shows that rapport helps to formalize a negotiator's intuition about objective outcomes and predict future objective value. Social scientist Spencer-Oatey (Spencer-Oatey, 2008b,a) explains the experience of establishing

rapport according to three interrelating components: face, social rights and interactional goals. People use a variety of strategies to manage these three factors, which have been categorized into five major domains: illocutionary domain, discourse domain, participation domain, stylistic domain and non-verbal domain. Most recent studies of rapport agents (Gratch et al., 2006; Huang et al., 2011) have investigated the importance of the non-verbal domain in human-agent interaction. However, we are interested in examining how the verbal channel produces rapport in the context of human-agent negotiation. Previously, (2012b) employed a hierarchical linear model to validate and quantify the contribution of the verbal channel in rapport management.

We believe that introducing social conversation into complex negotiation communication will help establish and maintain rapport while facilitating negotiation. To this end, we propose a two-phase method in our computation model of negotiation: the task phase and the social phase. The task phase generates the next system task intention/move (e.g., to request a book). The social phase provides opportunities for social intentions/moves (e.g., self-disclosing a personal preference) realized by different conversational strategies. Conversational strategies are units of discourse that are larger than speech acts, which have been demonstrated to contribute to building, maintaining or even destroying interpersonal (or human-agent) bonds (Romero et al., 2017b; Zhao et al., 2014b). This social phase is inspired by the work(Mattar and Wachsmuth, 2012; Bickmore and Cassell, 2001), who exploits structures of casual conversation in human-human communication to improve the system's capabilities of generating task-related social moves that are more than idle chit-chat. Its major function is to prepare the interlocutor for the next negotiation moverather than increase familiarity between interlocutorsby discussing some restricted common topics like the weather. For instance, the system discloses its personal interest in reading as the current social move (e.g., "I love reading"), which anticipates its next negotiation move of requesting the book (e.g. "Can I have all the books?")

Our work does not seek to develop negotiation tactics but rather to use rapport management strategies to facilitate the negotiating process. Thus, we do not develop our own negotiation reasoning module but adopt the end-to-end negotiation dialog model from Facebook AI Research (FAIR) group as our task reasoning backbone. FAIR has demonstrated the advancements of their neural dialog models on the multi-issue bargaining task. Because this neural dialog model allows users to negotiate with the system through natural language, we are able to insert social moves into dialogs. Further, FAIR leveraged reinforcement learning to maximize the outcome of negotiation rather than imitating the actions of human users in supervised learning. Finally, FAIR replaced traditional likelihood decoding with a rollout algorithm, a kind of beam search conducted in the dialog turn level, which allows the model to generate more effective negotiation strategies. Despite these advantages, FAIR suffers from low agreement with human users because of users' unwillingness to cooperate with such an uncompromising opponent. We believe that combining the negotiation tactics of the Facebook end-to-end model (FB) and our proposed rapport strategies will allow our social intelligent negotiation dialog system (SOGO) to overcome these limitations and improve task performance while constructing an interpersonal bond with humans. To compare outputs from both systems, refer to the sample dialogs in Table 6.1.

In the next section we review prior studies related to negotiation and rapport agents. Then, section 6.2 briefly describes the Facebook negotiation corpus. Sections 6.3 and 6.4 introduce our computational model of negotiation and provide an overview of the system architecture. To validate our model, we explain the set up of our experiment in section 6.5. Finally, in section 6.5.1,

we combine subjective metrics and objective metrics to evaluate the system's performance in task negotiation rapport building.

FB System
Agent: *I'll take the book and balls if you want the book.*
User: *I need all the balls and the hat*
Agent: *You can have the hat if i can have the rest*
User: *I prefer all the balls and you can have the rest*
Agent: *deal*

SOGO System
User: *Since you really like the ball last time, I am giving 2 to you.* I get the hat and the books. Deal?
Agent: *Thank you so much!* *I guess the book must be very important to you* *. Shall I give you the the balls and 1 book?*
User: *You get the balls I get 2 books and the hat. Deal?*
Agent: *Why are you so good at this?! Seriously! ball and hat! Great!* *..., can I have the one ball and one hat please?*
User: *okay I guess.*
Agent: *Playing with you is making me a smarter chatbot.*

Table 6.1: Sample human-agent negotiation dialog of the FAIR model (top) and our SOGO model (bottom)

6.2 Study Context

Facebook AI Group has collected a human-human negotiation dialog corpus (n=6118 human-human dialogs (Lewis et al., 2017) using Amazon Mechanical Turk. Pairs of participants completed a classic multi-issue bargaining task(Fershtman, 1990): They negotiated to divide items from three item-types (books, hats, balls) in a pool of five to seven total items. Participants were each given different and unseen value functions. The maximum value to each side is 10. The unique value function for each item was constrained to ensure that both participants could not receive its maximum value. Table 6.2 shows the statistics of the negotiation corpus.

Metric	Dataset
Train/Valid/Test	5048/544/526
Success/Fail	4801/1317
Average Turns per dialog	6.6
Average Words per Turn	7.6

Table 6.2: FAIR Negotiation Corpus Stats

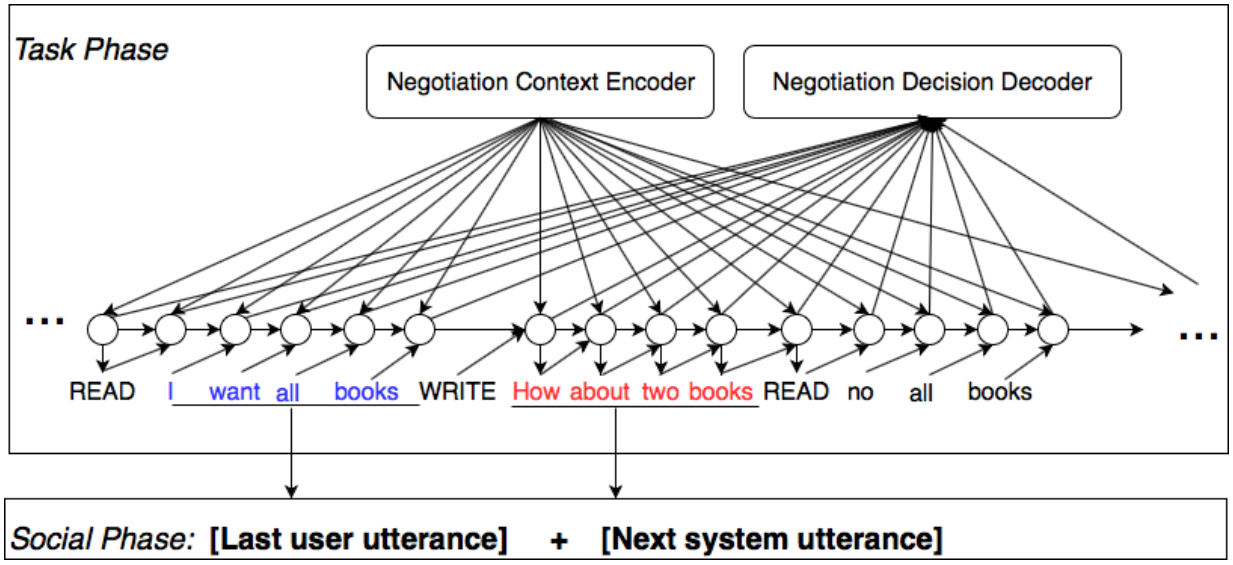


Figure 6.1: Overview of Two-Phase method

6.3 Computational Model

In the following, we outline the details of our two-phase formalism in a computational model of negotiation. First, the task phase and the social phase are performed sequentially. As described above, our task phase adopts the Facebook end-to-end negotiation dialog model (Lewis et al., 2017) which decides the system's next-task utterance. To make the system seem more human, in the social phase, our model displays all eligible strategies and realizes them into utterances that concord with former task utterances. These concordances are based on deep understand of users' prior utterances and the system's next-task utterances. Figure 6.1 demonstrates the overview of two phase method.

6.3.1 Task Phase: end-to-end dialog model

(Lewis et al., 2017) utilize two-stage learning strategies by pre-training the model with supervised learning, then fine-tuning the parameter using reinforcement learning. In this section, we briefly discuss their advanced reinforcement learning with dialog rollouts decoding based model which we deploy in our study (c.f. (Lewis et al., 2017)). Each dialog D is represented as a set of token x_t where the total number of tokens are T. Tokens are segmented by two special tokens WRITE and READ which indicates turn-taking between human and agent. The agent has an input goal g and generates the outcome of the negotiation o_i.We keep the structure of their four GRU-based recurrent neural networks: GRU_g(Agentś goal encoder), GRU_w(dialog token generator),$GRU_{\overrightarrow{o}}$(forward output encoder),$GRU_{\overleftarrow{o}}$(backward output encoder). In the first stage of supervised pre-training, given the word embedding W, (Lewis et al., 2017) firstly models the dependencies between language and input goals with the function (6.1):

$$p_\theta(x_t|x_0...t-1,g)) \tag{6.1}$$

Conditioning the input goals and generated dialog, they predict negotiation outcomes with the function (6.2)

$$p_\theta(o_i|x_{0...T}, g)) \tag{6.2}$$

Thus, the objective function in the supervised learning stage can be represented as:

$$L(\theta) = -\sum_{x,g}\sum_{t} logp_\theta(x_t|x_0...t-1, g)) - \alpha\sum_{x,g,o}\sum_{j} logp_\theta(o_i|x_{0...T}, g)) \tag{6.3}$$

α is a hyperparameter to balance token prediction loss and outcome prediction loss. Based on the negotiation outcome at the end of each dialog, the agent receives a reward $r(o)$. In the second stage of reinforcement learning, given the discounter factor as γ and a running average of completed dialog rewards μ, the objective is to optimize the expected reward of each token generated by the agent as follows:

$$L_\theta^{RL} = \mathbb{E}_{x_t\sim p_\theta(x_t|x_0...t-1,g))}[\sum_{x_{t...T}} \gamma^{T-t}(r(o) - \mu)] \tag{6.4}$$

In the decoding part, dialog rollout algorithm (Lewis et al., 2017) generates a small set of candidate utterances $U = \{u_i|u_i = x_{n,n+k}\}$ and chooses the utterance that maximizes the expected reward, with the following function:

$$u* = \underset{u_i=x_{n,n+k}, u_i\in U}{argmax} (\mathbb{E}_{x_{(n+k+1...T;o)}\sim p_\theta}[r(o)p_\theta(o|x_{0...T})]) \tag{6.5}$$

Finally, both $u*$ and current turn user input utterances are sent to our social phase.

6.3.2 Social Phase

Our social phase transforms task utterances by introducing social moves. To effectively plan these moves, we need two types of information stored in our defined conversation state: (a) user/agent task intention/move and (b) user/agent model (e.g. personal preference, dialog history and context). Based on this information, our social language generator selects an appropriate conversational strategy to realize social moves. To obtain that information on the fly and generate social language from it, we first apply a text classification method to understand task intention from the utterance. Secondly, we use traditional information extraction to construct a user/agent model. Finally, drawing on socio-psychology theory, we defined nine conversational strategies, which have a property of pre-conditions, from which we determine eligibility of specific conversational strategies given the conversational state. Meanwhile, we deploy emoticons to our generated sentences as indications of the illocutionary force in the textual utterances that they accompany.

Intention Recognition

In our study, understanding user and agent task intention/move is the baseline for transforming the task utterance into conversational strategy. Based on our definition, each task intention/move consists of one speech act (e.g., Request, Offer) and one or several affiliated entity mentions (e.g., two books). We leverage vector-based text representation to build a speech act classifier and utilize a keyword matching algorithm to extract the entities mentioned in the sentence. Following

Speech Act	Precision	Recall	F1
Request	0.922	0.935	0.928
Reject	0.824	0.590	0.688
Accept	0.826	0.858	0.842
Elicit preference question	0.776	0.422	0.547
Offer	0.913	0.859	0.885

Table 6.3: Performance Evaluation of our speech act classifier

these, the challenge might be multiple intentions in one utterance. For instance, "If you give me the ball, I will give you the book and two hats" refers to both Offer and Request. Thus, it is difficult for us to link the entity mentioned to its affiliated speech act. Our solution was to utilize the Stanford CoreNLP toolkit (Manning et al., 2014), which breaks the utterance into separate clauses with the smallest grammatical unit that can express a complete proposition, before training our speech act classifier. In this way, we guarantee that each clause includes only one intention. Both the human annotation and the trained classifier below are in the clause level.

Speech Act Annotation Based on empirical studies of human-agent negotiation dialogs(Konovalov et al., 2016; DeVault et al., 2015; Gratch et al., 2015), we discovered the five speech acts most closely related to rapport/face management and being widely used in negotiation.

- Elicit preference question: ask questions about the opponent's preferences that gain maximal information.
- Request: request a subset of items from the opponent.
- Offer: offer a subset of items to the opponent.
- Reject: reject the previous offer, in whole or in part.
- Accept: accept the previous offer, in whole or in part.

The annotation work was conducted on the Amazon Turk platform. Six out of ten MTurkers passed the qualification test, i.e., completed previous tasks with more than 80% accuracy. 2,500 dialogs were annotated and used to train our speech act classifier, which served to annotate the rest of the corpus.

Speech Act Classifier We leverage the sentence classification library fastText to train our supervised speech act classifier. fastText is essentially an extension of the word2vec model, which treats each word as a composition of character n-grams. We have set n as in the range between 3 and 6. We chose the fastText toolkit for two reasons: (1) The Facebook negotiation corpus is domain-specific with a small vocabulary. Out-of-vocabulary presents a considerable problem. fastText generates better word embeddings for rare words and even those out-of-vocabulary words since it constructs the embeddings in the character-level. (2) fastText is memory-consuming: the number of n-grams in the character-level grows exponentially with the growth of corpus size. Since we have a small vocabulary in the corpus, this is not the case.

User Model and Agent Model

Both user model and agent model contain the dialog history and context across the interactions that serve as long-term memory in human-agent interaction. Meanwhile, this memory offers the dialog content for specific conversational strategies (e.g., reference to shared experience) that could index their built relationship. User model and agent model share most parts of the schema: (1) preferences, (2) historical game results (e.g. scores, deal items, game context), (3) speech act sequences, and (4) sentiment sequences. Besides, agent model also includes the conversational strategy sequences. In order to obtain this key information in real-time, we developed a syntactic-based preference extractor and utilize the off-the-shelf rule-based sentiment classifier (Gilbert, 2014).

Preference Extractor In the clause breaking process, Stanford CoreNLP pipline (Manning et al., 2014) generated a dependency tree of each clause as one of the intermediates. Thus, we wrote several subject-verb-object (SVO) templates to extract user preference on the dependency tree.

Social Language Generator

In this part, we adopt a theory-driven template-based approach to generate social moves. As we discussed above, there are five domains of rapport management strategies based on the Spencer-Oatey's theory (Spencer-Oatey, 2008b,a). (Zhao et al., 2014b) advanced this theoretical framework to propose a computation model of rapport that explains how humans in dyadic interactions build rapport over time through conversational strategies. Specifically, (Zhao et al., 2014b) find four major conversational strategies that positively impact rapport: Self-Disclosure (**SD**), revealing personal information to decrease social distance; Reference to Shared Experiences (**RSE**), which indexes common history; Praise (**PR**), which increases self-esteem in the listener and therefore raises interpersonal cohesion; and Violation of Social Norm (**VSN**), where general norms are purposely violated to accommodate the other's behavioral expectations. However, the authors studied peer tutoring-a scenario that elicits far fewer face-threatening speech acts, such as requests or rejections, than do negotiation dialogs. Thus, most conversational strategies proposed in (Zhao et al., 2014b) belong to the discourse and stylistic domains, not the illocutionary domain. We add speech act strategies in the illocutionary domain which could boost politeness and appropriately address face-threatening speech acts. Concretely, based on (Blum-Kulka and Olshtain, 1984; Eisenstein and Bodman, 1986; Beebe and Takahashi, 1989), we include **Request**, **Reject**, **Gratitude**, **Greeting** and **Closing** strategies, each of which contains several sub-categories. For instance, head act is a core part of a request sequence. We tried to mitigate its face-threatening effect through different supportive moves: (1)*Preparator:* "I'd like to ask you something..." (2)*Grounder:* "I missed my book so much" (3)*Promise of reward:* "I will give you all the books in the next game." (4)*Imposition downgrader:* "Could you please give me the ball if you are not playing with it now?". We acquired several variations in sentence realization for each sub-category by hiring two native English writers from the Fivrr website. Table 6.4 shows some examples.

Notice that some templates are designed for a specific negotiation entity (in red) and others are more general (in green).

Strategies	Sub category	Realization
SD	Inner state	You know what, I really love reading.
RSE	Preference	Books are for you since you said you love reading last time
PR	Interaction	Negotiating with you is such fun.
VSN	Teasing	You messed up my thinking my friend....
Request	Grounder	Tomorrow is my creator's birthday and I do not get time to buy him a gift. Could you please give me the books for him?
Reject	Conditional	If you'd told me earlier, I could have given you the books.
Gratitude	Appreciation	You are a life savor!
Greeting	Friend	It is always a pleasure to play with you
Closing	SD Closing	Besides the game, I look forward to getting know you better

Table 6.4: Strategy Realizations

Category	Pre-conditions
System and User Speech Act	Request, Reject, Greeting, Closing, Gratitude
Sentiment	positive and negative
Dialog History	time of interaction, number of turns, historical game results
Entity	book-slot, hat-slot, ball-slot
System conv.strategies	SD, RSE, VSN, PR

Table 6.5: Preconditions

Preconditions Each strategy contains several pre-conditions that decide eligibility of usage given the current conversational state. In our model, pre-conditions are divided into categories, as shown in table 6.5.

Emoticons Since visual access between participants in this study was limited, we substituted non-verbal cues with emoticons. Emoticons are generally accepted as non-verbal indicators of emotions that map directly onto facial expressions (Rezabek and Cochenour, 1998; Walther and DAddario, 2001), yet they also indicate the illocutionary force of an utterance (Dresner and Herring, 2010). They do not contribute to the propositional meaning of a sentence but construct a context in pragmatics for the text. For instance, using a smile emoticon when violating social conversational norms signals joking or teasing (Zhao et al., 2016c), which can significantly enhance interpersonal rapport between friends (Ogan et al., 2012b). Following (Ekman and Friesen, 1986), who reveal that humans have six basic emotions, we provide six emoticon types: Happy, Sad, Fear, Anger, Surprise, and Disgust. Each type has two to three variants.

6.4 System Architecture

Currently, our SOGO (Social Intelligent Negotiation dialog System) system is in the Wizard-of-Oz (WOz) setup. Below we detail the architecture that operationalize our computational model,

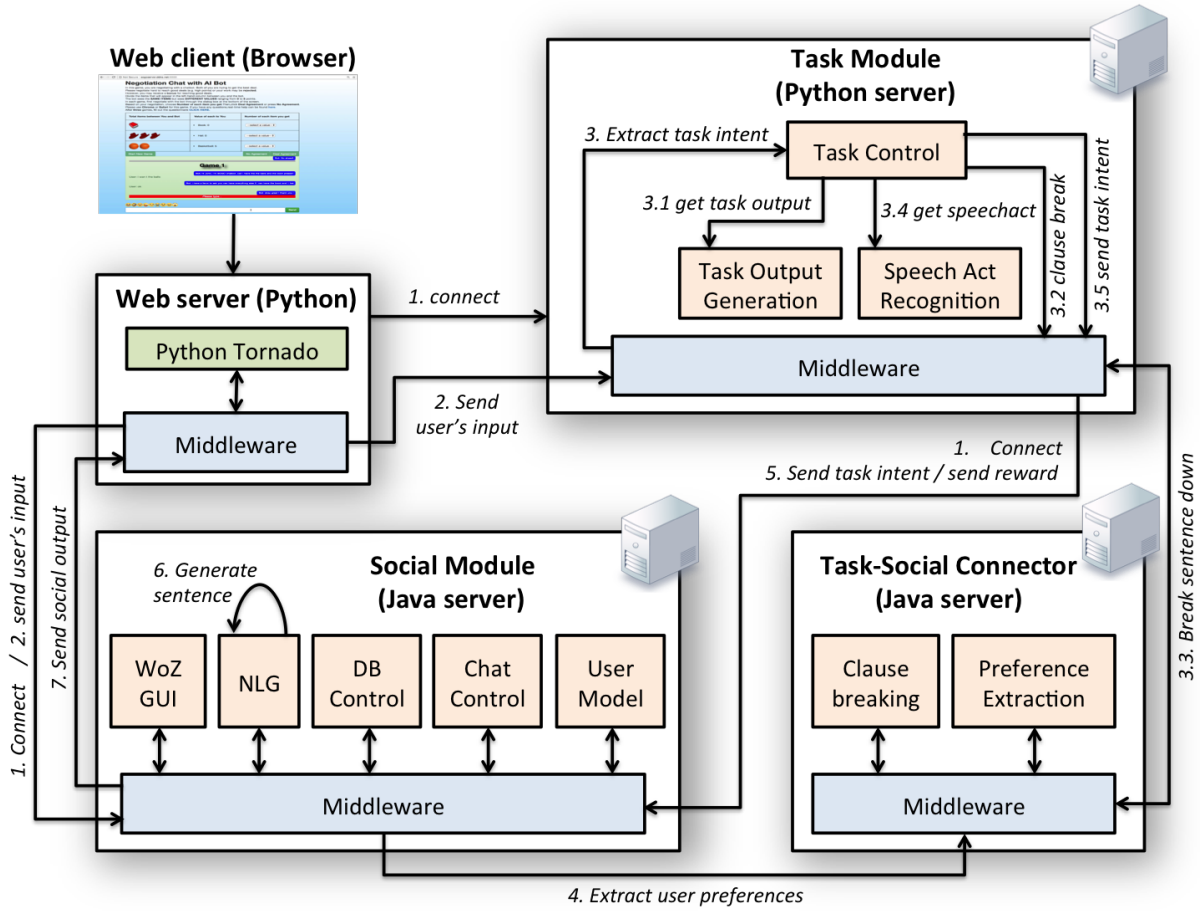


Figure 6.2: SOGO's Overall Architecture

shown in Figure 6.3.

SOGO modules are deployed across four server nodes (two Java servers and two Python servers), and the system can be accessed by users through a web browser (Chrome and Safari are supported). The Web Server processes web client requests and delegates who should process each request; the Task Module is in charge of generating the next task intent according to the negotiation state and last user's input; the Task-Social Connector module runs different NLP Stanford APIs and bootstraps the transition from task to social phase; and finally the Social Module generates a social output according to the current conversation state, the given task output, and the chosen conversational strategy. All servers use a middleware layer that guarantees interoperability across multiple languages and operating systems, hides the underlying complexity of the environment, and masks the heterogeneity of networking technologies to facilitate programming of high-level features. The middleware layer provides multiple capabilities such as communication, message passing, concurrency, logging, service discovery, session management, and component pluggability. The Middleware layer uses ZMQ, a high-performance asynchronous socked-based messaging library for use in distributed and concurrent applications with minimal latency footprint; it provides pre-built communication patterns and their implementations for more than 40 different

programming languages.

6.4.1 The system pipeline

the Task-Social Connector server begins as a daemon service (a long running process) which listens for incoming requests. Then, the Social Module is launched and waits for other modules to connect. After a user initializes the interaction through the web browser and starts the negotiation game, the servers connect (the Web Server, the Task Module, the Social Module and the Social-Task Connector – step 1 in Figure 6.3). During the user's turn, his/her input (e.g., user says: "I want two books and the ball") is sent to both the Task and Social Modules (step 2), and the two processes run in parallel (steps 3 and 4). The Task Module generates a task output by using Facebook end-to-end (Lewis et al., 2017) (e.g., Agent: "I need one book and one hat, you can take the rest", step 3.1), which splits the input into two clauses (e.g., "I want two books" and "I want the ball", steps 3.2 and 3.3, respectively) and extracts the corresponding speech acts (e.g., <request, 2, books> and <request, 1, ball>, step 3.4). The composite task intent (i.e., the task output plus speech acts) is then sent to the Social Module (steps 3.5, and later, step 5). Meanwhile, the Social Module is extracting user preferences in order to update the user model (step 4) and displays user and system interaction in real-time on a WoZ (Wizard-Of-Oz) dashboard GUI. Once it receives the task intent from Task Module (step 5), the Social Module executes a template-based Natural Language Generator (NLG) component which loads a set of pre-defined conversational strategies (using the DB Control). These strategies are then combined with a self-reflection mechanism based on a user's input parser and filtered using a rule-based system (step 6). Given the current dialog state, a set of plausible social sentences are shown to the Wozer, who chooses one (e.g., Agent: "This book looks exactly like one my grandpa gave me, would you mind giving me that book and the hat that looks really nice on me? you can have the rest. . . "). The Social Module sends this output to the Web Server, which in turn displays it on user's browser (step 7).

6.4.2 Logging System

Our middleware layer logs many types of events: changes to the user model (e.g., preferences, likes, dislikes), changes to the current game (e.g., speech act sequence, intentions, conversational strategy sequences), changes across multiple games (e.g., deal rates, success rates, scores), changes to the conversational state (e.g., number of turns, user's output, agent's output, deal items), and system errors. We developed a variety of log parsers that helped us to extract json messages that were embedded on those logs for further evaluation of the data collected from the experiments.

6.5 Pilot study with SOGO system

In the Wizard-of-Oz setup, our SOGO is semi-automatic: The functions described above are carried out automatically, but the Wozer decide which strategy to use when there are multiple available. This pilot study will serve as the proof of the concept for our proposed work. In our experiment, we use Facebook end-to-end dialog model as a baseline and compare it with our developed SOGO system. Thus, we recruited 60 English speakers on Amazon Mechanical Turk

who were equally and randomly assigned to one of the conditions. To obtain high quality data, those workers were be based in the United States or UK and had at least 95% approval rating and 5,000 previous HITs. Each participant played six games with the agent and completed a subjective questionnaire to reveal their feelings toward the game and interlocutor. As we explained above, our experiment sought to evaluate the effectiveness of SOGO on rapport building with a human user and its performance in the context of negotiation.

6.5.1 Evaluation

In this study, we combine subjective and objective measures to assess the effectiveness of our social intelligent negotiation dialog system on both rapport-building and negotiation performance. First, we conducted a two-tailed independent sample t-test on the questionnaires to explore the difference of mean value of users' rating on two systems in each question. For all significant results (p¡0.05), we also calculated effect size via Cohen's d to test for generalizability of results.

Subjective metrics

Based on items used in prior studies (Curhan et al., 2006; Gratch et al., 2015; DeVault et al., 2015), we developed a 15-item self-report questionnaire that characterizes the interaction into dimensions of rapport, such as coordination, attentiveness, positivity, and so on; question 14 asked users to directly rate the overall feeling of rapport during the interaction. Table 6.7 shows our list of questions. Responses were rated by each participant on a scale of 0 (Strongly Disagree) to 7 (Strongly Agree). Factor analysis proved only one factor for the 15 questionnaire items, which have high internal consistency with Cronbach's α= 0.94. Table 6.7 show the complete results. We describe our findings of differences between two grounds on each dimension of rapport as follows:

Coordination: We observed that users felt more in sync with the SOGO system(d=0.81), as they could say almost everything that they wanted to say during the interaction (d=-0.79). Effective sizes in these categories was high. Next, we noticed users felt a little bit frustrated in both setups but showed no significant differences toward the two systems.

Attentiveness: Users reported that the SOGO system paid more attention to them(d=0.89). Meanwhile, they also realized that the SOGO system was more respectful and considered their concerns (d=0.80). Both findings also have a high effective size. Users stated that they were interested in listening to the system in both conditions. Thus, there is no significant difference among the two groups of participants, but the mean values of this question are all low (Mean(SOGO)=1.63,Mean(FB)=1.70).

Positivity: Three questions in this dimension have significantly different responses between the two groups. Users liked the SOGO system more and felt warm toward to their partner (d=0.76). They experienced more of a sense of friendliness (d=1.00) and caring from the SOGO system, as well (d=0.87).

Face: Both groups reported a low degree of damaging their sense of pride but no significant difference was found across groups.

Feeling about the negotiation: We ameliorated the uncompromising and uncooperative impression in users from the FB system to the SOGO system (d=-0.35), though the rating of the

Objective Metrics	t-test value	Mean value	Effective Size
Win Times	t(29)=2.59*	SOGO=2.70 FB=1.80	d=0.67
Deal Rate	t(29)=7.74***	SOGO=0.90 FB=0.45	d=1.99
Average Dialog Length	t(29)=-1.50	SOGO=6.80 FB=7.59	d=-0.39
Average User Utterance Length	t(29)=2.59*	SOGO=7.17 FB=5.48	d=0.67
Pareto Optimal	t(29)=2.05**	SOGO=96.67 FB=80.00	d=0.53

Table 6.6: Complete t-test statistical analysis of negotiation performance of SOGO system versus Facebook Baseline system. Effect size assessed via Cohen's d. Significance: ***:$p<0.001$, **:$p<0.01$, *:$p<0.05$

SOGO system is still unsatisfied with the Mean(SOGO)=2.23. Users felt more satisfied about the instrumental outcome in the SOGO system with low effective size (d=0.31). Finally, we found that users regarded the whole negotiation process as a good foundation for a future relationship with the SOGO system. The attitude to these questions differ significantly in the two groups with a large effective size (d=0.99).

Perceived Rapport: User perceive significantly higher rapport with SOGO system (Mean(SOGO)=5.10) comparing to FB system (Mean(SOGO)=3.50).

Information Disclosure: Users prefer to share more of their personal information with the Facebook system rather than the SOGO system but the results suffer from low effective size (d=-0.2).

Objective metrics

In objective metrics, we first measured agent performance through three dimensions inherited from (Lewis et al., 2017): (1) Number win by the system (**Number of Win**). Obviously, the SOGO system wins more often than the Facebook system, with a moderate effective size (d=0.67). (2) Percentage of games that end up with an agreed-upon negotiation decision (**Deal Rate**). We know this is the major problem with the Facebook system in (Lewis et al., 2017) since users even prefer not to agree rather than capitulate to an uncooperative system. The SOGO system significantly improves this agreement rate from 0.45 to 0.90. The effective size is 1.99. (3) Percentage of Pareto optimal solutions for agreed deals (**Pareto Optimal**). The SOGO system performs well in this dimension likely because users prefer to adapt or even sacrifice themselves to agree with the system as a means of building rapport.

Next, since conversation length and response length are strong objective indicators of user engagement or interest (Yu et al., 2017), we include them here as well. We find that the average dialog length for both systems are similar, however, users reply with more words in each utterance when they negotiate with the SOGO system (d=0.67). Table 6.6 provides complete results.

Dimension	Subjective Questions	t-test value	Mean value	Effective Size
Coordination	1. I think that my agent and I were in sync with each other	t(29)=3.13**	SOGO=5.10,FB=3.77	d=0.81
	2. I felt uncomfortable and could not say everything that I wanted to say	t(29)=-3.05**	SOGO=1.63,FB=2.73	d=-0.79
	3. The interaction was frustrating	t(29)=-1.82	SOGO=2.90,FB=3.70	d=-0.47
Attentiveness	4. I felt that my agent was paying attention to what I was saying	t(29)=3.44**	SOGO=5.23,FB=3.53	d=0.89
	5. I was not really interested in what my agent was saying	t(29)=-0.24	SOGO=1.63,FB=1.70	d=-0.06
	6. My agent was respectful to me and considered to my concerns	t(29)=3.11**	SOGO=5.43,FB=4.10	d=0.80
Positivity	7. My agent was friendly to me	t(29)=3.90***	SOGO=5.97,FB=4.43	d=1.00
	8. I liked and felt warm toward my partner	t(29)=2.96**	SOGO=5.10,FB=3.80	d=0.76
	9. My agent cared about me	t(29)=3.36**	SOGO=4.73,FB=3.30	d=0.87
Face	10. Did you lose face (i.e., damage your sense of pride) in the negotiation?	t(29)=0.00	SOGO=1.76,FB=1.76	d=0.0
Feeling about the negotiation	11. My agent was very uncooperative.	t(29)=-1.36**	SOGO=2.23,FB=2.70	d=-0.35
	12. How satisfied are you with the balance between your own outcome and your agent's outcome(s)?	t(29)=1.21**	SOGO=5.07,FB=4.57	d=0.31
	13. Did the negotiation build a good foundation for a future relationship with your agent?	t(29)=3.83***	SOGO=5.37,FB=3.83	d=0.99
Perceived Rapport	14. I felt rapport between the agent and myself	t(29)=4.04**	SOGO=5.10,FB=3.50	d=1.04
Information Disclosure	15. I was willing to share information with my agent.	t(29)=-0.77**	SOGO=4.40,FB=4.73	d=-0.20

Table 6.7: Complete t-test statistical analysis of subjective questionnaire of rapport assessment by comparing SOGO system and Facebook end-to-end system. Effect size assessed via Cohen'sd. Significance: ***:p <0.001, **:p <0.01, *:p <0.05

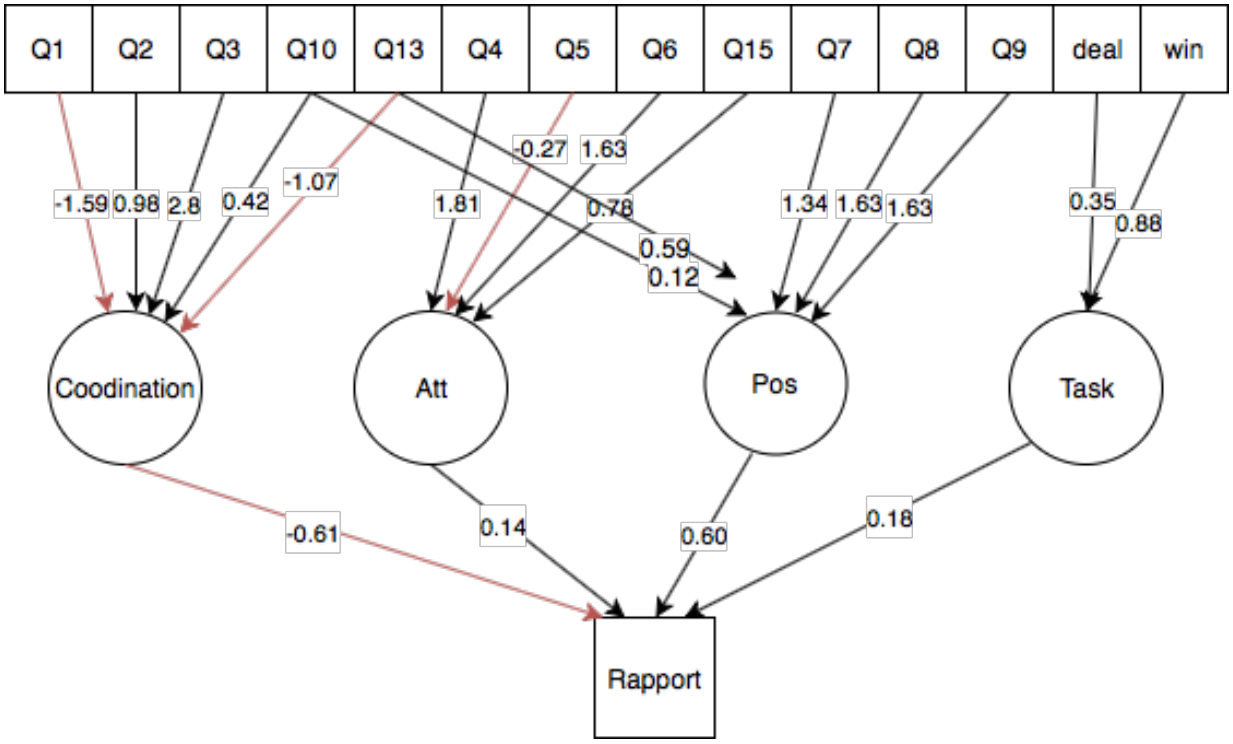


Figure 6.3: Structural Equation Model of Rapport

Factor analysis of Rapport

Since (Spencer-Oatey, 2008b; Zhao et al., 2014b) have provided a clear computational model of rapport, we proposed a four-factor model to explain the dynamics of rapport. Our goals were to validate this model and specify the variable loadings of each question to its corresponding factor of rapport. With respect to these goals, we conducted a confirmatory factor analysis, which assessed how well the proposed model captured the covariance between all variables in the model. In our case, subjective questions were observed variables, represented by square boxes; the subcomponents of rapport were the latent factors, drawn by circles. Overall, our proposed four-factor model is a relatively good fit based on the metrics listed in the table 6.5.1. In the top-level of the model, coordination and positivity correlate strongly to rapport in the context of negotiation. This result tells us that users value feelings of synchronicity and friendliness more than others. The negative sign of coordination loadings is due to the reversed-coded questions in the bottom level. Also, our findings confirm (Curhan et al., 2006), namely that the outcome of the task (e.g., win or lose) does not greatly affect the building of a social bond. Unexpectedly, attentiveness seems like the least important factor of rapport, which is contradicts rapport theory (Tickle-Degnen and Rosenthal, 1990). One explanation of this phenomenon is that attentiveness has a large covariance with both coordination and positivity. Thus, its explantory power toward the variance of rapport is reduced. In the bottom level, most questions have high variable loadings and figure 6.3 displays the complete results.

Name	Value
RMSEA	0.094
SRMR	0.063
CFI	0.937
TLI	0.915

Table 6.8: Model fit metrics. RMSEA = root mean square error of approximation; SRMR = Standardized Root Mean Square Residual; CFI = comparative fit index; TLI = tucker-lewis index

6.5.2 Discussion

In this section, we demonstrate the proof of the concept social intelligent negotiation dialog system, which can negotiate with people while building a social bond. These achievements come from our proposed two-phase computational model that blends social moves with task moves in an utterance. We leverage the off-the-shelf end-to-end dialog model to decide the next task move and a theory-driven template-based social language generator to introduce social skill to the system. The experiment demonstrates that our SOGO system behaves in sync with its user. People feel more comfortable and engaged during the interaction compared to the baseline Facebook system. Even though they stay in a semi-cooperative environment, people thought our SOGO system was friendly and cared about them.

The design of our computational model and system architecture is supported by empirical work in social psychology, which helps identify areas in which the system can develop human-like qualities. Our work operationalizes these theories to practical human-agent interaction. Especially, in the social phase, we realized abstract theoretical findings into conversational strategies and speech act strategies. Our findings improve understanding of how to instantiate rapport in human-agent negotiation. Broadly speaking, we validated the discovery that strategies for building buddy relationships in human-human communication could be transferred to human-agent interaction.

6.6 Proposed Work

6.6.1 Automatically assessing negotiation rapport

In our pilot study, objective metrics are evaluated automatically. Subjective metrics, however, are acquired from a human self-reported questionnaire, which is expensive. Such human evaluations restrict us to make our system to learn social skills in negotiation from self-play mode. Thus, before moving to build a fully-automatic SOGO system, we expect to develop a model that learns to predict human-like rapport scores to a dialog. Since the dynamics of rapport is affected by both propositional goals and interpersonal goals in conversation, our model should be able to understand both conversational semantics and infer a human user's social intention, which manifests by different conversational strategies based on our socially-aware framework in section 2. This evaluation model will serve as the ground truth of negotiation rapport for training our fully-automatic SOGO system.

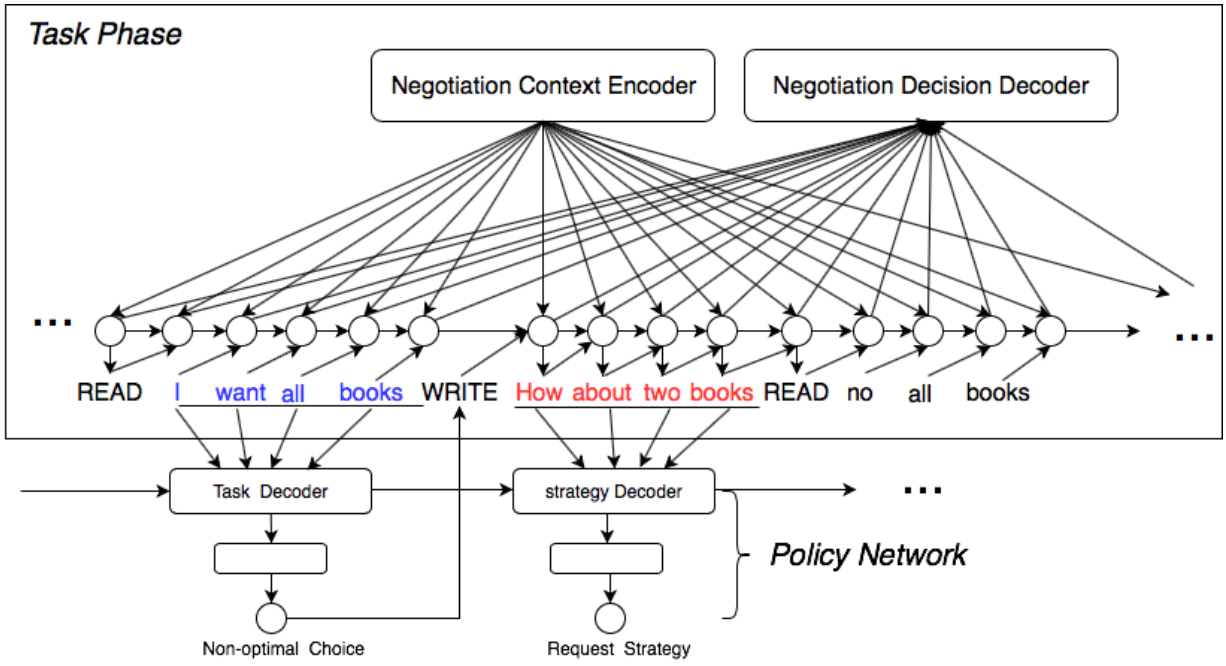


Figure 6.4: Policy Network for learning conversational strategy

6.6.2 SOGO 2.0: Fully-automatic SOGO System

Next, we seek to upgrade our semi-automatic SOGO system to be a fully-automatic system, SOGO 2.0. With the help of the rapport estimator developed in section 6.6.1, we can simulate agent-agent dialogs with the ground truth rapport ratings as the social outcome. In this way, we intend to design a policy network to optimize the social outcome. The model structure is shown in Figure 6.4. Specifically, we will freeze the parameters of the task phase model and tune the policy network for predicting the next system task choice and conversational strategy. Intuitively, task phase will generate the two best utterances in respect to the task outcome. A task decoder will predict the favor of selecting the next system utterance from these two options: optimal or non-optimal. A non-optimal choice allows the system to learn to yield to the human's repeated request, which makes our system more compromising. Subsequently, the strategy decoder will pick the conversational strategy for the social language generator.

6.6.3 SOGO 3.0: Fully-automatic SOGO system with off-task social conversation

Both semi-automatic SOGO and SOGO 2.0 system mainly introduce in-task social conversation, wherein the talk topic is closely related to negotiation (aside from general greetings and closings). Recent studies (Lucas et al., 2018) employ off-task social conversation (e.g., chit-chat) to facilitate rapport in the ranking task and demonstrate the effectiveness of the built rapport in recovery from errors. We intend to expand our conversational strategy set by including this chit-chat strategy and incorporate some off-task social conversations to our negotiation dialogs. However, most

chit-chat models suffer from a lack of specificity and an inconsistent personality. To alleviate this problem, (Lucas et al., 2018) presented a dialog model that generates utterances conditioned on a user's provided profile information. Drawing on this work, we plan to incorporate this chit-chat dialog model into our SOGO system. The chit-chat dialog model will engage the user at the beginning of the interaction by exchanging personal information and in an attempt to get to know its interlocutor. Subsequently, we will move to negotiation games and the generation task will be dominated by SOGO system.

Chapter 7

Timeline

7.1 Timeline

- Theoretical Framework of Rapport Management
  - Published at IVA 2014
- Socio-Cognitive Effects of Conversational Strategy Congruence
  - Published at ICMI 2015 (joint work with Tanmay Sinha)
- Predictive Model for Conversational Strategies Recognition
  - Published at SIGDIAL 2016
- Predictive Model for Rapport Assessment
  - Published at IVA 2016
- Discourse Planning for Social Dialog
  - Published at IJCAI 2017 (joint work with Oscar J Romero)
- SARA: A Socially-Aware Animated Intelligent Personal Assistant Agent
  - Published at SIGDIAL 2016 (joint work with Yoichi Matsuyama)
- SOGO: A Social Intelligent Negotiation Dialog System
  - Submitted to: IVA 2018
- SOGO 2.0: Fully-automatic SOGO System
  - Target Conference: SLT 2018
- SOGO 3.0: Fully-automatic SOGO system with off-task social conversation
  - Sep-Oct 2018
- Thesis writing and defense: Nov-Dec 2018

Appendices

Appendix A

Sample Dialogs of Sociall-Aware Intelligent Personal Assistant

Table A.1: Extract from an actual interaction

Turn	Speaker	Utterance	Intent	Strategy
1	System	I'm here to be your personal assistant and help you get the most out of the conference, but to be honest, I'm a little nervous. This is my first time here, is it yours as well?	introduce, do_attendance_elicitation	SD, QE
2	User	Actually, this is my first time here as well.	first_time_attending	SD
3	System	I'm finding it overwhelming. I bet you are too.	feedback_attendance_elicitation_first_time	SD, QE
4	User	Ya, there's just so much going on. But i'm getting a hang of it.	positive_confirmation	SD
5	System	Great. Now, If you wouldn't mind sharing your goals for attending the world economic forum this year, I can try to help you achieve them.	do_goal_elicitation	ASN
6	User	Well, I'm here to meet people who are passionate about the use of technology in education, and to know more about the latest developments in the field.	recommend_people, recommend_sessions	None
7	System	Great. Good goals. Can you tell me a little bit about your work so I understand your interests better?	feedback_goals, do_interest_elicitation	ASN
8	User	I am the founder of a non profit educational startup which helps deliver quality education to the remote areas.	like {startup, education}	None
9	System	Very Interesting. Let's start by finding some interesting people for you to meet that are relevant to your interests. This is my favorite part. Let me look this up. One moment.	feedback_interest_elicitation, do_person_recommendation	ASN, SD
10	User	Sure	positive_confirmation	None
11	System	I think you would enjoy meeting this attendee. On the screen are some more details. Well? What do you think?	outcome_person_recommendation, end_person_recommendation	ASN, VSN

Appendix B

Pre-conditions and Post-conditions of conversational strategies in social reasoner

Self-Disclosure
Pre-conditions: [low–rapport, medium–rapport, rapport–decreased], [sd–user, qesd–user], [smile, gaze–elsewhere], [introduce, start*], . . .
Post-conditions (add): [sd–history, smile, gaze–partner, rapport–increased, rapport–maintained], . . .
Post-conditions (delete): [rapport–decreased, sd–user, qesd–user, pr–history, vsn–history, introduce, start–*], . . .

Acknowledgement
Pre-conditions: [sd–user, vsn–user], [gaze–partner], [not–ack–history–user, not–ack–history–system], [feedback–*]
Post-conditions (add): [ack–history, rapport–maintained]
Post-conditions (delete): [not–ack–history, feedback–*]

Praise
Pre-conditions: [low–rapport], [not–pr–user], [not–pr–history–user, sd–history–system, turns–lower–thresh, not–pr–history–system, qesd–history–system], . . .
Post-conditions (add): [pr–system, pr–history, rapport–increased, rapport–maintained], . . .
Post-conditions (delete): [low–rapport, not–pr–history],

Question Elicitation Self-disclosure
Pre-conditions: [rapport–increased], [not–qesd–history, not–sd–history], [do–*, preclosing, ask–*] . . .
Post-conditions (add): [qesd–system, gaze–partner] . . .
Post-conditions (delete): [not–qesd–history–system, not–sd–history–system, do–*, preclosing, ask–*], . . .

Reference to Shared Experiences
Pre-conditions: [medium–rapport, high–rapport], [rse–user, sd–user, vsn–user], [vsn–history, not–rse–history–system], [available–shared–experiences] . . .
Post-conditions (add): [rse–history, rapport–increased, rapport–maintained, gaze–partner], . . .
Post-conditions (delete): [gaze–elsewhere], . . .

Adhere to Social Norm
Pre-conditions: [low–rapport, medium–rapport], [not–asn–history–system], [outcome–*–recommendation, preclosing, greeting, farewell, feedback–*, start–*, . . .]
Post-conditions (add): [asn–system, asn–history, rapport–maintained, gaze–partner, . . .]
Post-conditions (delete): [not–asn–history–system, [outcome–recommendation, farewell, feedback–*, . . .]

Violation of Social Norm
Pre-conditions: [high–rapport], [vsn–user], [smile, gaze–partner], [turns–higher–threshold], [once–vsn–history–user, not–vsn–history–system], [start–*, feedback–*,]. . .
Post-conditions (add): [vsn–history, rapport–increased,]
Post-conditions (delete): [not–vsn–history–system, greeting, start–*, feedback–*, do–*, . . .]

References

IRWIN Altman. 1973. Reciprocity of interpersonal exchange1. *Journal for the Theory of Social Behaviour* 3(2):249–261.

Irwin Altman and Dalmas Taylor. 1973. Social penetration theory. *New York: Holt, Rinehart &\ Mnston* .

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111(2):256.

Bernard J. Baars. 2003. *The global brainweb: An update on global workspace theory*. Science and Consciousness Review.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

John A Bargh, Katelyn YA McKenna, and Grainne M Fitzsimons. 2002. Can you see the real me? activation and expression of the "true self" on the internet. *Journal of social issues* 58(1):33–48.

Roy F Baumeister and Mark R Leary. 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin* 117(3):497.

Leslie M Beebe and Tomoko Takahashi. 1989. Sociolinguistic variation in face-threatening speech acts. In *The dynamic interlanguage*, Springer, pages 199–218.

Frank J Bernieri and John S Gillis. 2001. Judging rapport: Employing brunswik's lens model to study interpersonal sensitivity. *Interpersonal sensitivity: Theory and measurement* pages 67–88.

Frank J Bernieri and Robert Rosenthal. 1991. Interpersonal coordination: Behavior matching and interactional synchrony. *Fundamentals of nonverbal behavior* 401.

Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2008. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agents*. Springer, pages 262–269.

Timothy Bickmore and Justine Cassell. 1999. *Small Talk and Conversational Storytelling in Embodied Conversational Interface Agents*.

Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pages 396–403.

Timothy Bickmore and Justine Cassell. 2005. Social dialogue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, Springer, pages 23–54.

Timothy Bickmore, Laura Pfeifer, and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. In *Intelligent Virtual Agents*. Springer, pages 55–67.

Timothy Bickmore and Daniel Schulman. 2012. Empirical validation of an accommodation theory-based model of user-agent relationship. In *International Conference on Intelligent Virtual Agents*. Springer, pages 390–403.

Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. its just like you talk to a friendrelational agents for older adults. *Interacting with Computers* 17(6):711–735.

Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (ccsarp) pages 196–213.

Ilan Bronstein, Noa Nelson, Zohar Livnat, and Rachel Ben-Ari. 2012a. Rapport in negotiation: The contribution of the verbal channel. *Journal of Conflict Resolution* 56(6):1089–1115.

Ilan Bronstein, Noa Nelson, Zohar Livnat, and Rachel Ben-Ari. 2012b. Rapport in negotiation: The contribution of the verbal channel. *Journal of Conflict Resolution* 56(6):1089–1115.

Penelope Brown and Stephen Levinson. 1978. Universals in language usage: Politeness phenomena. questions and politeness: Strategies in social interaction, ed. by e. goody, 56-311.

Jerry M Burger, Jackeline Sanchez, Jenny E Imberi, and Lucia R Grande. 2009. The norm of reciprocity as an internalized social norm: Returning favors even when no one finds out. *Social Influence* 4(1):11–17.

Michael Burns. 1984. Rapport and relationships: The basis of child care. *Journal of Child Care* .

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction* 13(1-2):89–132.

Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, Hannes Vilhjalmsson, and Hao Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, pages 520–527.

Justine Cassell, Alastair J Gill, and Paul A Tepper. 2007. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*. Association for Computational Linguistics, pages 41–50.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, Springer, pages 163–185.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

William Coon, Charles Rich, and Candace L Sidner. 2013. Activity planning for long-term relationships. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013, Proceedings*. Springer, volume 8108, page 425.

Jared R Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology* 91(3):493.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* .

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 307–318.

Valerian J Derlega, Sandra Metts, Sandra Petronio, and Stephen T Margulis. 1993. *Self-disclosure*. Sage Publications, Inc.

David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction. AAAI Press, Stanford, CA*.

Eli Dresner and Susan C Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication theory* 20(3):249–268.

Aimee L Drolet and Michael W Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology* 36(1):26–50.

Miriam Eisenstein and Jean W Bodman. 1986. i very appreciate: expressions of gratitude by native and non-native speakers of american english. *Applied linguistics* 7(2):167–185.

Paul Ekman and Wallace V Friesen. 1986. A new pan-cultural facial expression of emotion. *Motivation and emotion* 10(2):159–168.

Paul D Ellis. 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Richard M. Emerson. 1976. Social exchange theory. *Annual Review of Sociology* 2:pp. 335–362.

Nick J Enfield. 2013. Reference in conversation. *The handbook of conversation analysis* pages 433–454.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*. ACM, pages 1459–1462.

Peyman Faratin, Carles Sierra, and Nick R Jennings. 1998. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems* 24(3-4):159–182.

Chaim Fershtman. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior* 2(3):224 – 238.

Leon Festinger. 1954. A theory of social comparison processes. *Human relations* 7(2):117–140.

Donna T Fujimoto. 2009. *Listener responses in interaction: A case for abandoning the term, backchannel*. ł.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software* 31(7):1–24.

Erving Goffman. 2005. *Interaction ritual: Essays in face to face behavior*. AldineTransaction.

Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *International Conference on Intelligent Virtual Agents*. Springer, pages 201–215.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. 2006. Virtual rapport. In *Intelligent virtual agents*. Springer, pages 14–27.

Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents*. Springer, pages 125–138.

Mathieu Guillame-Bert and James L. Crowley. 2012. Learning temporal association rules on symbolic time sequences. In *Proceedings of the 4th Asian Conference on Machine Learning, ACML 2012, Singapore, Singapore, November 4-6, 2012*. pages 159–174.

Mathieu Guillame-Bert and Artur Dubrawski. 2014. Learning temporal rules to forecast events in multivariate time sequences. In *NIPS, Modern Machine Learning and Natural Language Processing Workshop*.

Robert H Guttman and Pattie Maes. 1998. Agent-mediated integrative negotiation for retail electronic commerce. In *International Workshop on Agent-Mediated Electronic Trading*. Springer, pages 70–90.

A Hartholt, D Traum, SC Marsella, A Shapiro, G Stratou, A Leuski, LP Morency, and J Gratch. 2013. All together now: Introducing the virtual human toolkit. In *Int. Conf. on Intelligent*

Virtual Humans.

Charles T. Hill and Donald E. Stull. 1982. Disclosure reciprocity: Conceptual and measurement issues. *Social Psychology Quarterly* 45(4):pp. 238–244.

Koen Hindriks, Catholijn M Jonker, Sarit Kraus, Raz Lin, and Dmytro Tykhonov. 2009. Genius: negotiation environment for heterogeneous agents. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, pages 1397–1398.

George C. Homans. 1958. Social behavior as exchange. *American Journal of Sociology* 63(6):pp. 597–606.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual rapport 2.0. In *Intelligent Virtual Agents*. Springer, pages 68–79.

Adam N Ioinson and Carina B Paine. 2007. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology* page 2374252.

E Kacewicz, J. W Pennebaker, M Davis, M Jeon, and A. C Graesser. 2009. The language of social hierarchies .

Sin-Hwa Kang, Jonathan Gratch, Candy Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. 2012. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th ICAAMS-Volume 1*. pages 63–70.

Bilge Karacora, Morteza Dehghani, Nicole Krämer-Mertens, and Jonathan Gratch. 2012. The influence of virtual agents gender and rapport on enhancing math performance. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. pages 563–568.

Dacher Keltner and Brenda N Buswell. 1997. Embarrassment: its distinct form and appeasement functions. *Psychological bulletin* 122(3):250.

Dejun Tony Kong, Kurt T. Dirks, and Donald Ferrin. 2014. Interpersonal trust within negotiations: Meta-analytic evidence, critical contingencies, and directions for future research 57:1235–1255.

Vasily Konovalov, Ron Artstein, Oren Melamud, and Ido Dagan. 2016. The negochat corpus of human-agent negotiation dialogues. In *LREC*.

Robert E Kraut and Robert E Johnston. 1979. Social and emotional messages of smiling: An ethological approach. *Journal of personality and social psychology* 37(9):1539.

Karel Kreijns, Paul A Kirschner, and Wim Jochems. 2003. 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.

Joseph B Kruskal and Mark Liberman. 1983. The symmetric time-warping problem: from continuous to discrete. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* pages 125–161.

Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R Pietromonaco. 1998. 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.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *ArXiv e-prints* .

Gale M Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pages 344–351.

Pattie Maes. 1989. How to do the right thing. *Connection Science* 1(3):291–323.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar J. Romero, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *17th Annual SIGdial Meeting on Discourse and Dialogue*.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat. In *Annual Conference on Artificial Intelligence*. Springer, pages 119–130.

Johnathan Mell and Jonathan Gratch. 2017. Grumpy & Pinocchio: Answering Human-Agent Negotiation Questions through Realistic Agent Design. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Sao Paulo, Brazil, pages 401–409.

Youngme Moon. 2000. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research* 26(4):323–339.

Janice Nadler. 2003. Rapport in nd conflict resolution. *Marq. L. Rev.* 87:875.

Yukiko I Nakano, Sakiko Nihonyanagi, Yutaka Takase, Yuki Hayashi, and Shogo Okada. 2015. Predicting participation styles using co-occurrence patterns of nonverbal behaviors in collaborative learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 91–98.

Radoslaw Niewiadomski, Ken Prepin, Elisabetta Bevacqua, Magalie Ochs, and Catherine Pelachaud. 2010. Towards a smiling eca: studies on mimicry, timing and types of smiles. In *Proceedings of the 2nd international workshop on Social signal processing*. pages 65–70.

Neal R Norrick. 2003. Issues in conversational joking. *Journal of Pragmatics* 35(9):1333–1359.

Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012a. Rudeness and rapport: Insults and learning gains in peer tutoring. In *Intelligent Tutoring Systems*. Springer, pages 11–21.

Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012b. Rudeness and rapport: Insults and learning gains in peer tutoring. In *International Conference on Intelligent Tutoring Systems*. Springer, pages 11–21.

James W Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development

and psychometric properties of liwc2015.

Sandra Petronio. 2012. *Boundaries of privacy: Dialectics of disclosure*. Suny Press.

Isabella Poggi, Francesca D'Errico, and Laura Vincze. 2010. Types of nods. the polysemy of a social signal. In *LREC*.

Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice hall.

Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 23–30.

Landra Rezabek and John Cochenour. 1998. Visual cues in computer-mediated communication: Supplementing text with emoticons. *Journal of Visual Literacy* 18(2):201–215.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*. pages 704–714.

Carl R Rogers. 1966. *Client-centered therapy*. American Psychological Association.

Oscar J. Romero. 2011. An evolutionary behavioral model for decision making. *Adaptive Behavior* 19(6):451–475.

Oscar J. Romero and Angelica de Antonio. 2012. Evolving the way of doing the right thing. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC*. Springer, pages 1–8.

Oscar J. Romero, Ran Zhao, and Justine Cassell. 2017a. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. pages 3807–3813.

Oscar J. Romero, Ran Zhao, and Justine Cassell. 2017b. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proc. of the 26th Int. Joint Conf. on Artificial Intelligence, IJCAI-17*. pages 3807–3813.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 26(1):43–49.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of uh huhand other things that come between sentences. *Analyzing discourse: Text and talk* 71:93.

Anna M Sharpley, James W Irvine, and Christopher F Sharpley. 1983. 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.

C Sidner. 2012. Engagement: Looking and not looking as evidence for disengagement. *Workshop at HRI* 12.

Tanmay Sinha and Justine Cassell. 2015a. Fine-grained analyses of interpersonal processes and their effect on learning. In *Artificial Intelligence in Education*. Springer, pages 781–785.

Tanmay Sinha and Justine Cassell. 2015b. 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.

Tanmay Sinha, Ran Zhao, and Justine Cassell. 2015. 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*.

Roslyn M Sparrevohn and Ronald M Rapee. 2009. Self-disclosure, emotional expression and intimacy within romantic relationships of people with social phobia. *Behaviour Research and Therapy* 47(12):1074–1078.

Helen Spencer-Oatey. 2005. (im) politeness, face and perceptions of rapport: unpackaging their bases and interrelationships.

Helen Spencer-Oatey. 2008a. *Culturally speaking: Culture, communication and politeness theory*. Continuum International Publishing Group.

Helen Spencer-Oatey. 2008b. *Face,(im) politeness and rapport*. Continuum International Publishing Group.

Ilya Sutskever. 2013. *Training recurrent neural networks*. Ph.D. thesis, University of Toronto.

Henri Tajfel and John C Turner. 1979. An integrative theory of intergroup conflict. *The social psychology of intergroup relations* 33:47.

Dalmas A Taylor and Irwin Altman. 1987. Communication in interpersonal relationships: Social penetration processes. .

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1(4):285–293.

Stanislav Treger, Susan Sprecher, and Ralph Erber. 2013. Laughing and liking: Exploring the interpersonal effects of humor use in initial social interactions. *European Journal of Social Psychology* 43(6):532–543.

Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L Sidner, and Timothy Bickmore. 2012. Designing relational agents as long term social companions for older adults. In *Intelligent Virtual Agents*. Springer, pages 289–302.

Joseph B Walther and Kyle P DAddario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social science computer review* 19(3):324–347.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. Love ya, jerkface: using sparse log-linear models to build positive (and impolite) relationships with teens. In *13th annual SIGdial meeting on discourse and dialogue*. Association for Computational Linguistics, pages 20–29.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, pages 74–85.

Zhou Yu, Alan W Black, and Alexander I Rudnicky. 2017. Learning conversational systems that

interleave task and non-task content. *arXiv preprint arXiv:1703.00099* .

Zhou Yu, David Gerritsen, Amy Ogan, Alan Black, and Justine Cassell. 2013a. Automatic prediction of friendship via multi-model dyadic features. In *Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics, Metz, France, pages 51–60.

Zhou Yu, David Gerritsen, Amy Ogan, Alan W Black, and Justine Cassell. 2013b. Automatic prediction of friendship via multi-model dyadic features. In *14th Annual SIGdial Meeting on Discourse and Dialogue, Metz, France*.

Mark P Zanna. 1999. *Advances in experimental social psychology*, volume 31. Elsevier.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* .

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014a. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*. Springer, pages 514–527.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014b. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, pages 514–527.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *17th Annual SIGDIAL Meeting on Discourse and Dialogue*.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016b. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *16th International Conference on Intelligent Virtual Agents*.

Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016c. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*. Springer, pages 218–233.

Tiancheng Zhao, Ran Zhao, Zhao Meng, and Justine Cassell. 2016d. Leveraging recurrent neural networks for multimodal recognition of social norm violation in dialog. *arXiv:1610.03112* .

Barry J Zimmerman. 2000. Self-efficacy: An essential motive to learn. *Contemporary educational psychology* 25(1):82–91.