Socially-Aware Dialogue System

Ran Zhao

CMU-LTI-19-008

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

<u>Thesis Committee:</u>
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 Philosophy
In Language and Information Technologies

Keywords: socially-aware framework, multimodal machine learning, temporal sequence learning, cognitive architecture, neural dialogue model, rapport, conversational strategy, spoken dialogue system, discourse analysis, social reasoning, natural language generation, negotiation

Abstract

In the past two decades, spoken dialogue 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 dialogue systems that build interpersonal closeness over the course of a conversation while also carrying out the task. This thesis attempts to address that shortcoming. Specifically, research in computational linguistics (Bickmore and Cassell, 1999) has shown that people pursue multiple conversational goals in dialogue, which include those that fulfill propositional functions to contribute information to the dialogue; those that fulfill interactional functions to manage conversational turn-taking; and those that fulfill interpersonal functions to manage the relationships between interlocutors. Although spoken dialogue 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 insufficient models of interpersonal goals and strategies in human communication.

As dialogue systems become more common and are used more frequently, 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 dialogues that fulfill interpersonal functions.

In Zhao et al., (2014), 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 intentions 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 dialogue system focuses on blended conversations that mix a goal-oriented task with social chat. As a proof of concept, we exploit a knowledge-inspired socially-aware personal assistant to aid conference attendees by eliciting their preferences through building rapport, and then making informed personalized recommendations about sessions to attend and people to meet.

Finally, we leverage the power of neural networks to model negotiation dialogue within our socially-aware intelligent framework. We present a novel learning method and a two-phase computational model to blend negotiation utterances and social conversation. Our method requires less human efforts than traditional knowledge-inspired approaches. We conduct comprehensive experiments to show that the system can facilitate negotiation while building a social bond with a human user.

Acknowledgments

I would like to sincerely express my gratitude, first and foremost, to my wonderfully supportive and helpful committee, Prof. Alexander I. Rudnicky, Prof. Alan Black, Prof. Louis-Philippe Morency and Dr. Amanda Stent. My PhD advisor, Prof. Rudnicky, allowed me freedom to explore the field of human-computer interaction, guided my future research directions, and appreciated and supported my ideas. Our relationship is deeper than the typical advisor-advisee relationship. Prof. Rudnicky generously shared with me his insights on the fundamental philosophy and theory of human communication, which sparked my curiosity about dialogue systems and inspired me to learn state-of-the-art models. Our meetings left me imaging new research directions that I had previously not thought possible, and to ensure rigor, he guided me to craft evaluation metrics for each study. Without his help, I could not have produced the quality work that follows. Prof. Black offered helpful and creative advice when I strayed off path. He taught me how to ground computational theories of speech and language with methods of practical implementation. He made me believe that one day a computer could speak like a human. I am privileged to have had the opportunity to learn about multimodal machine learning from Prof. Morency. His excellent teaching and thoughtful questions equipped me with both high-level methodology design and low-level model development on the research of human communication dynamics. Dr. Stent at Bloomberg provided me an invaluable internship experience which engaged my great passion for applying deep learning to finance. She directed me to inspiring resources that broadened my horizons, and she taught me how to shape research ideas and design experiments. Even more, she helped me present our work at an international conference. I am extremely grateful for having four world-renowned experts on my committee. Thank you all for supporting my thesis work.

During my PhD journey, I was fortunate to work on different projects and collaborate with several faculty and talented researchers, including Prof. Justine Cassell, Prof. Mark Dredze, Dr. Arun Verma, Dr. David Rosenberg, Tanmay Sinha, Oscar J. Romero, Tiancheng Zhao, Yuntian Deng, Yoichi Matasuyama, Alexandros Papangelis, Arjun Bhardwaj, Sushma Akoju, Zhao Meng and Orson Xu. Thank you all for offering me precious research experience and lifelong friendship. As a student at CMU, I have learned so much about machine learning and natural language processing, and honed skills to implement that knowledge. Here, I want to thank professors who taught me while at CMU: Tom Mitchell, Manuela Veloso, Ruslan Salakhutdinov, Graham Neubig, Eduard Hovy, Eric Nyberg, Reid Simmons, Chris Dyer, Noah Smith and Carolyn Penstein Rose.

In the past six years, I have attended many conferences, which gave me opportunities to discuss and interact with many smart people, such as Prof. Jonathan Gratch, Prof. David Traum, Dr. Jason Weston, Prof. Stefan Scherer, Prof. Stefan Kopp, Prof. Timothy Bickmore, Prof. Yukiko Nakano, Prof. Hannes Vilhjalmssoon, Natasha Jaques, Dr. Bing Liu, Elena Corina Grigore, Dr. Arno Hartholt, and many more than I can list here. I am inspired by their passion and grateful for their contributions to our field.

In addition, I want to thank my sponsors, Yahoo! and Amazon, for their generous funding; Dr. Doug Phillips for his professional English revision; and my undergraduate research assistants for their annotation work. I am grateful for your assistance, which sped up my research and made my life much easier.

Finally, to my dearest family: Thank you. To my parents, for your unconditional love and support. Thank you, Jingjing, my wonderful wife and soulmate, for facing with me the challenges of the last six years and giving me the courage to overcome the difficulties. Thank you, Winston and Harry, my two cute sons for making my life meaningful and enjoyable. This dissertation is as much yours as it is mine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Spoken dialogue systems have been widely deployed as interactive interfaces in devices such as smartphones and personal home assistants. In these systems, agents, which consist of modules that understand natural language, manage dialogue and generate natural language, analyze and interpret users' speech to facilitate and sustain human-computer interaction. Primarily, researchers have developed three types of spoken dialogue systems/bots: (1) *task completion bots* that help users complete discrete tasks such as booking movie tickets, making restaurant reservations, and so on; (2) *information retrieval bots* that support an interactive Q&A system over a knowledge base; and (3) *social chatbots* that aim to engage users in dialogue with human-like social skills. Recent revolutions in deep learning have significantly advanced the development of the first two types of bots, which can be used to collaborate and execute complex tasks with human partners in many domains. However, social chatbots fall short in replicating the interpersonal function of communication. Moreover, many people fear that AI represents an increased threat to privacy and autonomy and could lead to a loss of genuine human interaction. Therefore, researchers have begun to value the role that social competence plays in interaction (Neururer et al., 2018). Specifically, as naturalistic and repeated interactions with dialogue systems increasingly become a part of daily life, the system's capacity to convey information smoothly is not enough. Systems should also be able to build intimacy and rapport with a human user, as those qualities are crucial to sustaining a long-term relationship, which in turn is intrinsic to successful collaboration.

Our primary interest in social intelligence is interpersonal rapport, which has been identified as an important function of human interaction. According to social science literature, *rapport*, or the feeling of harmony and connection with another, significantly impacts human performance in a variety of domains such as counseling (Kang et al., 2012), negotiation (Drolet and Morris, 2000), and education (Bernieri and Rosenthal, 1991). Therefore, we are motivated to ask a key question in this thesis: How can we make a dialogue system build rapport with human in a task? To address this broad question, we break it down into two dimensions. The first dimension is the scenario which characterizes the relationship between participants in a dialogue. Two types of scenarios have been widely studied: cooperative and semi-cooperative. Cooperative scenarios include personal assistants, peer tutoring, etc. Semi-cooperative scenarios include negotiation, etc. The second dimension is the modality which describes the specific kind of information a system can access to during a dialogue. Generally, three sets of information are presented such as verbal, visual and vocal cues.

After introducing our key research question, I outline the evolution of our research. In Chapter 2, we study face-to-face peer tutoring interactions (Yu et al., 2013) to develop a dyadic computational model (**Zhao, Ran** et al., 2014)[1] that explains how interlocutors manage rapport via specific conversational strategies to fulfill the intermediate goals that lead to rapport: face management, mutual attentiveness, and coordination. Specifically, we use the foundational work by (Spencer-Oatey, 2008b) to conceptualize the interpersonal nature of *face* as a desire to be recognized for one's social value and individual positive traits. Face-boosting strategies such as *praise* enhance self-esteem and increase interpersonal cohesiveness or rapport in the dyad. (Spencer-Oatey, 2008b) also posits that over time, interlocutors intend to increase *coordination* by adhering to behavioral expectations, which are guided by sociocultural norms in the initial stages of interaction and by interpersonally-determined norms afterwards. In later stages of interaction, sociocultural norms may be purposely violated to accommodate the other's behavioral expectations. Meanwhile, in the increasing trajectory of interpersonal closeness, *referring to shared experience* allows interlocutors to increase coordination by indexing common history and differentiating between in-group and out-group individuals (Tajfel and Turner, 1979). To better learn about the other person, *mutual attentiveness* is crucial (Tickle-Degnen and Rosenthal, 1990). We have seen in our own corpora that mutual attentiveness is fulfilled by prompting one's interlocutors to provide information about themselves through the strategy of *eliciting self-disclosure*. As the relationship proceeds and social distance decreases, these self-disclosures become more intimate. Broadly, our developed socially-aware theoretical framework will help us select the appropriate conversational strategy for each social goal.

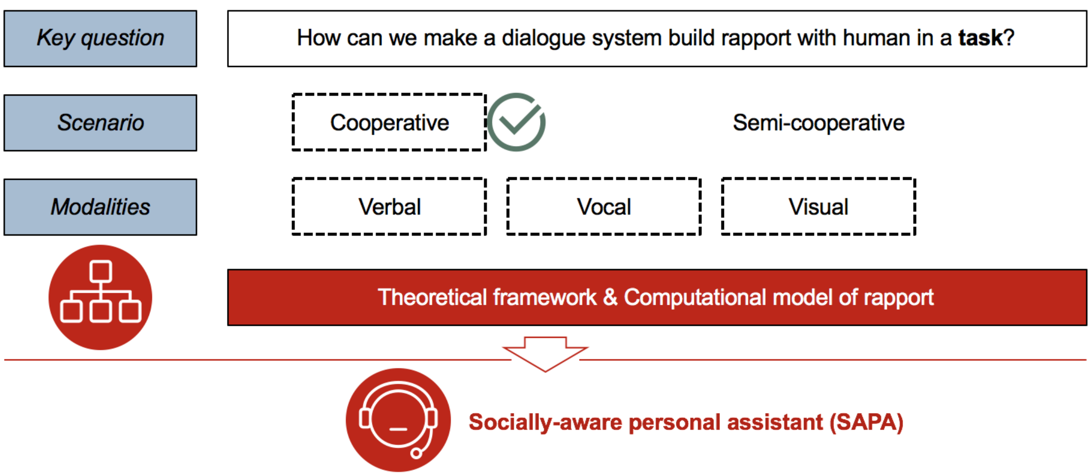

Figure 1.1: Setup of socially-aware personal assistant (SAPA)

Motivated by this theoretical rationale and our empirical findings concerning interpersonal closeness between human and agent, we next design socially-aware dialogue systems in both cooperative scenarios (Chapter 3) and semi-cooperative scenarios (Chapter 4). We begin with

[1]Zhao, Ran, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In Proceedings of the International Conference on Intelligent Virtual Agents (IVA'14).

cooperative scenarios (peer tutoring and personal assistant) because users who share a common goal are more likely to achieving rapport. Meanwhile, we allow the system to obtain all three modalities of information (visual, verbal, and vocal) in a dialogue. In Chapter 3.1, we review the architecture of our knowledge-inspired socially-aware dialogue system, a personal assistant that helps introduce conference attendees to other attendees and informs them about sessions that fit their interests (Matsuyama et al., 2016)[2]. To incorporate socially-aware intelligence to a traditional dialogue system, we argue that three extra capabilities should be built. First, systems must understand users' conversational strategies to infer their social intentions. Thus, in Chapter 3.2, we use quantitative methods to automatically recognize different conversational strategies that contribute to rapport (**Zhao, Ran** et al., 2016a)[3]. Once our system can understand a user's underlying intention, it should also be capable of estimating the dynamics of rapport. To achieve this second capability, in Chapter 3.3, we conduct a data-driven discovery of the temporally co-occurring and contingent behavioral patterns that signal high and low interpersonal rapport (**Zhao, Ran** et al., 2016b)[4]. We validate the discovered behavioral patterns by predicting rapport against our ground truth data (30-second thin slice rapport) via a forecasting model involving a two-step fusion of learned temporal associated rules. Finally, after assessing the social context of a dialogue, our system should be able to proactively manage the relationship with a human user by selecting the appropriate conversational strategy in response to the users input. In Chapter 3.4, we (Romero et al., 2017)[5] describe how our system 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 task goals (e.g., information-seeking, teaching, and calendar management) and social goals (e.g., managing rapport).

The promising results prompted us to explore more complicated settings. So, we moved to a semi-cooperative environment and modeled interpersonal functions of communication in negotiation. Even though face-to-face dialogue is natural, in most cases we are only able to access to the verbal channel since obtaining large high-quality multimodal data is impractical. Therefore, we examined verbal strategies within our theoretical framework on building rapport in negotiation dialogues. In Chapter 4, we instantiated a neural social -intelligent negotiation system called SOGO (**Zhao,Ran** et al., 2018)[6], equipped with a two-phase computational model that leverages the power of a neural network and reinforcement learning paradigm to blend social and task conversation (**Zhao, Ran** et al., 2019)[7]. This method aims to optimize both negotiation and

[2]Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar J. Romero, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In Proceedings of the 17th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL'16).

[3]Zhao, Ran, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialogue system. In Proceedings of the 17th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL'16).

[4]Zhao, Ran, Tanmay Sinha, Alan Black, and Justine Cassell. 2016b. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In Proceedings of the International Conference on Intelligent Virtual Agents (IVA'16).

[5]Oscar J. Romero, Ran Zhao, and Justine Cassell. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17).

[6]Zhao, Ran, Oscar J. Romero, and Alex Rudnicky. 2018. SOGO: A social intelligent negotiation dialogue system. In Proceedings of the International Conference on Intelligent Virtual Agents (IVA'18).

[7]Zhao, Ran, Oscar J. Romero, and Alex Rudnicky. 2019. Learning to blend social and task goals in conversational

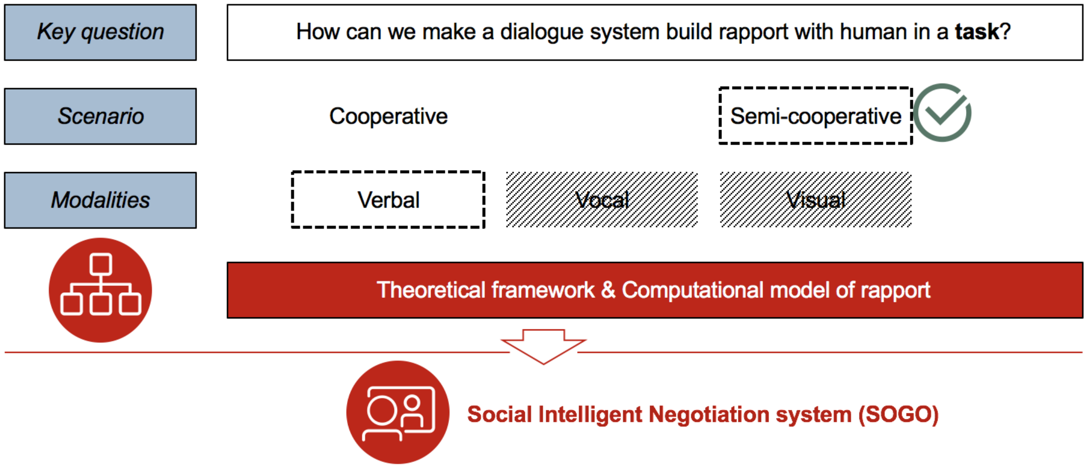

Figure 1.2: Setup of social intelligent negotiation system (SOGO)

rapport. Our empirical evaluation demonstrates promise in facilitating interpersonal rapport while improving negotiation performance.

1.1 Thesis Statement

In this thesis work, we design a socially-aware dialogue system that builds interpersonal closeness (rapport) by understanding human behaviors and generating appropriate responses over the course of a conversation along while completing a task.

1.2 Thesis Structure

- **Chapter 2 - Theoretical Framework**
  This chapter proposes a theoretical framework for rapport to explain how humans in dyadic interactions build, maintain and even destroy rapport via specific conversational strategies that function to fulfill special social goals and that are instantiated in particular verbal and nonverbal behaviors. In collaboration with Alexandros Papangelis and Justine Cassell, I contributed to designing theoretical computational model of rapport management based on the social science literature. This research work was presented in the following publication:
  - **Zhao, Ran**, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'14)* .

- **Chapter 3 - A Knowledge-inspired Socially-aware Personal Assistant (SAPA)**
  This chapter reviews our knowledge-inspired socially-aware dialogue system in a cooperative scenario, a personal assistant designed under the proposed theoretical framework.

agents through self-play (under review).

Chapter 3.1 - Overview of Architecture In this section, we lay out the building blocks for socially-aware dialogue systems, including conversational strategy classifier, rapport estimator and social reasonser. In collaboration with Yoichi Matsuyama, Arjun Bhardwaj, Oscar J. Romero, Sushma Anand Akoju and Justine Cassell, I contributed to integrating the Conversational Strategy Classifier, Rapport Estimator, BEAT and Smart Body modules to the whole architecture. Additionally, I contributed to the design of the social reasoner and the message passing infrastructure. This research work was presented in the following publications:

– Alexandros Papangelis, **Zhao, Ran**, and Justine Cassell. 2014. Towards a computational architecture of dyadic rapport management for virtual agents. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'14)* .

– Yoichi Matsuyama, Arjun Bhardwaj, **Zhao,Ran**, Oscar J. Romero, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL'16)* .

Chapter 3.2 - Conversational Strategy Classifier This section investigates patterns of conversational strategy usage in human dialogue that contribution to rapport-building. We incorporate these as the input into our understanding modules. Specifically, we introduce a machine learning model to automatically recognize conversational strategies through discovering the interplay of multimodal human behaviors. In collaboration with Tanmay Sinha, Alan Black and Justine Cassell, I contributed to writing the coding manual for conversational strategy annotation and conducting statistical analysis. I also contributed to proposing and extracting multimodal features, running parts of statistical tests, training and evaluating the machine learning model. This research work was presented in the following publications:

– Tanmay Sinha, **Zhao, Ran**, and Justine Cassell. 2015. Exploring socio-cognitive effects of conversational strategy congruence in peer tutoring. In *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI'15)* .

– **Zhao, Ran**, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL'16)* .

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

Chapter 3.3 - Rapport Estimator This section presents a temporal association rule-based model to forecast the dynamics of rapport. I contributed to proposing the idea, implementing the temporal association rule using the toolkit developed by Mathieu Guillame-Bert, extracting temporal association rules and validating temporal association rules by training predictive model for rapport estimation. This research work was presented in the following publication:

- **Zhao, Ran**, Tanmay Sinha, Alan Black, and Justine Cassell. 2016b. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'16)* . **(Best Paper Award)**

Chapter 3.4 - Conversational Strategy Planning for Social Dialogue This section focuses on how to automatically select appropriate conversational strategies for a dialogue system to realize the system's social intentions. We detail our behavior network formalism to model the decision-making process of the conversational strategy. In collaboration with Oscar J. Romero and Justine Cassell, I contributed to proposing the idea of using a spreading activation model, designing pre-conditions and post-conditions for each conversational strategy and conducting experiments. This research work was presented in the following publication:

- Oscar J. Romero, **Ran Zhao**, and Justine Cassell. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)* .

- **Chapter 4 - A Neural Social Intelligent Negotiation Dialogue System (SOGO)**
  This chapter demonstrates our socially-aware dialogue system in semi-cooperative scenario - negotiation dialogue. Within our socially-aware framework, we introduced a new paradigm of training a socially-aware neural dialogue model that leverages expert supervision to blend task and social goals in negotiation.
  - **Zhao,Ran**, Oscar J Romero, and Alex Rudnicky. 2018. SOGO: A social intelligent negotiation dialogue system. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'18)* . **(Best Paper Award)**
  - **Zhao, Ran**, Oscar J Romero, and Alex Rudnicky. 2019. Learning to blend social and task goals in conversational agents through self-play *(under review)* .

- **Chapter 5 - Conclusion and Future Work**
  This chapter summarizes our main contributions and discusses promising directions for future research.

Chapter 2

Theoretical Framework

2.1 Introduction

Humans are deeply interdependent and adept at navigating social interactions. Nevertheless, modeling social interaction between humans can be difficult. It stands to reason, then, that modeling social interaction between a human and a computer system is especially complicated. To begin, we review socio-psychological literature which holds that individuals and society are intrinsically interdependent; this interdependence is manifested through social phenomena, such as rapport, trust and interpersonal closeness. From there, we develop a dyadic computational model of rapport to explain how humans in dyadic interactions build, maintain, and destroy rapport via specific conversational strategies that fulfill specific social goals, instantiated in particular verbal and nonverbal behaviors. In a broad view, this theoretical framework will guide us to further design and implement a socially-aware dialogue system.

Rapport, a feeling of connection and closeness with another, feels good, but it also has powerful effects on human performance in a 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 computer agents increasingly take over 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 in human-computer interaction, there have existed no rigorous models of the mechanisms underlying the relationship between social and cognitive functioning in tasks (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 multiple interactions between a human and an agent. In Zhao et al., (2014), we claim that one obstacle to models of this sort is the fact that, as Bernieri and Gillis (2001) have 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 computational 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 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 be destroyed. Drawing on these distinctions has also allowed us to move toward an implementable computational architecture, described in this chapter, that accounts for both participants' cognition, intentions, actions and beliefs, and their interplay, within one person and across the dyad.

In this chapter, our study context is CMU reciprocal peer tutoring, a domain in which rapport has been shown to have a positive effect on student learning (Ogan et al., 2012). Prior work demonstrates that peer tutoring is an effective paradigm that results in student learning (Sharpley et al., 1983), making this a good context to study dyadic interaction with a concrete task outcome. 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 each for 5 one-hour sessions over as many weeks (a total of 60 sessions, and 5400 minutes of data), tutoring one another in algebra (Yu et al., 2013). 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 tutor and tutee roles reversed, and then the final social time. Our student-student data demonstrates that a tremendous amount of rapport-building takes place during the task of reciprocal tutoring (Sinha and Cassell, 2015b).

In what follows we first review prior literature from the social sciences on the components that make up the experience of rapport, the way people assess rapport in others, and the goals and strategies people use to build, maintain and destroy rapport. Bringing these components together, we next propose a model for rapport enhancement, maintenance, and destruction in human-human and human-agent interaction. Our contributions in this chapter are two-fold: (1) an analysis of the social functions and conversational strategies that go into building, maintaining and breaking rapport; and (2) a computationally viable dyadic model of rapport over time built from that analysis.

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. In 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 a virtual agent. The system, however, did not assess the user's level of trust and only built trust through verbal behaviors (small talk, primarily). While the model was successful in building trust – particularly with extroverts – Bickmore and Cassell (2005) later demonstrated the need for incorporating nonverbal behavior. Since then, Bickmore and his colleagues have developed 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 models of 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 construction of engagement (Bickmore et al., 2011). Bickmore and Schulman (2012) relied on accommodation theory to design conversational strategies intended to generate discourse that matches a user's level of intimacy, and then 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. Still, accommodation theory provided a successful means for assessing the user's level of intimacy, which bears keeping in mind for future work. Following 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 targeted 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 attempts to elicit the experience of rapport by working at the level of observable phenomena, coordinating its nonverbal behavior with the human user. Rather than treating rapport as a dyadic or interpersonal construct, they addressed it similarly to other display functions and perhaps not surprisingly, as with other engaging displays, they found increased user engagement. Most recently they 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 ignored significant differences in conversational strategies between interlocutors with different levels of power in the relationship.

In contrast to previous work, we distinguish between conversational goals (overarching goals such as "create rapport" or sub-goals such as "index commonality"), 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, embarrassed laughter, or insults). This tripartite 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 Theoretical Framework for Rapport Management

Tickle-Degnen and Rosenthal (1990)'s work on the changing nonverbal expression of rapport over the course of a relationship has significantly impacted 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 expressions or actions. Tickle-Degnen and Rosenthal (1990), however, propose that the molar level 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 "stereotypical behavior" or "more diversity in the ways they communicate"? How should we represent the goals of interactants and conversational strategies to fulfill the goals? In the current work, then, we discuss research 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.

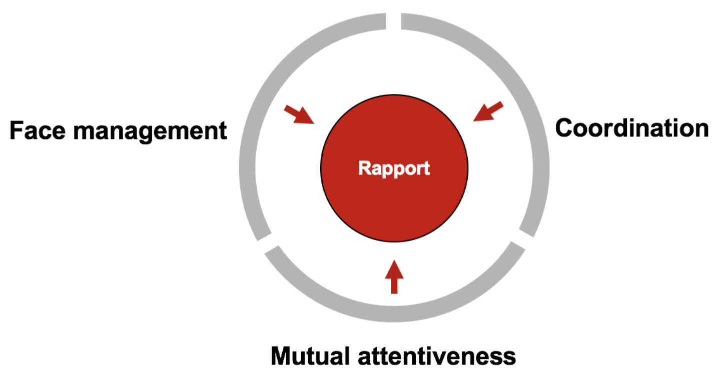

Figure 2.1: Spencer-Oatey's rapport theory

Spencer-Oatey (2005) offers an alternative approach 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 (Yu et al., 2013) support the tremendous importance of face and contains 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 attentiveness, 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. Conversely, *face-boosting acts* can increase self-esteem in the individual, and increase 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 (e.g., 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 (e.g., when one teen asks "wait why do you have to keep your hat on," 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 increase rapport.

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). As a relationship deepens, the breadth and depth of the topics disclosed become wider and deeper, according to social penetration theory (Taylor and Altman, 1987). This helps the interlocutor gain common ground as a basis for an interpersonally-specific set of behavioral expectations. Self-disclosure, however, plays another role in rapport-building, and when successful it is reciprocal (Derlega et al., 1993). In our data, self-disclosure 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 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 data.

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 increase mutual gaze and smiling, and proceed to more intimate topics, such as their poor performance at keeping their pets alive. In fact, Bronstein et al., (2012) found that in a negotiation setting not reciprocating negative self-disclosure led to decreased feelings of rapport. Further, Treger et al., (2013) point out that humor 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," that is, 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 and 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 social 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) suggest 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. 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., 2012). We know that rapport between teacher and student increases learning. When the tutor and tutee are strangers, their behavior complies with sociocultural norms. Impoliteness may challenge rapport by violating those sociocultural behavioral expectations and thus reduce the learning gain in strangers. When the 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 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 as well as 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.

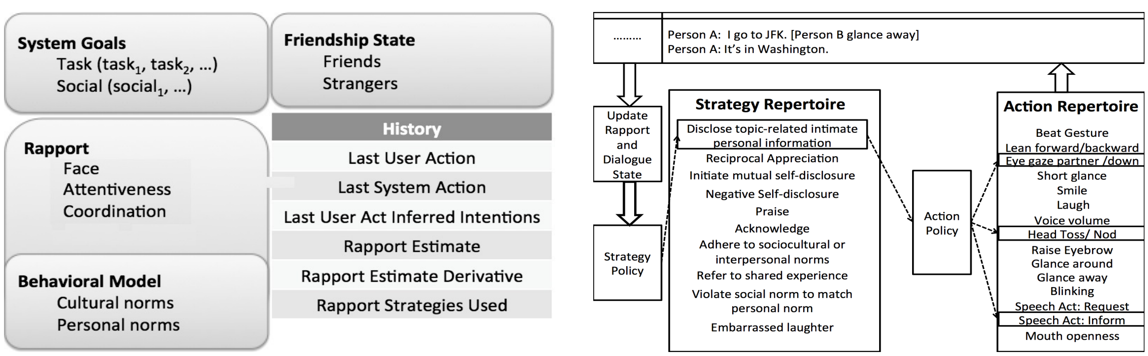

Figure 2.2: Dyadic state (left) and Strategy/Action repertoire (right)

In order to represent these goals and desires in a computational model, (**Zhao, Ran** et al., 2014) stress that while rapport is dyadic, it nevertheless depends on the cognition, actions, beliefs and intentions of each interlocutor, and on the perception by each of these aspects of the mind of the other. In our computational model, therefore, we represent the state of each participant, and of that participant's perception of the state of their 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 incrementally or after each user's turn. Figure 2.2(right) displays how a user and system state lead 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

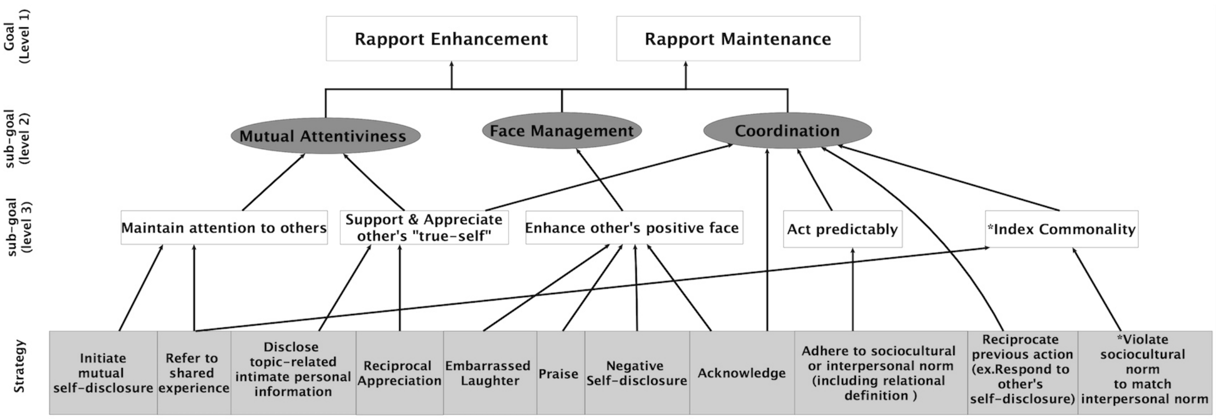

Figure 2.3: Social functions and conversational strategies for rapport enhancement and maintenance

need to detect the goals and conversational strategies of the interlocutors on the basis of observed behaviors, 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 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 incomplete. 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 desire to build rapport with each other. 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 embarrassed 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 behavioral 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-disclosures. 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 increases coordination (**Zhao, Ran** et al., 2014).

In the **rapport-maintenance** orientation (Figure 2.3), people are assumed to begin at state T_2 (Acquaintance) and 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, 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 that detecting it will help us choose appropriate rapport "recovery" strategies. (**Zhao, Ran** et al., 2014) provide more details about rapport-destruction strategies.

2.5 Examples from Corpus Data

In order to demonstrate the functioning of the computational model, six examples are taken from our data, collected in (Yu et al., 2013) (see Table 2.1). In this experiment, 12 dyads of 12-15 year-old students (half boys and half girls, half friends and half strangers) tutor each other in algebra over a period of 5 weeks. Table 1 (left) shows how dyads of strangers interact early in the 5-week period. Table 2.1 (right), shows dyads of friends. Labels indicate how the computational model would generate the same output, based on our annotation of the data for nonverbal behavior and conversational strategies such as disagreement and agreement, politeness and rudeness, and on- and off-task talk.

2.6 Conclusions

In this chapter, 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-human interaction. Based on this framework we have designed a computational model of rapport that can be applied to interactions between humans and virtual agents. In turn, that computational model allows us to proceed towards a dyadic computational architecture for a virtual agent. A first sketch of the details necessary to realize this work computationally is described in the following chapter, in which we lay out the building blocks for our socially-aware dialogue systems.

Stranger-Example 1

P1: b equals nineteen over nine
$[s_1 =$ N/A,$t_d = 1]$
P2: {laughter} good job
$[s_2 =$ praise, $t_d = 1]$
R=Increase

Stranger-Example 2

P1:I suck at negative numbers
$[s_1 =$ negative self-disclosure, $t_d = 1]$
P2: it's okay so do I
$[s_2 =$ reciprocate self-disclosure, $t_d = 1]$
R=Increase

Stranger-Example 3

P1: x equals sixty-four over three
$[s_1 =$ N/A,$t_d = 1]$
P2: yep
$[s_2 =$ acknowledge, $t_d = 1]$
R=Increase
P1: x all right thanks .. all right
$[s_1 =$ adhere to sociocultural norm, $t_d = 1]$
P2: it was a complicated one
$[s_2 =$ face-boosting acknowledgment, $t_d = 1]$
R=Increase

Friend-Example 1

P1: are there any girls you like
$[s_1 =$ elicit self-disclosure, $t_d = 3$
more personal topic]
P2: all of them are not the best looking
$[s_2 =$ reciprocate self-disclosure,$t_d = 3]$
R=Increase

Friend-Example 2

P1:remember you went to Connecticut
$[s_1 =$ Refer to shared experience,$t_d = 2]$
P2:that was just to visit my cousin
$[s_2 =$ disclose topic-related intimate personal information, $t_d = 2]$
R=Increase

Friend-Example 3

P1: silly goose that's a backwards two
$[s_1 =$ violate sociocultural norm to adhere to interpersonal norm, $t_d = 1]$
P2:two
$[s_2 =$ N/A, $t_d = 1]$
R=Increase

Table 2.1: Stranger examples (left) and Friend examples (right) session, where s is a rapport strategy, t_d is topic depth and R is dynamics of rapport. During the first session, most topics are discussed superficially; during the second session, more personal information is disclosed.

Chapter 3

A Knowledge-inspired Socially-aware Personal Assistant (SAPA)

3.1 Overview of Socially-aware Personal Assistant

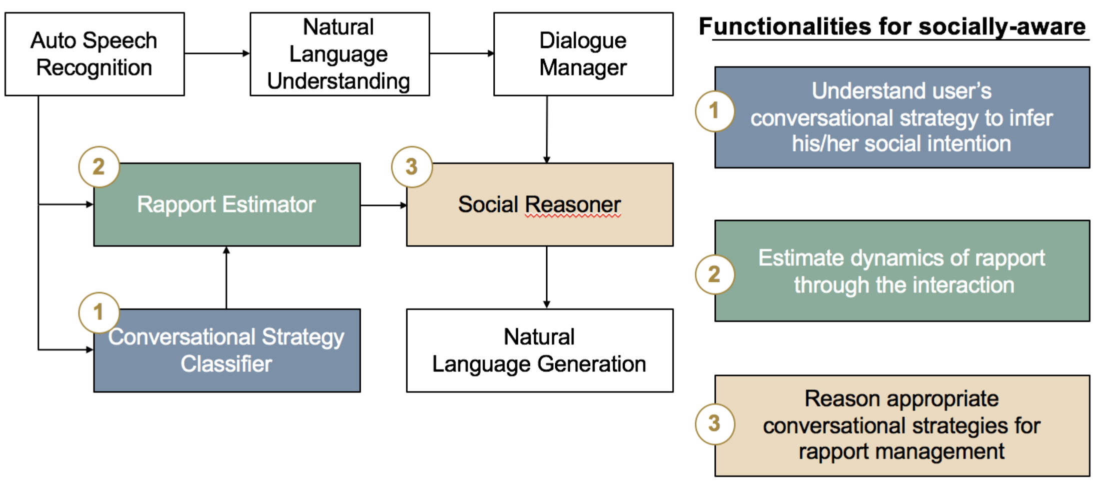

Figure 3.1: Function-level architecture of a socially-aware dialogue system

3.1.1 Introduction

In this chapter, we induce a knowledge-inspired computational architecture to operationalize our theoretical framework from last chapter in a cooperative scenario, which will serve as a personal assistant at a conference. The personal assistant is designed to facilitate rapport with a human user while recommending a session to attend and people to meet. Specifically, the system is designed to leverage rapport to elicit personal information from the user that can be used to improve the helpfulness and personalization of its responses. To achieve this goal, we argue that three augmented functions needs to be added to a traditional dialogue system, as shown in Figure 3.1.

The white boxes represent common modules in a traditional system, whereas the colored boxes indicate three proposed modules to achieve the targeted functions: (1) a Conversational Strategy Classifier is trained to detect the human user's social intention, (2) a Rapport Estimator is used to automatically assess the dynamics of rapport, which forms a feedback loop of rapport management, and (3) a Social Reasoner selects an appropriate conversational strategy to respond to the user's underlying intention. While Figure 3.1 describes our system at the function-level, the following section presents our fully-realized architecture.

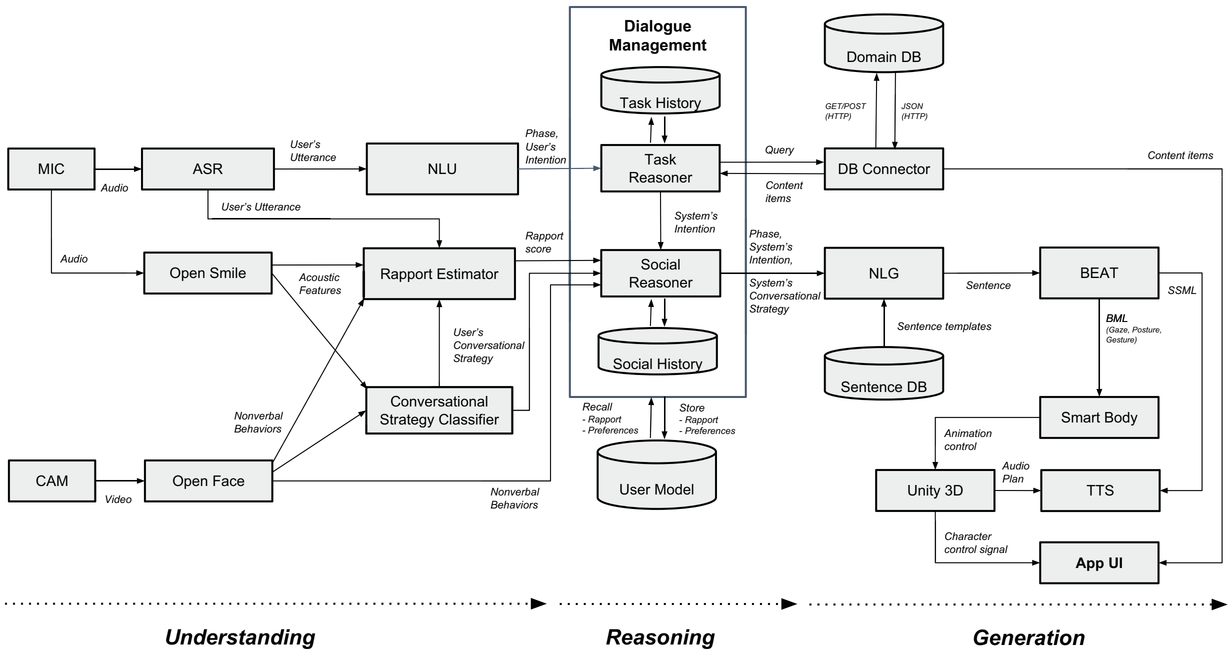

Figure 3.2: Realized full architecture of dyadic rapport (Matsuyama et al., 2016)

3.1.2 Computational Architecture

Figure 3.2 shows an overview of the architecture, which is from our work (Matsuyama et al., 2016). Our developed modules in this thesis are a **Conversational Strategy Classifier** (Chapter 3.2), a **Rapport Estimator** (Chapter 3.3) and a **Social Reasoner** (Chapter 3.4). The modules are integrated with other common components of a dialogue 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 stages, we process verbal, vocal and visual inputs. 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. OpenFace (Baltrušaitis et al., 2016)) detects 3D facial landmarks, head pose, gaze and Action Units. By analyzing the extracted multimodal information of the speaker and interlocutor in both the current and previous turns, our conversational strategy classifier can automatically recognize particular styles and strategies of talking that contribute to building, maintaining or sometimes destroying a budding relationship. Our rapport estimator scores the current relational state between the user and system. In the decision-making stage, a task reasoner focuses on obtaining information to fulfill the user's goals, and a social reasoner chooses ways of talking that are intended to build rapport in the service of better achieving the user's goals. Finally, based on the output of the dialogue 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 behaviors.

3.1.3 Sample Dialogues

We demonstrated our integrated system at SIGDIAL 2016 (Matsuyama et al., 2016). Table 3.1 shows sample dialogues extracted from an actual interaction with the system and annotated with the outputs of the different modules as the system works to meet social and task goals.

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

Table 3.1: Sample dialogues from actual interactions
(**SD**: self-disclosure. **QE**: question elicitation **ASN**: adhere to social norm. **VSN**: violate social norm)

3.2 Predictive Model for Conversational Strategy Recognition

3.2.1 Introduction

Improved understanding of conversational strategies may help us build dialogue systems that exhibit and evoke behaviors not just as conversationalists, but also as confidants to whom we can relay personal and emotional information with the expectation of acknowledgement, empathy and sympathy in response (Boden, 2010). These social strategies improve the bond between interlocutors which, in turn, can improve the efficacy of their collaboration. Efforts to experimentally generate interpersonal closeness (Aron et al., 1997) to achieve positive task and social outcomes depend on advances beyond behavioral channels in isolation and leveraging the synergy and complementarity provided by multimodal human behaviors.

Thus, we strive to better understand conversational strategies in and of themselves, and in employing automatic recognition of conversational strategies to improve the capability of a natural language understanding module to capture the user's interpersonal goals, such as those of building, maintaining or destroying rapport. In this section, we perform quantitative analysis of our peer tutoring corpus. Based on our findings, we propose a machine learning model of conversational strategies. We demonstrate that these conversational strategies are most effectively recognized when verbal (linguistic), visual (nonverbal) and vocal (acoustic) features are all taken into account.

3.2.2 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). 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.

Prior work on quantifying social norm violation has been heavily data-driven (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 dialogues. This work successfully predicted impoliteness and positivity in the next turn of the dialogue. Their success with both annotated and automatically extracted features suggests that a dialogue 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.

In terms of hedges or indirect language, (Prokofieva and Hirschberg, 2014) proposed a preliminary approach to automatic detection, relying on a simple lexical-based search. Machine learning methods that go beyond keyword searches are a promising extension, as they may be able to better capture language used to hedge as a function of contextual usage.

However, a common limitation of the above work is its focus on only the verbal modality, as studies have shown conversational strategies to be associated with specific kinds of nonverbal behaviors as well. 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 (see Chapter 2), which gives us the opportunity to look at conversational strategies as they are used in both a task and social context.

3.2.3 Ground Truth

We assessed our automatic recognition of conversational strategies against the peer tutoring corpus annotated for those strategies. 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. Appendix A and B describe the definitions of each conversational strategy and nonverbal behavior that was annotated. It is worthwhile to note that we achieved high inter-rater reliability for the praise annotation due to its obvious and simple sentence structure (e.g. "you are great!").

3.2.4 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) was 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 was 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, and 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 spoken conversational strategy, we might have multiple types of eye gaze represented. We performed χ^2 test to see whether the appearance of visual annotations was 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. The 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, 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, because self-disclosures are often likely to comprise emotional expressions when revealing one's likes and dislikes (Sparrevohn and Rapee, 2009), we used the LIWC dictionary to capture words that represent negative emotions (d=0.32) and positive emotions (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 considered. Finally, as expected, we found self-disclosure utterances used significantly more 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 their 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, and male and female references (d=0.78), as well as 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 swear 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., 2014) 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 3.2 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 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, students sometimes 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 the 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).

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

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 a student violates social norms, their behaviors are likely to become outliers compared to their normative behaviors. In fact, we noticed a "joking" tone of voice (Norrick, 2003) and different pitch than usual to signal a social norm violation. When the content of the utterance violated social norms, students also tried to lower their voice, which could be a way of hedging these violations. Table 3.2 provides the 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 their partner (o(gP)=0.44) or worksheet (o(gN)=4.29) or gazing elsewhere (o(gE)=0.30) were among the non-verbal behaviors that were significantly present in praise utterances. Table 3.3 provides complete set of results for the speaker (as discussed above) and for the listener.

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	Smile - χ^2(1,1013)=18.63*** - o=1.63
	Gaze (gP) - χ^2(1,1013)=93.04*** - o=2.39	Gaze (gP) - χ^2(1,1013)=131.34*** - o=2.84
	Gaze (gN) - χ^2(1,1013)=35.1*** - o=0.49	Gaze (gN) - χ^2(1,1013)=73.23*** - o=0.38
	Gaze (gO) - χ^2(1,1013)=173.88*** - o=0.29	Gaze (gO) - χ^2(1,1013)=152.12*** - o=0.31
	Gaze (gE) - χ^2(1,1013)=120.77*** - o=1.8	Gaze (gE) - χ^2(1,1013)=78.92*** - o=2.37
2. Shared Experience	Smile - χ^2(1,183)=4.73* - o=1.75	Smile - χ^2(1,183)=7.53** - o=2.07
	Gaze (gP) - χ^2(1,183)=25.37*** - o=3.02	Gaze (gP) - χ^2(1,183)=33.36*** - o=3.59
	Gaze (gN) - χ^2(1,183)=3.73* - o=0.58	Gaze (gN) - χ^2(1,183)=17.68*** - o=0.32
	Gaze (gO) - χ^2(1,183)=27.87*** - o=0.31	Gaze (gO) - χ^2(1,183)=16.55*** - o=0.41
	Gaze (gE) - χ^2(1,183)=38.13*** - o=4.19	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 (gP) - χ^2(1,166)=14.22*** - o=0.39
	Gaze (gN) - χ^2(1,166)=37.52*** - o=4.29	Gaze (gN) - χ^2(1,166)=15.19*** - o=0.33
	Gaze (gO) - N.S	Gaze (gO) - χ^2(1,166)=24.23*** - o=3.30
	Gaze (gE) - χ^2(1,166)=14.44*** - o=0.30	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	Smile - χ^2(1,7469)=869.29*** - o=3.37
	Gaze (gP) - χ^2(1,7469)=911.89*** - o=2.75	Gaze (gP) - χ^2(1,7469)=609.06*** - o=2.27
	Gaze (gN) - χ^2(1,7469)=34.82*** - o=0.8	Gaze (gN) - χ^2(1,7469)=239.22*** - o=0.55
	Gaze (gO) - χ^2(1,7469)=515.26*** - o=0.47	Gaze (gO) - χ^2(1,7469)=110.48*** - o=0.70
	Gaze (gE) - χ^2(1,7469)=195.17*** - o=1.67	Gaze (gE) - χ^2(1,7469)=12.38** - o=1.14
	Head Nod - χ^2(1,7469)=8.06** - o=0.77	Head Nod - χ^2(1,7469)=44.51*** - o=0.56

Table 3.3: 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 a non-verbal behavior is to occur in conversational strategy utterances compared to non-conversational strategy utterances. Significance: ***:$p < 0.001$, **:$p < 0.01$, *:$p < 0.05$

.

3.2.5 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 discussed above. While LIWC features helped categorize words used during a particular conversational strategy, they did not capture contextual usage of words within

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 3.4: Comparative performance evaluation using accuracy (Acc) and kappa (κ) for logistic regression (LR), support vector machine (SVM) and naive bayes (NB)

the utterance. Thus, we also added bigrams, part of speech bigrams and word-part of speech pairs from the speaker's utterance.

In addition to the speaker's behavior, we also added two sets of interlocutor behaviors 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[1] measures. Table 3.4 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 better recognize 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.

3.2.6 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) also positively predicted 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

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

set f_3 predictive of speakers disclosing included no smiling (-0.30), no head nodding (-0.15) and softer 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 ones 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 softer voice (-0.58), smiling (0.47), gazing elsewhere (0.59) or at the speakers own worksheet (0.27) or at the partner (0.22), but not at the 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 the partner's worksheet (0.68) to direct attention to the partner's (perhaps the tutee's) work, and smiling (0.51), perhaps to mitigate potential embarrassment 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 back-channeling and acknowledgement and not gazing at the 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), softer voice (-0.91) and overlap (-0.66) from the interlocutor in the last turn.

Social norm violation: 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 violation to provoke forgiveness from the interlocutor, in addition to indicating hedging. (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 loudness (0.86) and jittery 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.

3.2.7 Post-experiment Analysis

The kappa of detection violation of social norm is comparatively lower than other conversational strategies and slightly better than a random guess. Thus, we conducted a post-experiment analysis and proposed a more advanced model to improve the prediction performance. 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., 2012). In such a sway, we hypothesize that the performance of detecting social norm violation should attribute to the fact that logistic regression fails to model the dialogue 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. Since the appearance of violation social norm is much more than other conversational strategies, in this section, we did not rebalance the dataset but directly trained our model with the original dataset. Using long-term temporal and contextual information, our model achieves an F1 score of 0.705 comparing to 0.578 of baseline logistic regression model.

Model

We treated a dialogue 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 $([w^i_0, ...w^i_m], e_i)$, where $[w^i_0, ...w^i_m]$ 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 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 norms. 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 norms, given the entire history of the dialogue.

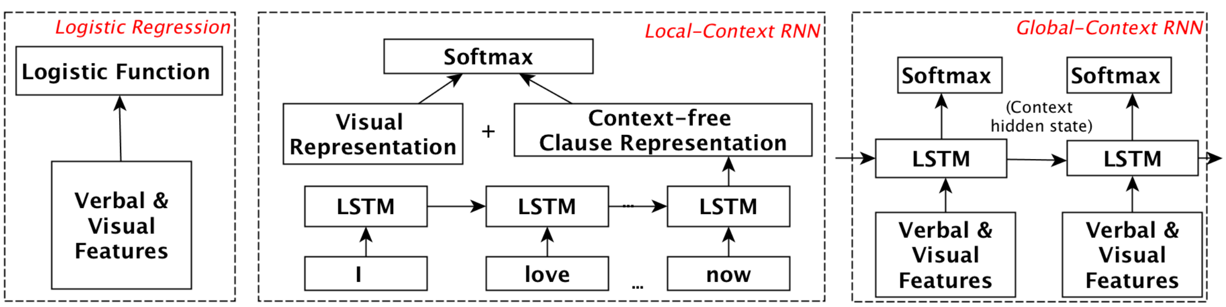

Figure 3.3: Three proposed computational models.

Logistic Regression Model We first trained a L2 regularized logistic regression model using the proposed verbal and visual features f_i as inputs (Figure 3.3, left). This model serves as our baseline.

Local/Global-Context RNN Model Past empirical results suggest two possible hypotheses of improving the model performance: (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 the 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 3.3, 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 dialogue. The average number of clauses in training dialogue 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 Results

We observed that Global-Context RNN with 2 LSTM layers outperformed other models as showed in Table 3.5. First, by comparing the logistic regression model with our best model, the result indicates the strong predictive power of long-term temporal contextual information in detecting social norm violation in dialogue. On the other hand, Local-Context RNN model did not achieve significant improvement on overall performance regarding 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 this could be an insufficient training set to learn generic clause representation. To avoid the issue of overfitting, after 2 layers, we did not further complicate our model.

3.2.8 Conclusions

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

Model	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**

Table 3.5: Performance comparison for the 3 evaluated models

We found that interlocutors typically accompany 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 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 without nodding when they violate a social norm. 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 fulfill interpersonal goals in a dialogue.

We acknowledge some methodological limitations in the current work. Specifically, we under sampled the negative examples to create a balanced dataset. In our future work, we will work with corpora that have a more natural distribution and deal with the sparsity of the phenomena through machine learning methods. This will improve applicability to a real-time system where conversation strategies are likely to be less frequent than in our training dataset. Moreover, in the current work, we looked at individual modalities in isolation initially, and fused them later via a simple concatenation of feature vectors. Including sequentially occurring features may better exploit correlation and dependencies between features from different modalities. In the following section, we have thus started to investigate the impact of temporal ordering of verbal and visual behaviors that lead to increased rapport (**Zhao, Ran** et al., 2016b).

3.3 Predictive Model for Rapport Assessment

3.3.1 Introduction and Motivation

As we introduced in Chapter 2, conversational strategies in our computational model of rapport 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. In this section, 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 account for interaction among different modalities and between interlocutors. There are three key issues that we believe should be considered when performing such assessment.

(1) When the foundational work by (Tickle-Degnen and Rosenthal, 1990) described the nature of rapport, three interrelated 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 the relationship between interlocutors, and therefore it is necessary to account for the relationship status of a dyad when extracting patterns of rapport. (2) while (Ogan et al., 2012) discovered some of the common behaviors exhibited by dyads in peer tutoring to build or maintain rapport (e.g. playful teasing, face-threatening comments, attention-getting, etc.), tutors and tutees were looked at separately, and each of these behaviors was examined in isolation. We are interested in moving beyond individual behaviors to focus on temporal sequences of such behaviors in the dyadic context. Likewise, (Ogan et al., 2012) did not distinguish between rapport management during task (tutoring) vs social activities. We suspect that interactions between verbal and nonverbal behaviors manifest differently in social and tutoring periods, since the roles of a tutor and tutee are more evident in 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-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.

3.3.2 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'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).

For instance, (Kang et al., 2012) presented a study of co-occurrence patterns of human nonverbal behaviors during intimate self-disclosure. However, contingent relations between different nonverbal behaviors was not considered, which could extensively contribute to the design of a social agent that interacts with a human over time. (Wörtwein et al., 2015) learned behavioral indicators that were correlated to expert judges opinions of each key performance aspect of public speaking. They fused the modalities by utilizing a least squared boosted regression ensemble tree and predicted speaker performance. However, this work also did not consider the effect of interactions among different modalities and their temporal relations. Similarly, (Chiu et al., 2015) introduced deep conditional neural fields to model the generation of gestures by integrating verbal and acoustic modalities, while using an undirected second-order linear chain to preserve temporal relations between gestures. However, this approach only modeled individual co-verbal gestures, without considering interaction between the speaker and the interlocutor.

In (Heylen et al., 2007) temporal combinations of individual facial signals (such as nod, smiles, etc.) were used to infer positive (agree, accept, etc.) and negative (dislike, disbelief, etc.) meanings via human ratings. An interesting take-away from this work was that a combination of signals could significantly alter the perceived meaning. For instance, facial tension or frown alone did not mean "dislike", but the combination frown and tension did; tilt and gaze right down alone did not mean "not interested" as significantly as did the combination of tilt and gaze. However, while a combination of these nonverbals signaled higher level constructs (that were in turn associated with some pragmatic meaning), the authors were more interested in how these combinations were perceived by humans, rather than necessarily in a predictive task or testing these combinations in a human-agent dialogue.

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 "hmm", 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 works are discussed in (**Zhao, Ran** et al., 2016b).

For instance, (Gravano and Hirschberg, 2009) looked at observable lexical, acoustic and prosodic cues produced by the speaker, followed by backchanneling from the listener. The authors found that the likelihood of occurrence of a backchannel from the interlocutor appeared to increase with the simultaneous occurrence of one or more cues by the speaker, such as final rising intonation, higher intensity and pitch levels, longer inter-pausal units (maximal sequence of words surrounded by silence longer than 50 ms), etc. However, in this work, no attempt was made to use the temporal sequence or co-occurrence of observables preceding a backchannel to predict higher-level social constructs such as positivity, coordination, attentiveness, or underlying psychological states such as rapport or trust.

(Allwood and Cerrato, 2003) explored the interplay between head movements, facial movements like smile and eye brow raising, and verbal feedback in a range of conversational situations, including continued attentiveness, understanding, agreement, surprise, disappointment, acknowledgment and refusing information. As the situations became more negative (disappointment, refusing information), the accompanying nonverbals became more extensive - no longer just a head nod, but a series of movements. The authors claim that this series of movements added extra information or emphasized or contradicted what had been said, but ground truth was not provided for these claims.

Finally, (Chollet et al., 2014) used sequence mining methods to automatically extract nonverbal behavior sequences of the recruiters that represented interpersonal attitudes. Then, Bayesian networks were deployed to build a generation model for computing a set of nonverbal sequence candidates, which were further ranked based on the previously extracted frequent sequences. Even though this work considered the effect of sequencing of nonverbal signals, their model could be improved by adding temporal information inside these sequences or verbal signals, and modeling listeners' behaviors as well.

3.3.3 Study Context

In this study, we conduct our experiment on CMU reciprocal peer tutoring dataset (Yu et al., 2013), as 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). Appendix A and B describe the definitions of each conversational strategy and nonverbal behavior that was annotated.

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 annotator bias, account for label over-use and under-use and select a single rapport rating for each 30-second video segment. The segments were randomly presented to the annotators 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) of the dyad (Sinha and Cassell, 2015a,b) and that greater likelihood of being in high rapport in the next 30-second segment (improvement in rapport dynamics over the course of the interaction) is positively predictive of the dyad's problem-solving performance.

3.3.4 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 a 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 $E_i = \{t^i_j \in \mathbb{R}\}$ 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 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.

$$confidence_r = P((t \in E_1)|(t' \in E_2), t' - t \leq w) \tag{3.1}$$

$$support_r = \frac{\{\#e_2|\text{r is active}\}}{\#e_2} \tag{3.2}$$

$$precision_r = \frac{1}{\text{standard deviation}_r} \tag{3.3}$$

3.3.5 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 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 3.4, 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, referring to shared experience, smiling while violating a social norm and overlapping speech are associated with high rapport. Examples are:

FH 1 *One student smiles while the other violates a social norm (Social period)*

FH 2 *One student 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, after 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, after which the tutor smiles and violates a social norm (Task period)* **[shown in Figure 3.4]**

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 violates a social norm, 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 worksheet 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 interlocutor 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 interlocutor 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 3.5]**

3.3.6 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 that 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 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

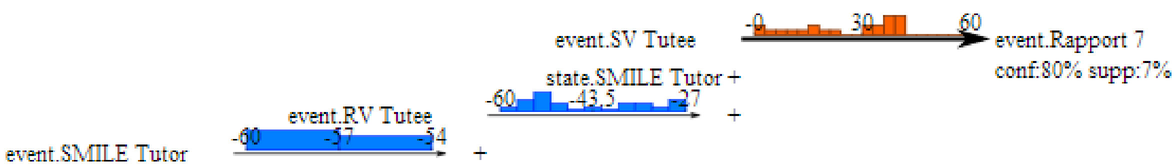

Figure 3.4: Friends in high rapport - The tutee reciprocates a social norm violation while overlapping speech with the tutor, after which the tutor smiles while the tutee violates a social norm.

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 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;

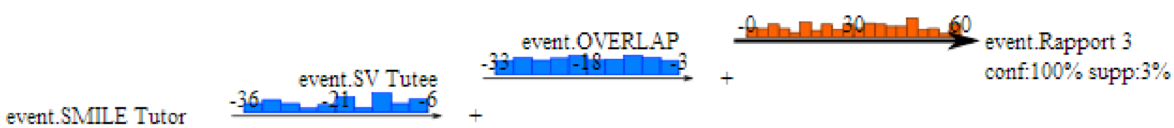

Figure 3.5: Strangers in low rapport - The tutor smiles and the tutee violates a social norm within the next 30 seconds, before their speech overlaps within the next 30 seconds.

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 exactly the same{smile}, 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.

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

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 number of conditions: 5, minimum use: 10. Second, we fused those rules based on the algorithm discussed above and applied them to the 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 3.6 suggest that the Titarl-based regression method has a significantly lower mean square error than the naive baseline linear regression method. The high effect size in both strangers (d=-0.62) and friends (d=-0.42) further prove the substantial improvement on accuracy of assessing rapport by Titarl-based regression compared 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 violation and serving to provoke forgiveness from the interlocutor. Overlap in this context may index 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. This also represents an advance on work by (Huang et al., 2011) that improved rapport through nonverbal and para-verbal channels, but did not take linguistic information or temporal co-occurrence across modalities into account. We included linguistic information in our rules and in the previous section we have shown that the linguistic information (conversational strategies) that formed an essential part of the temporal rules presented here can be automatically recognized (**Zhao, Ran** et al., 2016a). As noted above, while our current work focused on developing an interpretable and explanatory model of temporal behaviors to serve as a building block for our socially-aware (rapport) dialogue systems, the framework can be applied for prediction of other social phenomena of interest in virtual agent systems (such as trust and intimacy), in domains as diverse as survey interviewing, sales, and health.

3.3.7 Conclusions and Future Work

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. We benchmarked our Titarl-based regression approach with a linear regression model which does not capture the sequence and temporal ordering of different behaviors. The baseline seems a bit weak due to the absence of modeling temporal information. To address this limitation, we will consider other baseline models that include temporal information such as a recurrent neural network model, auto-regressive model, etc. Another limitation of the current work is the lack of interpretability. Given that we extracted a very large number of patterns and fused them to predict the rapport state, it is non-trivial to explain the underlying reasons of the phenomena by tracing back to specific rules. In the future, we would like to shrink the space by selecting the most representative rules (e.g. through KNN method) for the sake of interpretability. 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.

3.4 Conversational Strategy Planning for Social Dialogue

3.4.1 Introduction and Motivation

In the previous chapter, we focused on endowing a system with the capability of detecting social intentions and understanding dynamics of rapport in conversation. In this chapter, we present a novel decision-making module that allows a socially-aware dialogue system to reason about appropriate response to social intentions. Conventionally, a dialogue manager pick ups the next system action with respect to the task. Similarly, we develop a module named social reasoner that focuses on managing relational bonds with human users through reasoning the usage of the conversational strategy for the system. Specifically, here we are interested in seven common conversational strategies shown to positively impact rapport (Tajfel and Turner, 1979; Spencer-Oatey, 2008b)(some of them are included in the previous chapters): Self-Disclosure (**SD**), revealing personal information to decrease social distance; Question Elicitation of Self-Disclosure (**QESD**) to encourage the other interlocutor to self-disclose; Reference to Shared Experiences (**RSE**) to index common history; Praise (**PR**) to increase self-esteem in the listener and therefore interpersonal cohesiveness; Adhere to Social Norm (**ASN**) to increase coordination by adhering to behavioalr 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**) 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 is fulfilled by our trained multimodal *Conversational Strategy Classifier* introduced in Chapter 3.2. The second prerequisite is fulfilled by our temporal association rule-based *Rapport Estimator* explained in Chapter 3.3.

The Social Reasoner module takes 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 generates appropriate social conversational responses with the system's goal of always keeping rapport high in order to increase trust and long-term engagement. While there are several potential approaches, most 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 whose results are almost intractable to solve with a purely symbolic approach such as production rule systems or classic planners. On the other hand, (Romero et al., 2017) 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.

3.4.2 Related Work

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

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

Similarly, (Coon et al., 2013) focused 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 specific activities for each stage, it is a challenge to scale up their algorithm.

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

3.4.3 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 a) to 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.

Module Descriptions

The Social Reasoner's architecture is depicted in Figure 3.6. 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 chooses 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 dialogue systems and conversational agents, such as ASR, NLU, NLG, etc.

The 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.

3.4.4 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 continuous domains. Behavior selection is modeled as an emergent property of activation/inhibition dynamics among all behaviors. A

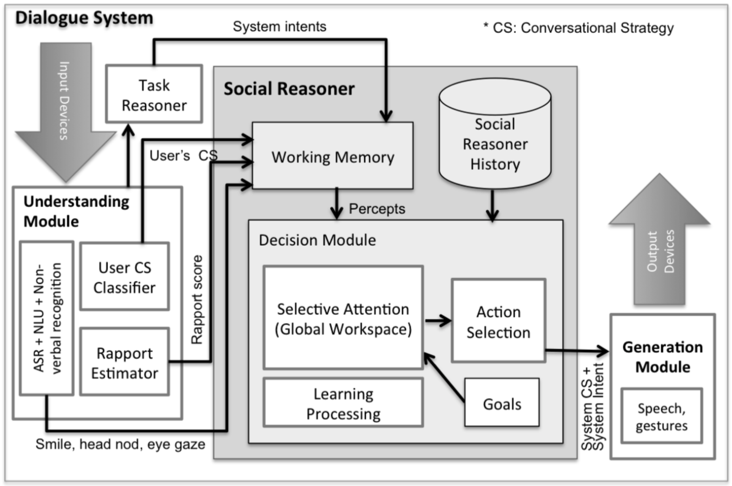

Figure 3.6: System architecture

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 a member of the add list of i and a 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 the 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.

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.
dialogue 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.

Table 3.7: Pre-condition and post-condition categories

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 3.7, and post-conditions are defined in terms of what the expected states are after performing the current conversational strategy (e.g., rapport score increases, user smiles, etc.). This kind of chain reasoning based on linked pre-conditions and post-conditions endows the system with plan-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 prevent incorrect conversational strategies from being triggered. The Social Reasoner is adaptive enough to respond to unexpected user 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.

3.4.5 Design of the Decision-Making module

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.

1. 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 norms 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.

2. Data-driven Sources

Data-driven discovery by temporal association rule: (**Zhao, Ran** 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, and at later stages by friend rules. For instance, one rule that strangers followed was: *one interlocutor 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 conference guide that recommends sessions to attend and people to meet. In each session, a dyad consisting of a user and the virtual assistant (using a Wizard-of-Oz protocol) interacted through a dialogue 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 WoZ-er's assessment of the 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.

Encoding of Pre-conditions & Post-conditions

(Romero et al., 2017) 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 pre-condtions and post-conditions (from Table 3.7) based on the theoretical foundations provided in Section 3.4.5; 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 statistical analysis, more specifically, a bidirectional elimination stepwise regression that allowed us, through a series of partial F-tests, 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 C.

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 $\delta = 75$.

3.4.6 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.

Experiment 1: Social Reasoning Validity

H_0 : *Social Reasoning doesn't contribute significantly to building rapport and increasing conversational engagement compared to traditional dialogue systems.*

For this experiment, we divided the WoZ study dataset of 228 sessions (Section 3.4.5) into two groups: dialogue turns that used conversational strategies and dialogue turns that did not use any conversational strategy (plain behavior). Then, we observed the rapport score (1-7), our variable of interest. We ran a one-way ANOVA analysis in order to determine whether there is a statistically significant difference between the groups at $p < .05$. The ANOVA is shown in Table 3.8.

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

Table 3.8: ANOVA for experiment 1.

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 used social reasoning was higher ($5.65 \pm 0.4, p = .032$) in comparison with the group that used 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 building rapport. Likewise, we conclude that using social reasoning may improve social bonds (rapport) on a 35.4% during a conversation.

Experiment 2: Social Reasoner's Accuracy

H_0 : *Data-driven tuning process does not improve the Social Reasoner's accuracy*

In 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 n one-way ANOVA analysis. Results are shown in Table 3.9.

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 ($4.83 \pm 0.5, p = .027$) compared to the group that only used a theoretical-based design ($3.05 \pm 0.4, p = .033$). 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 the Social Reasoner up to a 25.4%.

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

Table 3.9: ANOVA for experiment 2.

Experiment 3: Social Reasoner's Performance

In this experiment, we chose four well-characterized conversational sessions from the dataset log files in the post-experimental evaluation to test system performance. Below we describe 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 becomes 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 loses interest.

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 3.10: Social Reasoner's performance. MSE rate: [1 - ($MSE_{SR} \div MSE_{TD}$)]

Table 3.10 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 dialogue system that doesn't use conversational strategies) and MSE_{SR} (a dialogue system that uses our Social Reasoner). It is important to note that, for the experiments executed, the proposed Social Reasoner model improves performance results obtained by a traditional dialogue system a rate between 20% and 34%.

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

Intuitively, one can deduce that the Social Reasoner emergently tailors a plan as the combination of SD, PR and QESD strategies when it detects that 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 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 these kinds of strategies, even if they are not strategies with the highest activation level. ACK is more likely to appear when there is evidence of progressive increase in use 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 continues to accumulate energy during the whole interaction so it can be easily triggered if the system realizes that a previous action (because of a particular CS) diminishes the rapport level. Finally, PR is continually used when the Social Reasoner detects no significant changes in the user's verbal and non-verbal behaviors that can increase rapport, especially when another conversational strategy such as SD and QESD has been used without success.

3.4.7 Conclusions

In this chapter, we proposed a hybrid adaptive Social Reasoner component that determines which conversational strategy should be used 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 consequently expand the number of possible combinations between matched pre-conditions and triggered conversational strategies. In our future work, we hope to: 1) continue collecting data from user interactions to fine tune the system and improve its performance; and 2) explore an alternative to learn and adjust pre-conditions and post-conditions. Rather than using a fixed set of pre-conditions and post-conditions we will use our data-driven model as a cold-start solution while more suitable pre-conditions and post-conditions are discovered over time by a learning process that may personalize the interaction with the user. One approach is to assign weights to pre-conditions and post-conditions based on saliency properties observed from data. That is, during the stepwise regression analysis some variables produced stronger effects on spreading activation dynamics than others, for instance, variable "past–experiences–available" had a stronger effect on RSE than "low–rapport", so the former could have a weight of, e.g., 0.93 while the latter a weight of 0.12. RSE would then be triggered faster when the former variable is present. After that, weights could be adjusted through reinforcement learning.

Chapter 4

A Neural Social Intelligent Negotiation Dialogue System (SOGO)

So far, we have applied our socially-aware framework to a personal assistant scenario that stems from a fully-cooperative interpersonal relationship. The results indicate that our proposed conversational strategies are effective at building rapport within our theoretical framework. Thus, to explore the generalization capabilities of our framework, in this chapter, we focus on negotiation, a field that blends cognitive (task) and interpersonal (social) skills (Gratch et al., 2015) in a semi-cooperative environment where both interlocutors' relationship and negotiation tactics are equally important. Our studies on socially-aware personal assistant (SAPA) have shown that reasoning an appropriate conversational strategy as a response relies on understanding the user's social intention, the dynamics of rapport, as well as the learned strategy policy for discourse planning. However, as we discussed in previous chapters, the architecture of socially-aware personal assistant (SAPA) to achieve those functions is complicated. More important, each module must be tweaked separately when moving to a new domain. Therefore, we embraced a neural approach to replicate those functions, which will significantly flatten our technology stack and learn the whole model in one learning pass.

4.1 Introduction and Motivation

Whether deciding between a salad or fast food, or asking a coworker to complete a project ahead of schedule in exchange for help later, we negotiate every day. When we hold conflicting interests, we must negotiate to pursue our ultimate goals. Thus, negotiation is an act wherein participants with unique motives cooperate and compete to maximize their own benefits. Unfortunately, most of us are poor negotiators. Prior literature has documented a range of cognitive biases that undermine the quality of agreements in human-human negotiation (Thompson and Hastie, 1987). By helping people avoid these limitations, virtual agents have proven to be powerful tools for teaching negotiation skills and modeling negotiation (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. Further, most studies regard negotiation solely as a 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. Negotiators are encouraged to share crucial information and cooperate to reduce the risk of impasse and build *rapport*, a feeling of connection and closeness with another (Bronstein et al., 2012; 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. Therefore, in this study, we leverage different linguistic devices to build rapport between a human and a dialogue system, which thus fosters integrative agreement during negotiation.

Although face-to-face dialogue is the most natural interaction, it is difficult to collect and access to high-quality multimodal data is limited. Meanwhile, the result from our previous study of SAPA and prior literatures (Bronstein et al., 2012) highlight the contribution of the verbal channel on rapport building. Therefore, we focus on verbal strategies within our theoretical framework on building rapport in the context of human-agent negotiation. We believe that social conversation 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., 2017; **Zhao, Ran** et al., 2014). 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 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?").

In this study, we move from a knowledge-inspired approach to a neural approach for three reasons. First, the knowledge-inspired approach requires significant human effort to define symbolic representation. Instead, the neural approach learns a specific neural space where linguistic information is automatically implicitly encoded using low-dimensional continuous vectors. Second, the architecture of our SAPA is extremely complicated. More important, it is hard to implement and extend to other domains. The neural approach will help us dramatically flatten the technology stack so that reapplying model to other tasks with one learning pass. Third, unlike a knowledge-inspired dialogue system, which consists of different key components that separately optimize their specific objective function, the neural dialogue model adopts one objective function for end-to-end learning. The downside of neural approach is that we need to collect intensive data to train the model. Since our research objective is to investigate the effectiveness of verbal strategies in building rapport, it is sufficient and efficient to use a text-based system setup for data collection. As humans can only access the verbal channel, we believe they will encode all propositions and interpersonal functions in linguistic devices, which aligns with our research goals.

The remainder of this chapter is organized as follows: In the next section, we briefly describe the context of the study. We then review prior studies related to negotiation and rapport agents. In Section 4.4, we develop a semi-automated system (SOGO 1.0) for conducting a Wizard-of-Oz study on human-computer negotiation to validate the effectiveness of our socially-aware framework in a negotiation scenario. We also collected data from human expert behaviors as the first step toward a fully automated social intelligent negotiation system (SOGO 2.0). Consequently, in Section 4.5, we introduce a new paradigm of training a socially-aware neural dialogue model that leverages expert supervision to automatically blend task and social goals for a fully-automated system. We combine subjective and objective metrics to evaluate the system's performance in negotiation and rapport building.

4.2 Study Context

Since a neural approach requires intensive data to train the model, in this chapter, we utilize a human-human negotiation dialogue corpus(Lewis et al., 2017) constructed by Facebook AI Research (FAIR). 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 5-7 total items. Participants were each given different and unseen value functions where the maximum value to each side was 10. The unique value function for each item was constrained to ensure that both participants could not receive its maximum value. Table4.1 summarizes the negotiation corpus.

Metric	Dataset
Number of Dialogues	5808
% Agreed	80.1
Average Turns per Dialogue	6.6
Average Words per Turn	7.6

Table 4.1: FAIR negotiation corpus stats

4.3 Related Work

In this section, we describe research on building a negotiation agent and a social intelligent agent.

4.3.1 Negotiation Agent

Modeling negotiation in an agent has become a popular area of research. Different negotiation frameworks(Hindriks et al., 2009; Mell and Gratch, 2016; Fabregues and Sierra, 2011) have been established to serve as benchmarks for evaluating practical negotiation strategies. For instance, the generic negotiation framework Genius(Hindriks et al., 2009) facilitates research in bilateral multi-issue negotiation. Alternately, IAGO (Mell and Gratch, 2016) allows a human to negotiate with a multimodal virtual agent, whose facial expressions and nonverbal cues are accessible to its negotiating partner. Unlike the single-agent Genius and IAGO frameworks, DipGame (Fabregues and Sierra, 2011) is a multi-agent system developed on the Diplomacy Game, where negotiation and relationships between players are essential for success. These frameworks provide a testbed for researchers to investigate different negotiation tactics and social strategies in human-agent interaction.

Prior research indicates that modeling negotiation in virtual humans requires the system to demonstrate cognitive skills for reasoning and social skills for communication, as well as express its internal state (Gratch et al., 2015). Gratch and his colleagues introduced a Conflict Resolution Agent (CRA) that allows students to engage with virtual human role-players across multi-issue bargaining problems. They showed that participants perceived the same levels of satisfaction, cooperation and rapport when paired with a CRA as with a human. Furthermore, they confirmed that the CRA could help people improve their negotiation skill through practicing interpersonal skills. To build on these promising findings, subsequent research investigated ways to design an agent with better communication and negotiation skills. For instance, (DeVault et al., 2015) employed a wizard-controlled system to improve turn-taking skills, and (Lucas et al., 2016) investigated deceptive strategies to optimize one party's benefit while preserving the illusion of fairness towards the other party to improve negotiation policy. The analysis demonstrates that deceptive strategies increase feelings of satisfaction, trust and fairness from one's partner, which seems more credible and even preferential to honesty. Meanwhile, this work also ameliorates an agent's capability to understand human behavior by identifying multimodal signals of trustworthiness in human-human negotiation (i.e., different sets of behavior indicators for predicting objective trustworthiness and perceived trustworthiness). This finding enables the system to infer human negotiation behaviors and display specific behavioral signals on demand to manipulate trustworthiness. Additionally, (Johnson et al., 2017) developed an autonomous agent that teaches people to negotiate through visualized feedbacks. Basically, the researchers encoded theoretical negotiation principles into several automatic quantifiable metrics that have been validated in pedagogical negotiations.

Recent work in negotiation agent algorithms has moved away from a traditional rule-based expert system to statistical machine learning based system. This new method reduces human effort extensively and promises to build a fully-automatic agent. The work on which our task module is based (Lewis et al., 2017) introduces an end-to-end model that employs a reinforcement learning algorithm that was shown to have learned both linguistic and reasoning skills without human annotations. However, this model is perceived to be too uncompromising and aggressive: it suffers from a low agreement rate with human users. Our work addresses this problem by building interpersonal rapport through conversational strategies, which could increase the human user's tolerance to his or her agent partner. Alternately, (Sun et al., 2018) suggests designing a benevolent agent based on a reward-shaping method that diminishes rewards that make the agent feel less satisfied for consecutive rewards. In addition to the reward-shaping method, (Lerer and Peysakhovich, 2017) try to solve social dilemmas by exploiting the tit-for-tat (TFT) strategy. Specifically, they utilize a reinforcement learning method to approximate the cooperative and defect policies as well as the switching policy. This work encourages the agent to become more adaptive to solve social dilemmas in many environments.

Although promising, this prior research is limited by a focus on modeling negotiation in a single interaction; it does not leverage the collected knowledge across the interaction to facilitate the process. Also, these studies solely consider and optimize the negotiation outcome but ignore the social outcome and its contributions to the task. Furthermore, most negotiation agents communicate with users through pre-defined symbols, not natural language, which detracts from the naturalness of the interaction and removes social cues from language.

4.3.2 Social Intelligent Agent

Many prior studies have addressed the issue of rapport between human users and virtual agents. An early paper (Bickmore and Cassell, 2001) implemented a Real Estate Agent to build trust with users while selling houses. Based on sociolinguistics and social psychology, its computational architecture blended small talk and task talk in conversation, and demonstrated that social dialogue affects trust for extroverted users. Elsewhere, (Coon et al., 2013) created an algorithm to plan appropriate activities based on the user-agent relationship, ranging from stranger to companion. Each activity had its own required closeness level, and an activity planner optimized its activities to increase intimacy over time. Similarly, (Bickmore and Schulman, 2012) categorized relationships into four intimacy levels and defined a set of dialogue acts for each level. Inspired by accommodation theory, the researchers modeled long-term user-agent relationships in the context of health counseling. Their proposed reactive algorithm selects the most appropriate dialogue act for the assessed intimacy level yet fails to increase intimacy. (Gratch et al., 2006) offers an alternative way of building rapport. They developed a sensitive listener agent that generates listening feedback behaviors (e.g. head nod, eye gaze) to effectively coordinate with human users. In subsequent studies, (Huang et al., 2011) proposed a next generation rapport agent capable of enhancing mutual attention by predicting backchannels, increasing coordination by generating smooth turn-taking behaviors and maintaining positive communication through affective response and facial expressions.

Unfortunately, like research in negotiation agents, this work does not model the user across interactions and use that knowledge to build interpersonal bonds.

4.4 SOGO 1.0: A Semi-automated Socially-aware Negotiation System

4.4.1 Two-phase Computational Model

Our work does not seek to develop negotiation tactics but rather to use rapport management strategies to facilitate the negotiating process. Thus, in SOGO 1.0 system, we do not develop our own negotiation reasoning module but adopt the end-to-end negotiation dialogue model from Facebook AI Research (FAIR) as our task reasoning backbone. Because this neural dialogue model allows users to negotiate with the system through natural language, we can insert social moves into dialogues. Further, FAIR leveraged reinforcement learning to maximize the outcome

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 1.0 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 4.2: Sample human-agent negotiation dialogue of the FAIR model (top) and our SOGO 1.0 model (bottom)

of negotiation rather than imitating the actions of human users in supervised learning, and replaced traditional likelihood decoding with a rollout algorithm, a kind of beam search conducted in the dialogue 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 semi-automated social intelligent negotiation dialogue system (SOGO 1.0) to overcome these limitations and improve task performance while constructing an interpersonal bond with users. To compare outputs from both systems, refer to the sample dialogues in Table 4.2.

In the following, we outline the details of our Wizard-of-Oz study and 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 dialogue 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. Subsequently, a human expert decides the final sentence to be uttered by the dialogue system. Figure 4.1 provides an overview of two-phase method.

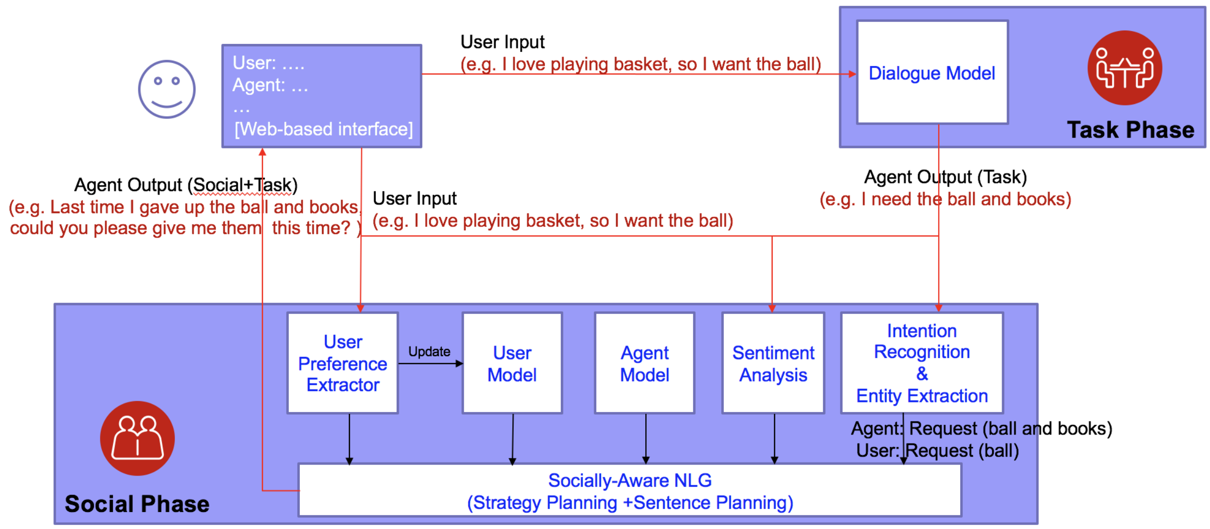

Figure 4.1: Overview of the two-phase method

Phase 1: Task Phase

(Lewis et al., 2017) utilize two-stage learning strategies by pre-training the model with supervised learning, then fine-tune the parameter using reinforcement learning. In this section, we briefly discuss their advanced reinforcement learning with dialogue rollouts decoding based model which we deploy in our study. Each dialogue D is represented as a set of tokens 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 user and agent. The agent has an input goal g and generates the negotiation outcome o_i. We keep the structure of their four GRU-based recurrent neural networks: GRU_g (Agent's goal encoder), GRU_w (dialogue 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) model the dependencies between language and input goals with the function (4.1):

$$p_\theta(x_t|x_0...t-1,g)) \tag{4.1}$$

Conditioning the input goals and generated dialogue, they predict negotiation outcomes with the function (4.2)

$$p_\theta(o_i|x_{0...T},g)) \tag{4.2}$$

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

$$L(\theta) = -\sum_{x,g}\sum_{t} logp_\theta(x_t|x_0...t-1,g)) - \alpha\sum_{x,g,o}\sum_{j} logp_\theta(o_i|x_{0...T},g)) \tag{4.3}$$

wherein α is a hyperparameter to balance token prediction loss and outcome prediction loss. Based on the negotiation outcome at the end of each dialogue, 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 dialogue rewards μ, the objective is to optimize the expected reward of each

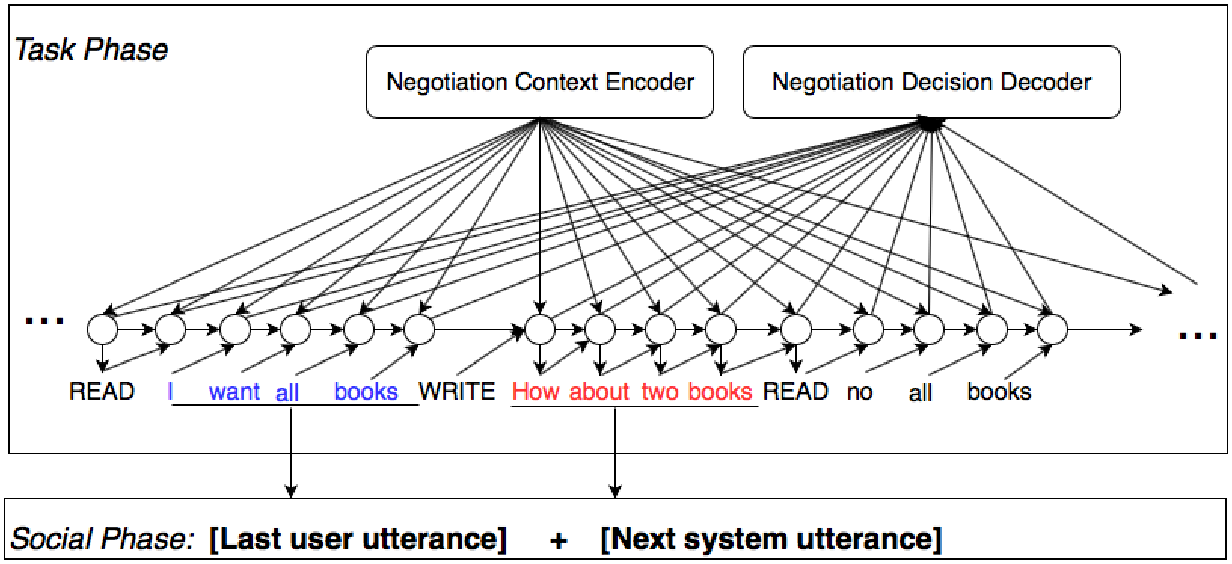

Figure 4.2: Task phase

token generated by the agent as follows:

$$L_\theta^{RL} = \mathbb{E}_{x_t \sim p_\theta(x_t|x_{0...t-1,g}))}[\sum_{x_{t...T}} \gamma^{T-t}(r(o) - \mu)] \tag{4.4}$$

In the decoding stage, a dialogue 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:

$$u* = \underset{u_i = x_{n,n+k}, u_i \in U}{argmax} (\mathbb{E}_{x_{(n+k+1...T;o)} \sim p_\theta}[r(o) p_\theta(o|x_{0...T})]) \tag{4.5}$$

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

Phase 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, dialogue 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 & Entity Extraction

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 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 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 (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 dialogues(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 widely used in negotiation.

- Elicit preference question: ask questions about the opponent's preferences that maximize information gain.
- 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 of ten MTurkers passed the qualification test, i.e., completed previous tasks with more than 80% accuracy. 2,500 dialogues were annotated and used to train our speech act classifier, which served to annotate the rest of the corpus.

Speech Act Classifier We leveraged 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 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 can address this issue by generating better word embeddings for rare words and even those out-of-vocabulary words since it constructs the embeddings at 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 our corpus contains a limited vocabulary, fastText fits well to our case.

User Model & Agent Model

Both the user model and agent model contain the dialogue history and context across the interactions that serve as long-term memory in human-agent interaction. Meanwhile, this memory offers

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 4.3: Performance evaluation of our speech act classifier

the dialogue content for specific conversational strategies (e.g., reference to shared experience) that could index their built relationship. The user and agent models 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. The 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 utilized the off-the-shelf rule-based sentiment classifier (Gilbert, 2014).

Preference Extractor In the clause breaking process, Stanford CoreNLP pipeline (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.

Socially-Aware NLG

In this section, we adopt a theory-driven template-based approach to generate social moves. Building on Spencer-Oatey's five domains of rapport management strategies (Spencer-Oatey, 2008b,a), (**Zhao, Ran** et al., 2014) proposed a computation model of rapport that explains how humans in dyadic interactions build rapport over time through conversational strategies. Specifically, (**Zhao, Ran** et al., 2014) 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 Norms (**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 dialogues. Thus, most conversational strategies proposed in (**Zhao, Ran** et al., 2014) 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. Specifically, 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

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 4.4: Strategy realizations

Category	Pre-conditions
System and User Speech Act	Request, Reject, Greeting, Closing, Gratitude
Sentiment	positive and negative
Dialogue 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 4.5: Preconditions

variations in sentence realization for each sub-category by hiring two native English writers from the Fiverr website. Table 4.4 shows some examples.

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

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 4.5.

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, Ran** et al., 2016a), which can significantly enhance interpersonal rapport between friends (Ogan et al., 2012). 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.

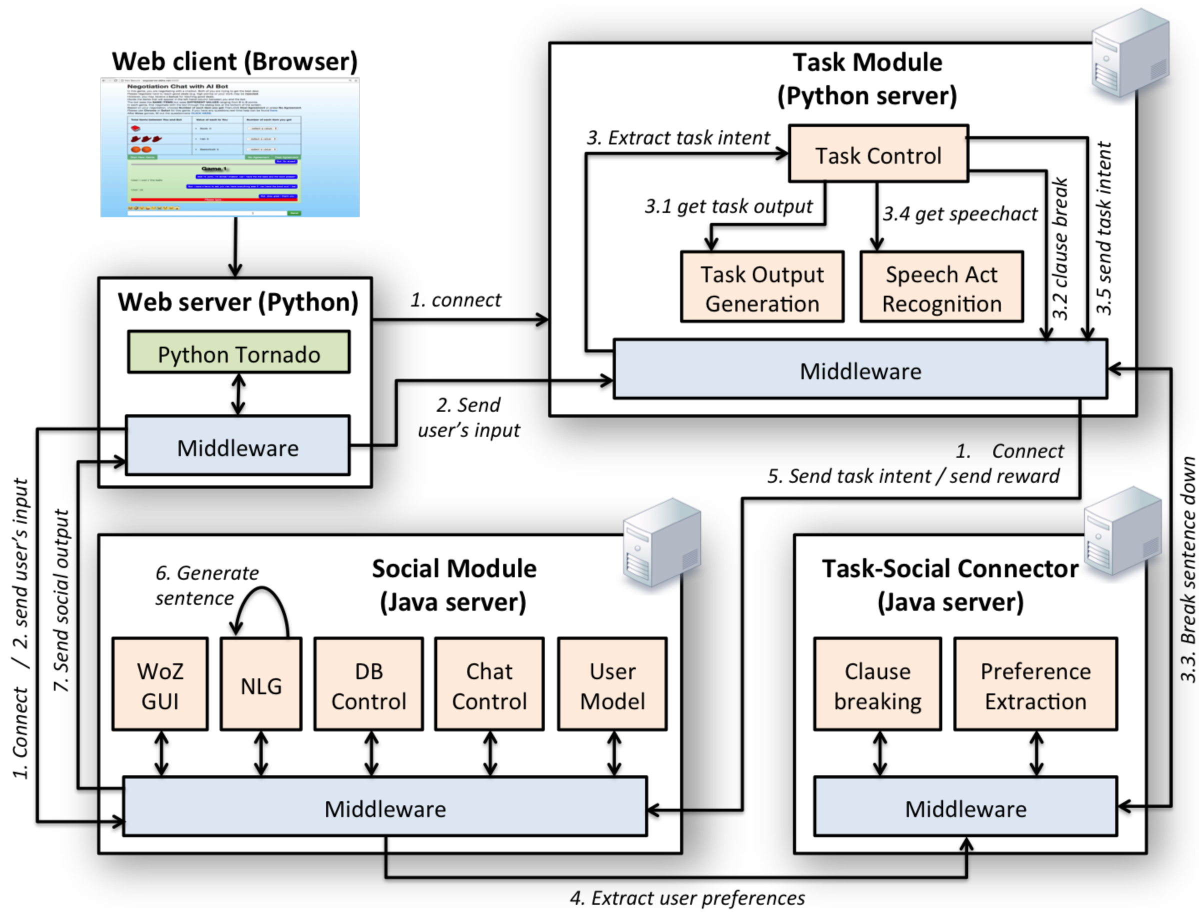

Figure 4.3: Overall architecture of SOGO 1.0

4.4.2 System Architecture

Our SOGO 1.0 system is in the Wizard-of-Oz (WoZ) setup. Below we detail the architecture for our computational model, shown in Figure 4.3.

SOGO 1.0 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 generates 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. 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.

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 launches 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 4.3). 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 extracts user preferences in order to update the user model (step 4) and displays user and system interaction in real-time on a 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 dialogue state, a set of plausible social sentences are shown to the WoZ-er, 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 the user's browser (step 7).

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 to extract json messages that were embedded on those logs for further evaluation of the data collected from the experiments.

4.4.3 Pilot Study with SOGO 1.0 System

In the Wizard-of-Oz setup, the functions described above are carried out automatically, but the WoZ-er decides which strategy to use when there are multiple available. This pilot study serves as the proof-of-concept for our proposed work. In our experiment, we use fully-automated Facebook end-to-end dialogue model as a baseline and compare it with our developed SOGO 1.0 system.

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 1.0 on rapport building with a human user and its performance in the context of negotiation.

4.4.4 Evaluation

In this study, we combine subjective and objective measures to assess the effectiveness of our social intelligent negotiation dialogue 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-reporting 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. 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 4.6 shows the list of questions and 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 1.0 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 observed that users felt slightly frustrated in both setups but showed no significant differences toward the two systems.

Attentiveness: Users reported that the SOGO 1.0 system paid more attention to them (d=0.89), was more respectful and better attended their concerns (d=0.80). These findings also have a high effective size. Users stated that they were interested in listening to the system in both conditions. Thus, we found no significant difference among the two groups of participants, even though the mean values of this question were low (Mean(SOGO)=1.63, Mean(FB)=1.70).

Positivity: Three questions in this dimension showed significantly different responses between the groups. Users liked the SOGO 1.0 system more and felt warmer toward their partner (d=0.76). They experienced a greater sense of friendliness (d=1.00) and caring from the SOGO 1.0 system, as well (d=0.87).

Face: Both groups reported a low degree of damage to 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 1.0 system (d=-0.35), though the rating of the SOGO 1.0 system is still unsatisfied with the Mean(SOGO)=2.23. Users felt more satisfied

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 4.6: Complete t-test statistical analysis of subjective questionnaire of rapport assessment by comparing SOGO 1.0 system and Facebook end-to-end system. Effect size assessed via Cohen's d. Significance: ***:p <0.001, **:p <0.01, *:p <0.05

about the instrumental outcome in the SOGO 1.0 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 1.0 system. The attitude to these questions differ significantly in the two groups with a large effective size (d=0.99).

Perceived Rapport: User perceived significantly higher rapport with SOGO 1.0 system (Mean(SOGO)=5.10) as compared to FB system (Mean(SOGO)=3.50).

Information Disclosure: Users preferred to share more of their personal information with the Facebook system rather than the SOGO 1.0 system but the results suffer from low effective size (d=-0.2).

Objective metrics

We first measured agent performance through three dimensions inherited from (Lewis et al., 2017): (1) Number of wins by the system (**Number of Win**). Obviously, the SOGO 1.0 system wins more often than the Facebook system, with a moderate effective size (d=0.67). (2) Percentage of games

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 Dialogue 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 4.7: Complete t-test statistical analysis of negotiation performance of SOGO 1.0 system versus Facebook Baseline system. Effective size assessed via Cohen's d. Significance: ***:p<0.001, **:p<0.01, *:p<0.05

that end with an agreed-upon negotiation decision (**Deal Rate**). As in (Lewis et al., 2017) show, the major problem with the Facebook system is that users even prefer not to agree rather than capitulate to an uncooperative system. The SOGO 1.0 system significantly improves the 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 1.0 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 dialogue length for both systems are similar, however, users reply with more words in each utterance when they negotiate with the SOGO 1.0 system (d=0.67). Table 4.7 provides complete results.

Factor analysis of rapport

Our goals were to quantify the computational model of rapport provided in our previous work (**Zhao, Ran** et al., 2014) and specify the variable loadings of each question to its corresponding factor of rapport. With respect to these goals, we conducted a confirmatory factor analysisa special form of structural equation model (SEM)which is most commonly used to test whether the data fit a hypothesized measurement model by assessing 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 4.8. In the top-level of the model, coordination and positivity correlate strongly to rapport in the context of negotiation. This 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),

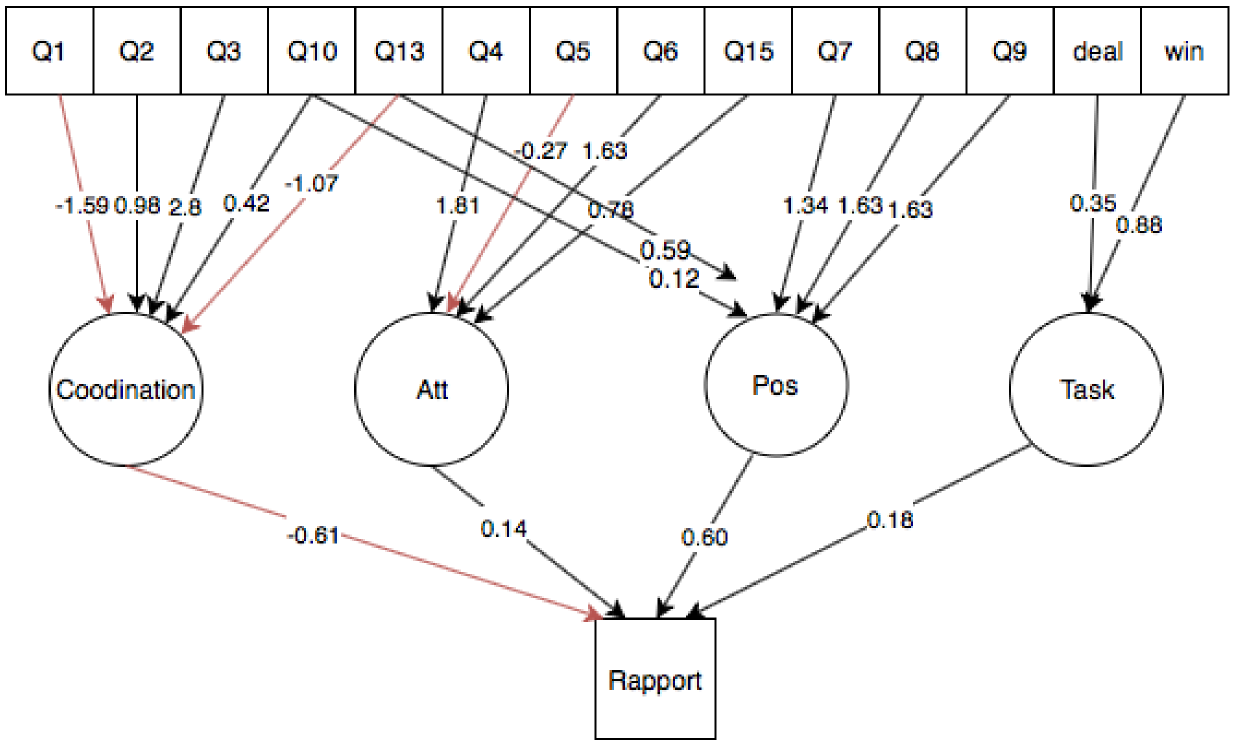

Figure 4.4: Structural equation model of rapport

Name	Value
RMSEA	0.094
SRMR	0.063
CFI	0.937
TLI	0.915

Table 4.8: Model fit metrics. RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker-Lewis index

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 contradicts rapport theory (Tickle-Degnen and Rosenthal, 1990) perhaps because attentiveness has a large covariance with coordination and positivity. Thus, its explanatory power toward the variance of rapport is reduced. In the bottom level, most questions have high variable loadings. Figure 4.4 displays the complete results.

4.4.5 Discussion

In this section, we demonstrate the proof-of-concept social intelligent negotiation dialogue 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 dialogue model to decide the next task move and a theory-driven template-based social language generator to introduce social skills into the system. The experiment demonstrates that our SOGO 1.0 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 1.0 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.

4.5 SOGO 2.0: A Fully-automated Socially-aware Negotiation System

In the previous section, we presented a semi-automated dialogue system (SOGO 1.0) and introduced rapport as the social outcome in human-agent negotiation. We demonstrated that different linguistic devices effectively build rapport and foster integrative agreement with human users. While natural, SOGO 1.0 system requires a human expert to decide the final sentence per turn, relying considerably on human effort and expertise and thus hindering its scalability and transferability. In this section, we present a novel framework for developing a fully-automated socially-aware negotiation system (SOGO 2.0) that leverages expert supervision to blend task and social goals in negotiation (see Table 4.9). Within the framework, two computational models are introduced: a rapport estimator and a socially-aware dialogue model. The rapport estimator learns to assess dialogue with respect to interpersonal closeness. The socially-aware dialogue model initially imitates the actions of human users to improve itself against task and social goals respectively.

Several studies have attempted to model negotiation with a fully-automated system. In the last section, we introduce FAIR's model that learns both linguistic and reasoning skills (without annotation) via reinforcement learning. However, entangling linguistic form and meaning is problematic for this paradigm since backpropagation caused by strategy decisions can adversely affect the linguistic quality of text, such as lack of semantic and linguistic diversity. To address this issue, FAIR introduced an approach to learn latent message representation that decouples semantics from surface text. Then, with the latent message representation, they leveraged a hierarchical reinforcement learning model for strategic planning and language generation. Similarly, (Zhao, Tiancheng et al., 2019) solved the problem with their latent action framework by treating dialogue acts as latent variables and developing unsupervised methods to induce action space from the data. Both works, however, framed negotiation as a purely non-cooperative task designed to maximized the task outcome. Conversely, (He et al., 2018) considered the perspective of human-likeness on modeling negotiation agents and placed "fairness" and "dialogue length" into reward functions. They adopted a modular approach centered around coarse dialogue acts (e.g., propose<book=1, hat=0, ball=2>) that decoupled task strategy from language

Human-Human

[Items available: 2 books, 3 balls, 2 hats]

User A: *I would like just 2 books and 1 hat.*

User B: *Can I have 1 hat and 3 balls?*

User A: *Actually I would like 2 books and 2 hats.*

User B: *No, sorry, I don't mind the books but i need 1 hat.*

User A: *Ok, I'll keep 2 books and 1 hat u have the rest*

User B: *Deal!*

Human-SOGO(2.0)

[Item available: 2 books, 2 balls, 1 hat]

Agent: *It's always a pleasure to play with you* 🙂*. Is that ok if I give you 2 balls?*

User: *I need the hat and 2 balls.*

Agent: *The hat looks great on me my friend* 🙂*. So would you mind if I take 1 ball and the hat ?*

User: *Seriously? I need the hat and you can have the rest.*

Agent: *Ok, remember to pay me back next time* 😛*. Playing with you is making me a smarter chatbot.*

Table 4.9: Sample human-human (top) and human-SOGO(2.0) (bottom) negotiation dialogues

generation (e.g., *"I'd like to keep a book and 2 balls, you can keep the rest"*). Yet, their approach is limited in three ways: (1) Though dialogue length and fairness are indicators of user engagement, no interpersonal goal (e.g., rapport) is explicitly defined for the agent; (2) The rule-based parser is not robust enough to capture different modality and negation forms (e.g., *"I can make a deal"* and *"I can't make a deal"* produce the same dialogue act); (3) They lose linguistic features (e.g., sentiment) of an utterance when parsing it into a coarse dialogue act form.

Based on our observation, building a fully-automated system is non-trivial due to several practical challenges. For instance, a dialogue model for a fully-automated system should be designed to simultaneously optimize task and social goals in conversation. However, training this kind of dialogue model requires significant human-agent data, which is generally unavailable or expensive to collect. Additionally, to further advance the social capability of the model, an accurate and prompt human-derived reward is required at the end of each interaction, which is impractical.

To alleviate this problem, we propose a computational framework to build a socially-aware dialogue model that learns to blend task and social goals. The key idea is to pre-train the dialogue model with explicit human supervision and fine-tune against derived reward through self-play. Task reward is easy to compute, and social reward can be inferred from an end-to-end model within the framework. Our dialogue model instantiates a task goal with different negotiation tactics in the format of coarse dialogue acts. To fulfill interpersonal goals, we leverage prior work on conversational strategy (**Zhao,Ran** et al., 2018), a kind of task-related social move that prepares the interlocutor for the next negotiation move (i.e., Self-Disclosure (SD) that reveals personal information, Reference to Shared Experiences (RSE) that indexes common history, Praise (PR) that serves to increase self-esteem in the listener, and Violation of Social Norm (VSN) where general norms are purposely violated to accommodate the others behavioral expectations). We evaluated the performance of our system against two baseline systems. The result demonstrates that our system performs better on task and social outcomes in subjective and objective evaluation metrics. The results suggest that our framework can be extended beyond negotiation tasks to other domains.

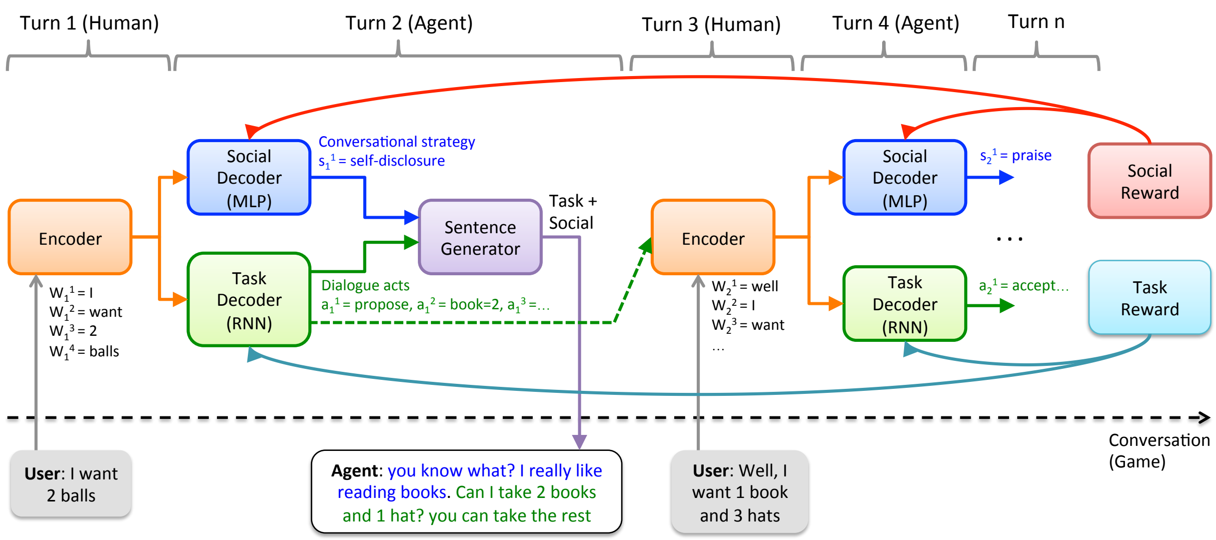

Figure 4.5: Illustration of the framework for socially-aware dialogue model

4.5.1 Proposed Framework

Figure 4.5 illustrates the proposed framework, which includes a rapport estimator and a socially-aware dialogue model. The rapport estimator takes a sequence of tokens as input, including the agent's coarse dialogue act along with its own realized conversational strategy and the user's surface-level words. Then it assesses rapport to serve as the social reward for an interaction. The socially-aware dialogue model performs two phases (the task phase and social phase) in parallel. It incorporates a dialogue encoder, which combines a social decoder and a task decoder that learn the dialogue context. In the task phase, our task decoder decides the next high-level coarse dialogue act based on its negotiation policy that has been trained via maximizing the derived task reward. In the social phase, social decoder deciphers an appropriate conversational turn by maximizing the social reward that is computed by our rapport estimator.

4.5.2 Rapport Estimator

Before discussing the negotiation dialogue model, it is worth detailing how the rapport estimator computes the social reward by determining whether a sequence of words increases or decreases rapport. Collecting significant human-agent negotiation dialogues is expensive and time consuming. To tackle these challenges, we generated a synthetic human-agent negotiation corpus by bringing together (in a self-play setting) an agent and a user simulator grounded in our previous Wizard-of-Oz study (Section 4.4). We then trained the rapport estimator using a corpus annotated by Amazon Mechanical Turk. Both agent and user simulators reason next task strategy from an end-to-end negotiation model (Lewis et al., 2017). Below we describe how we modeled the social strategy decision-making process.

User Simulator: Most approaches for modeling human behavior in user simulators are model-based and rule-based (Li et al., 2016b; Cuayáhuitl et al., 2005; Eckert et al., 1997). However, they do not properly capture the complexity of whole discourse. According to our observations, conversational strategies used by humans when negotiating with an agent are sparse (see Table4.12). This dearth of representative training examples (only 7.8% of data points use conversational strategies) further complicates attempts to model through classic AI or machine learning techniques. For this reason, our user simulator relies on cognitive theory rather than purely AI data-driven approaches. In particular, we extend the work proposed by (Romero et al., 2017), where a social reasoner module mimics a human expert making decisions about the correct conversational strategy per dialogue turn. This model is based on a Behavior Network (BN) (Maes, 1989), a hybrid approach that takes advantage of the features of AI planners governed by spreading activation dynamics. Based on the Global Workspace theory of consciousness and human attention (Baars, 2003), the social reasoner is endowed with both short 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. A BN comprises a set of behaviors that compete and collaborate between them for getting the focus of attention via excitatory and inhibitory links that connect behaviors' pre/post-conditions. Each behavior in Behavior network (BN) corresponds to a specific conversational strategy, represented as a tupple `<pre, pos, neg, act>`, where `pre` is a set of pre-conditions that must be true to activate the behavior; `pos` and `neg` are the positive (expected) and negative (unexpected) behavior's post-conditions; and `act` is the behavior's activation. Behaviors are interconnected through excitatory and inhibitory links, so behavior A will inhibit behavior B if there is a premise p such that $(p \in A.neg) \wedge (p \in B.pre)$, and A will excite B if $(p \in A.pos) \wedge (p \in B.pre)$.

Category	Pre- and Post-conditions
# of Games	low, medium, high
# of Turns	low, medium, high
Sentiment	negative, positive, neutral
User/agent intent	request, reject, et al.
Agent/User CS	praise, self-disclosure, et al.
Game performance	win, lost, equal

Table 4.10: Pre- and Post-condition Categories

In our study, we annotated 2,000 human-agent utterances and found that only 144 of them used conversational strategies. Using the annotated dialogues, we identified the pre- and post-conditions of each behavior (conversational strategy) – See Table 4.10. Then, using a semi-automatic 5-fold cross-validation process, we fine-tuned the model and proved that the parameterized learning (pre- and post-conditions, global parameters, etc.) generalize conveniently across the population samples, demonstrating that BN can model conversational strategies in different data settings while achieving 79% F1 score.

Conversational Strategy	Precision	Recall	F1
SD	0.89	0.81	0.85
RSE	0.72	0.83	0.77
PR	0.77	0.73	0.75
VSN	0.85	0.76	0.80
Request	0.81	0.85	0.83
Reject	0.92	0.89	0.90
Gratitude	0.81	0.85	0.83
Greeting	0.96	0.98	0.97
Closing	0.93	0.98	0.95

Table 4.11: Results of 10-fold cross validation on predicting conversational strategies (Agent Simulator)

Agent Simulator: The agent simulator is designed to imitate human decision-making in the previous Wizard-of-Oz study (Section 4.4) in selecting conversational strategy. We applied a L2-logistic regression model to construct a binary classifier for each conversational strategy individually. The prediction is decided by comparing their confidence levels. Performance is evaluated via 10-fold cross validation, as shown in Table 4.11. The complete set of hand-crafted features includes:

- Agent current turn speech act
- Agent last turn conversational strategy
- User last turn speech act
- User last three turns sentiment (average)
- Number of turn
- Number of game
- Winner of last game
- Deal or no deal last game

Rapport Annotation: Our synthetic human-agent dialogue corpus contains 2,200 dialogues that were generated based on scenarios from (Lewis et al., 2017). We leverage objective metrics introduced in (Schatzmann et al., 2005; Lewis et al., 2017) to evaluate the quality of our synthetic corpus in Table 4.12. Overall, the results demonstrate that our simulated dialogues are good proxies for human-agent dialogue. Then, Amazon Mechanical Turk assessed rapport using a 7-point Likert scale. Native raters were provided with a simple definition of rapport and three raters annotated each dialogue. Inter-rater reliability, computed via Krippendorff's alpha, is 0.79. Weighted majority rule was deployed to mitigate bias from the ratings of different annotators and account for label over-use and under-use. To evaluate the quality of third-party rapport annotation, we asked the Turkers to annotate the human-agent dialogues collected in the Section 4.4.4 and compared them with self-report rapport. Since rapport is a subjective variable, we allowed a tolerance in measurement of +/-1 in our case. The results in Figure 4.6 prove that third-party rapport annotation is analogous to self-reported rapport.

Metrics	Hu-Hu	Hu-WoZ	Sim
System win ratio	N/A	0.5	0.5
Deal Rate	0.8	0.9	0.8
Dialogue Length	6.6	6.6	6.4
User Utterance length	7.6	7.2	6.8
Pareto Optimal	76.9	96.7	89.3
Human CS ratio	6.9%	7.8%	9.2%

Table 4.12: Evaluation results on Human-Human (Hu-Hu), Human-Agent (Hu-Agent) and Simulated (Sim) dialogues.

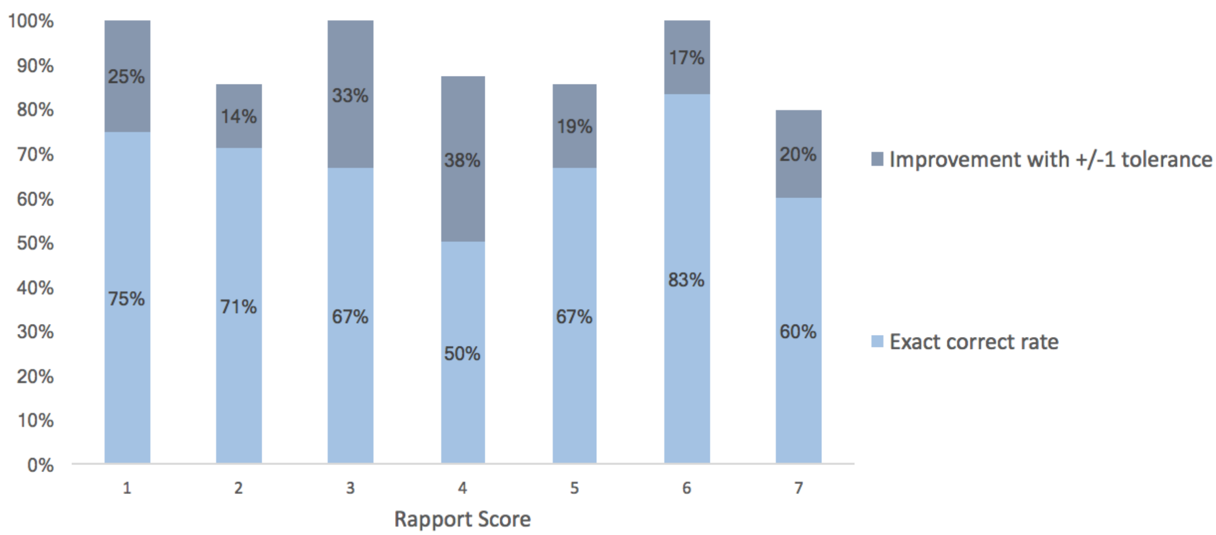

Figure 4.6: Results of agreement between **third-party** and **self-report** rapport

Bidirectional LSTM Model of rapport estimator: Each dialogue is composed of human turns and agent turns. Agent turns are represented by the coarse dialogue act and the chosen conversational strategy. Each coarse dialogue act is represented as a sequence of tokens a_i (He et al., 2018), e.g., "propose, 2 books, 1 ball, 2 hats." The set of conversational strategies s_i is defined in (**Zhao,Ran** et al., 2018). Each human turn is represented by a sequence of words $[w_1, w_2, ...w_n]$. Figure 4.7 illustrates how inputs are blended into the rapport estimator. We created a unified vocabulary $X = w_i \cup a_i \cup s_i$ and converted each dialogue as a sequence $[x_1, x_2, x_3, ...x_n]$, which was then assigned a class of rapport. We deployed a long-short-term memory recurrent neural network (LSTM) to overcome gradient vanish. To further boost the power of memorizing long sequences of data like a dialogue, we used a bidirectional LSTM, which processes a sentence in forward and reverse directions to exploit information both from the past and the future. Our rapport estimator automatically encodes the high level abstract

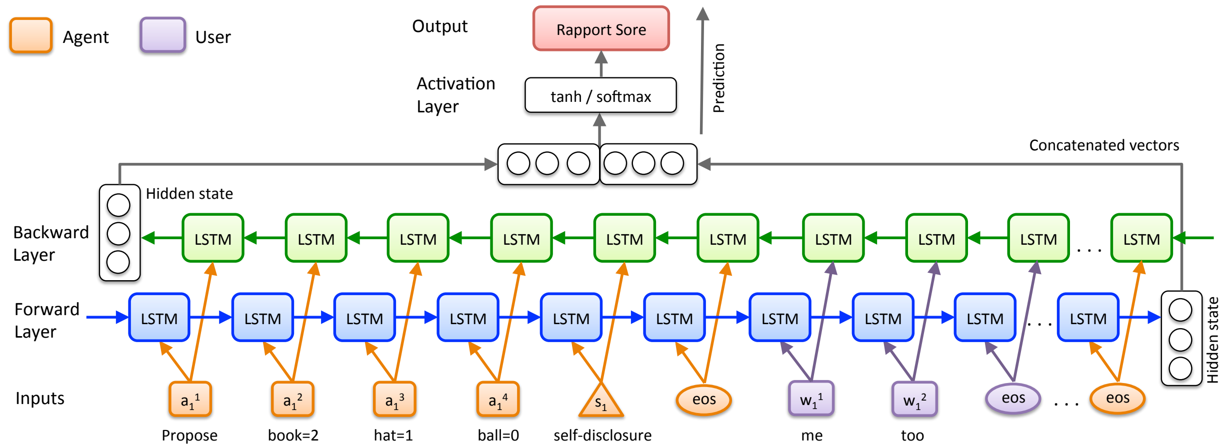

Figure 4.7: Bidirectional LSTM model for rapport estimation. $a_1^1 ... a_1^a n$: dialogue acts, s_1: conversational strategy, eos: end of sentence, $w_1^1 ... w_1^m$: user sentence's words.

information of the whole dialogue with a bidirectional LSTM to produce a dense representation. Specifically, the mechanism of LSTM is defined as:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ j_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W[h_{t-1}, x_t]$$
$$c_t = f_t \odot c_{t-1} + i_t \odot j_t$$
$$h_t = o_t \odot tanh(c_t)$$

The forward layer will generate h_{if} as the last hidden state and backward layer will end up with h_{ib} as the last hidden state. Then we concatenate these two vectors and utilize a softmax classifier to predict label $\hat{y}$ from a discrete set of class Y for a dialogue D.

$$u = tanh(W_u[h_{if}, h_{ib}] + b_u)$$
$$\hat{y} = softmax(W_s u + b_s)$$

Figure 4.7 shows the unfolded bidirectional LSTM structure for rapport classification. The network objective is trained to minimize the cross-entropy in the predicted distributions of rapport.

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j log(\hat{y}_i^j) \tag{4.6}$$

where N denotes the number of training samples and C is the rapport class number. Performance is evaluated via 10-fold cross validation, as shown in Figure 4.8. The result demonstrates that our rapport estimator could assess negotiation rapport in nearly human-level performance. Therefore, we are confident in employing the rapport estimator to induce social reward for a dialogue.

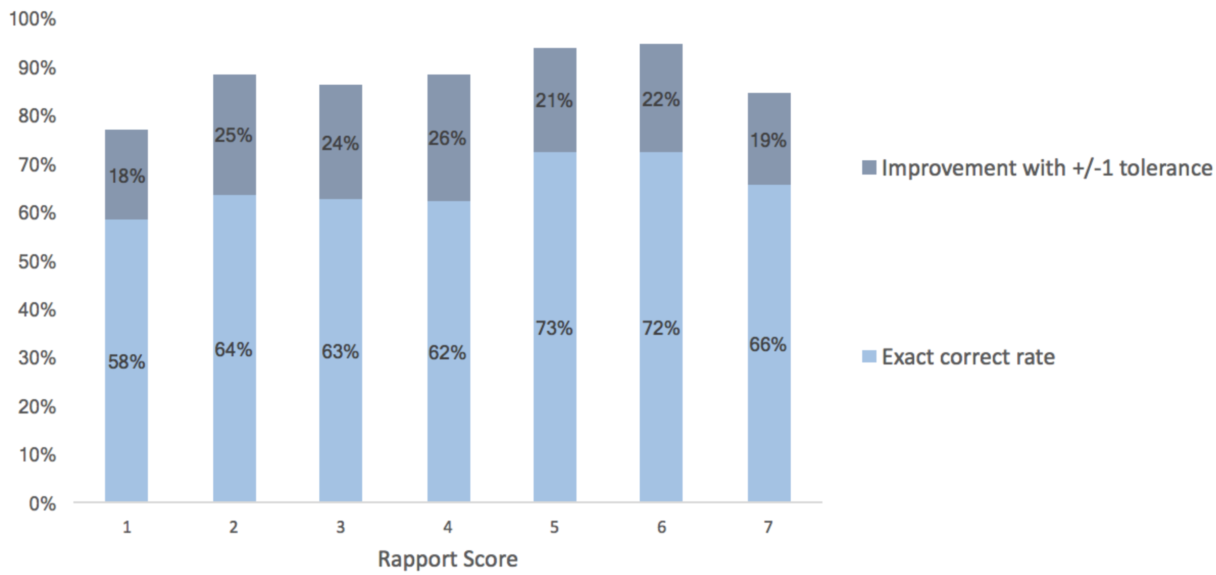

Figure 4.8: Results of agreement between **human** and **rapport estimator**

4.5.3 Socially-aware Dialogue Model

Drawing on the work of (He et al., 2018), we model negotiation dialogue in the act space that decouples strategy and language generation. Compared to common sequence-to-sequence models, this approach highlights three advantages in our work. First, we separately model coarse dialogue acts (e.g., propose(book=2,hat=1,ball=0) for task goals and conversational strategy (e.g., self-disclosure) for social goals, which enables us to interpret and incorporate model outputs into sentence generation. Second, since act space is much smaller than word-level space, most model capacity will be consumed by negotiation tactics instead of complex language models. Finally, we optimize both task and social goals by fine-tuning acts, which has been proven to be more efficient than words (Zhao, Tiancheng et al., 2019). Figure 4.5 depicts the network structure of our socially-aware dialogue model. We followed the two-step learning mechanism by pre-training the model with supervised learning, then tweaking parameters using policy gradient reinforcement learning (Lewis et al., 2017; Williams et al., 2017).

Supervised Learning

In this step, our goal is to imitate human behavior by maximizing the log likelihood on the training dialogues. Specifically, the encoder network converts human utterances into fixed-length vectors to serve as the dialogue context c. The task decoder network's targets generate coarse dialogue acts (what to say) by modeling the conditional distribution $p_\theta(a_t|a_{<t}, c)$ (parameterized by θ). We use a standard RNN-based decoder and an additive attention mechanism as explained in (Bahdanau et al., 2014). The attention will be applied to the hidden states from encoder network to generate context representation vector c. The weighted arithmetic mean of the $\{h_i\}$ is calculated according to the relevance of each h_i to the next negotiation strategy, thus

$$
\begin{aligned}
u_t =& tanh(w_w h_t + b_w)\\
\alpha_t =& \frac{exp(u_t^T u_w)}{\sum_t exp(u_t^T u_w)}\\
c =& \sum_t \alpha_t h_t
\end{aligned}
$$

Given that the dialogue context c captures the meaning of the input sequence, our RNN-based decoder produces target token sequences as the coarse dialogue acts word-by-word. We slightly modify the LSTM formula that we defined in Section 4.5.2 by adding the context vector c_t:

$$
\begin{aligned}
h_t &= LSTM(h_{t-1}, [w_{i_{t-1}}, c_t])\\
s_t &= FC(h_t)\\
p_t &= softmax(s_t)\\
i_t &= argmax(p_t)
\end{aligned}
$$

In parallel, the social decoder network produces conversational strategy (how to say something) where we use a multilayer perceptron to learn the conditional probability $p_\phi(s|c)$.

Reinforcement Learning

After pre-training our encoder-decoder model with human behavior, we aim to explicitly maximize agent task and social rewards in negotiation.

- **Task reward** (r_{task}): the number of points agent received in the game.
- **Social reward** (r_{soc}): rapport score perceived by our automatic estimator.

In terms of optimization, we chose a policy-gradient approach, in which a model generates a distribution from which actions (coarse dialogue act and conversational strategy) are sampled at each time step. At the end of each dialogue, the task reward is automatically calculated and social reward is inferred by our developed rapport estimator described in Section 4.5.2. The gradients of the probabilities of the actions taken with respect to the model weights are computed. Intuitively, given the dialogue context from the encoder output, an appropriate dialogue act relative to the task reward and an appropriate conversational strategy relative to the social reward will receive a positive gradient step and become more prevalent. Specifically, we utilize Monte-Carlo policy gradient (REINFORCE) (Williams, 1992) to update the model parameters against our objectives. Let us assume the model has access to a social reward signal r_t^{soc} and a task reward signal r_t^{task} throughout the dialogue. Then we can write the expected discounted task reward as $J(\theta) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t^{task}]$ and expected discounted social reward as $J(\phi) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t^{soc}]$. T is the dialogue length and γ is the discount rate. To reduce the variance of the updates(Greensmith et al., 2004), we also apply a running average of completed dialogue rewards in a batch μ. We update the parameters ϕ and θ by:

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta[\sum_{t=0}^{T}\sum_{i=0}^{Z_t} \nabla_\theta logp_\theta(a_{ti}|a_{<ti}, c_t)\sum_{k=0}^{T}\gamma^k(r_k^{task} - \mu_{task})]$$
$$\nabla_\phi J(\phi) = \mathbb{E}_\phi[\sum_{t=0}^{T} \nabla_\phi logp_\phi(s_t|s_{<t}, c_t)\sum_{k=0}^{T}\gamma^k(r_k^{soc} - \mu_{soc})]$$

where Z_t is the number of tokens in the coarse dialogue act at turn t.

4.5.4 Experiments

Training Details

We trained our socially-aware negotiation model in four stages. The original DealOrNoDeal dataset was used in the first two stages and the synthetic dataset for the last two stages. We hold out 236 scenarios for testing.

1. SL-Task: Pre-train encoder and task decoder with teacher forcing method (Williams and Zipser, 1989).
2. RL-Task: Fine-tune encoder and task decoder with a computed task reward.
3. SL-Soc: Freeze the parameter of task decoder, train encoder and social decoder with expert supervision on conversational strategy. (see Table 4.13 for performance of SL-Soc stage model)
4. RL-Soc: Fine-tune encoder and social decoder with social reward (rapport) computed by rapport estimator (see Figure 4.9).

Conversational Strategy	Precision	Recall	F1
SD	0.77	0.85	0.80
RSE	0.77	0.76	0.77
PR	0.75	0.79	0.77
VSN	0.73	0.76	0.75
Request	0.84	0.81	0.83
Reject	0.95	0.77	0.85
Gratitude	0.83	0.84	0.84
Greeting	0.98	0.99	0.99
Closing	0.95	0.99	0.97

Table 4.13: Performance of SL-Soc stage model on predicting different conversational strategies

Both the encoder and task decoder are two-layer LSTM with 200 hidden units. The social decoder is implemented as a multilayer perceptron which has a hidden layer of 128 with $tanh$ activation function. The drop rate for both $tanh$ and LSTM layer outputs is 0.3. Parameters are initialized sampling from a uniform distribution between -0.1 and 0.1. For optimization, we chose AdaGrad (Duchi et al., 2011). In supervised learning stages, the learning rate is 0.01. The mini-batch size is 64. We trained the model for 30 epochs and the best model with the lowest validation loss was forwarded to the next stage. In the reinforcement learning stage, we changed our learning rate to 0.001. The partner model is fixed and is derived from the SL-Task stage. The other agent is updated by policy gradient every episode (dialogue) with a discounting factor 0.95. The network is evaluated every 20 episodes. The max length of each episode is 20. The model was trained up to 5000 episodes. Eventually, the model with the highest reward on the validation set is chosen.

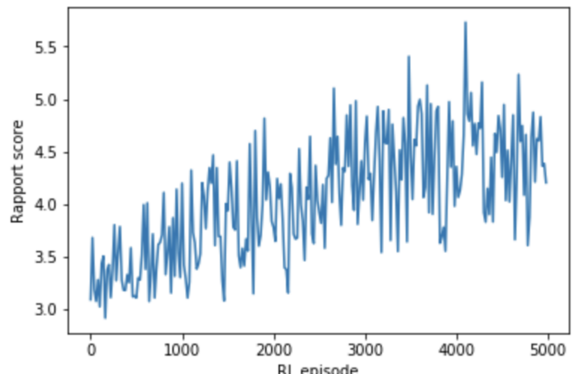

Figure 4.9: Social reward (rapport) learning curve over the course of RL training

Task Setup

Our goal is to evaluate performance on blending task and social goals in negotiation. Specifically, we conducted two-round evaluation. In Evaluation I, we introduced two baseline systems: (1) a **Task agent** that carries out only a task goal (He et al., 2018) and (2) an **SL-rapport agent** (the product after SL-Soc training stage) designed to imitate expert decision-making in the previous Wizard-of-Oz study (Section 4.4) on selecting conversational strategy. In Evaluation II, we investigate if rapport in negotiation is achieved through strategic discourse planning or just simply being polite or teasing each other like friends. Toward this end, we proposed two stronger baseline systems: (1) a **Politeness agent** to only apply the praise strategy (2) a **Rudeness agent** to only apply the violation social norm/teasing strategy. In both evaluations, we recruited 45 English speakers on Amazon Mechanical Turk who were equally and randomly assigned to a system. To obtain high quality data, those workers were based in the US or UK and had at least 95% approval rating with more than 5,000 previous Human Intelligence Tasks (HITs). Each participant played five games with the agent and completed a subjective questionnaire to reveal their feelings toward the game and interlocutor.

4.5.5 Evaluation I

In this study, we operationalize objective measures to quantify the task goal and subjective measures to quantify the social goal in negotiation. For a fair comparison, we followed the same metrics as (**Zhao,Ran** et al., 2018). Since our goal was to explore the mean value differences of users ratings towards three systems, we ran a one-way ANOVA with Tukey's post hoc test

Dimension		$\mu_{\mathbf{Task}}$	vs. $\mu_{\mathbf{SL-rap}}$	vs. $\mu_{\mathbf{RL-rap}}$	RL-rap is better than SL-rap
Coordination	Q1	3.1	**4.5**	**5.1**	
	Q2	2.5	2.4	**1.7**	✓
	Q3	3.7	3.1	**2.2**	✓
Attentiveness	Q4	4.1	4.9	**5.5**	
	Q5	1.7	2.5	2.1	
	Q6	3.5	3.6	**5.4**	✓
Positivity	Q7	4.8	**5.9**	**6.2**	
	Q8	4.2	5.3	5.5	
	Q9	2.9	3.5	**4.6**	✓
Face	Q10	2.3	2.6	2.1	
Feeling about the negotiation	Q11	3.8	**2.7**	**2.2**	✓
	Q12	4.7	4.5	5.3	✓
	Q13	4.1	**5.1**	**5.5**	
Perceived Rapport	Q14	3.9	**4.8**	**5.2**	✓
Information Disclosure	Q15	4.9	4.1	4.8	

Table 4.14: Human subjective evaluation results on different dimensions of rapport by comparing the Task agent (Task), the SL-rapport agent (SL-rap) and the RL-rapport agent (RL-rap) through one-way ANOVA with Tukey's post hoc test. In Tukey's post hoc test, the SL-rapport agent and RL-rapport agent are separately compared with the Task agent. Scores with statistical significance are **bold**. Performance of the SL-rapport agent is also compared with RL-rapport agent. Scores with statistical significance are highlighted with ✓.

for pairwise comparison. The level of significance was 0.05. In Tukey's post hoc test, the SL-rapport agent and RL-rapport agent separately compare with the task agent. Scores with statistical significance are bold. We also compared the performance between the SL-rapport agent and RL-rapport agent. Scores with statistical significance are highlighted with ✓. For all significant results, we calculated effect size via Cohen's d.

Subjective Metrics

After playing five consecutive games, each participant assessed the negotiation by completing a 15-item self-reporting questionnaire that characterizes the interaction into dimensions of rapport discovered in previous studies (Curhan et al., 2006; Gratch et al., 2015; DeVault et al., 2015). Responses were rated on a scale of 1 (Strongly Disagree) to 7 (Strongly Agree). Factor analysis proved single factor for the 15 questionnaire items, which have high internal consistency with Cronbach's α= 0.94. Table 4.14 shows the complete list of questions and results. We describe our findings on each dimension of rapport as follows:
Coordination: Users felt more synchronicity with the SL-rapport agent (d=1.05) and RL-rapport agent (d=1.41) as compared to the Task agent. The RL-rapport agent was shown to be more advanced than the SL-rapport agent by reducing feelings of frustration (d=-0.67) and uncomfortableness (d=-0.69) during the interaction.

Attentiveness: Compared to the Task agent, the RL-rapport agent paid more attention to users (d=0.82) and behaved respectfully in consideration of their concerns (d=1.37). The SL-rapport agent did not show significant improvement. Users claimed that they were interested in listening to all three systems without obvious differences.
Positivity: Users felt warmest toward the RL-rapport agent (d_{task}=0.74, d_{SL-rap}=0.23). They experienced a greater sense of friendliness (d_{task}=0.91, d_{SL-rap}=0.16) and caring from RL-rapport agent (d_{task}=1.31, d_{SL-rap}=0.65) as well.
Face: Neither group reported damage to their sense of pride, nor did significant differences surface across groups.
Feeling about the negotiation: Overall, the RL-rapport agent offered the most impressive experiences about the goods and relationship exchange in negotiation. In particular, it ameliorated uncompromising and uncooperative characteristics (d_{task}=0.54, d_{SL-rap}=1.21) though could still be improved (μ_{RL-rap}=2.2). Users felt more satisfied about the instrumental outcome (d_{task}=0.43, d_{SL-rap}=0.65) and believed the whole negotiation process was a good foundation for further relationship development (d_{task}=0.89, d_{SL-rap}=0.52).
Perceived Rapport: Users perceived significantly higher rapport in the SL-rapport agent (d=0.57) and the RL-rapport agent (d=1.13) compared to the Task agent, likely because they explicitly maximized social goals in conversation. Concurrently, the RL-rapport agent demonstrated greater capacity to build rapport than the SL-rapport agent (d=0.69).
Information Disclosure: Users suggested that they would not mind sharing their personal information with all three systems. We did not find significant differences on users attitudes across groups.

Objective Metrics	$\mu_{\mathbf{Task}}$	$\mu_{\mathbf{SL-rap}}$	$\mu_{\mathbf{RL-rap}}$	RL-rap is better than SL-rap
Win Times	1.6	2.3	**2.7**	
Deal Rate	0.4	**0.8**	**0.9**	✓
Pareto Optimal	72.2	**86.9**	**84.9**	

Table 4.15: Objective evaluation results on different dimensions of rapport by comparing the Task agent (Task), the SL-rapport agent (SL-rap) and the RL-rapport agent (RL-rap). The SL-rapport agent and the RL-rapport agent separately conducted pairwise t-tests with the Task agent. Scores with statistical significance ($p<0.05$) are **bold**. The pairwise t-test between the SL-rapport agent and the RL-rapport agent is conducted and the significant result is indexed with ✓.

Objective metrics

We quantified the task goal of negotiation with three dimensions (see Table 4.15): (1) Number of wins by the system (**Win Times**). Comparatively, the RL-rapport agent wins significantly more times than the Task agent with high effective size d=0.88. (2) Percentage of games that end up with an agreed-upon negotiation decision (**Deal Rate**). As (Lewis et al., 2017; He et al., 2018) describe, a major problem in negotiation in that users are not willing to compromise with a stubborn agent who insists on its offer with even the same words. Compared to a purely task-focused agent, both the RL-rapport agent and the SL-rapport agent have doubled the agreement rate with high effective size (d=2.17, d=1.35), perhaps because established rapport increases compliance (Pederson, 2013). (3) Percentage of Pareto optimal solutions for agreed deals (**Pareto**

Optimal). High scores indicate efficient negotiation. The result delineates that users are more likely to collaborate with the RL-rapport agent (d=0.88) and the SL-rapport agent (d=1.04) to optimize negotiation.

Dimension	Subjective Questions	$\mu_{\mathrm{RL-rap}}$	μ_{Pol}	μ_{Rud}
Coordination	1. I think that my agent and I were in sync with each other.	5.1	**3.7**	**2.8**
	2. I felt uncomfortable and could not say everything that I wanted to say.	1.7	2.2	**4.5**
	3. The interaction was frustrating.	2.2	2.5	**3.1**
Attentiveness	4. I felt that my agent was paying attention to what I was saying.	5.5	**4.3**	**2.9**
	5. I was not really interested in what my agent was saying.	2.1	**3.1**	**3.4**
	6. My agent was respectful to me and considered to my concerns.	5.4	**4.3**	**2.3**
Positivity	7. My agent was friendly to me.	6.2	6.0	**1.9**
	8. I liked and felt warm toward my partner.	5.5	5.2	**2.1**
	9. My agent cared about me.	4.6	4.3	**2.4**
Face	10. Did you lose face (i.e., damage your sense of pride) in the negotiation?	2.1	2.0	**5.1**
Feeling about the negotiation	11. My agent was very uncooperative.	2.2	2.5	**4.6**
	12. How satisfied are you with the balance between your own outcome and your agent's outcome(s)?	5.3	**4.2**	**3.1**
	13. Did the negotiation build a good foundation for a future relationship with your agent?	5.5	4.9	**2.1**
Perceived Rapport	14. I felt rapport between the agent and myself.	5.2	**3.8**	**2.2**
Information Disclosure	15. I was willing to share information with my agent.	4.8	4.9	**3.1**

Table 4.16: Human subjective evaluation results on different dimensions of rapport by comparing the Politeness agent (Pol), the Rudeness agent (Rud) and the RL-rapport agent (RL-rap). The Politeness agent and the Rudeness agent are separately conducted pairwise t-tests with the RL-rapport agent. Scores with statistical significance ($p<0.05$) are **bold**.

4.5.6 Evaluation II

In evaluation I, we discovered that the agent who meets both task and social goals is more competent in negotiation. This promising result prompts us to further think about how to build rapport. Therefore, in evaluation II, we want to find out if rapport in negotiation is achieved through strategic discourse planning or just simply being polite or teasing each other like friends. To reach this goal, we introduce a Politeness agent that only uses the Praise strategy and a Rudeness agent that only applies the Violation Social Norm strategy. Given our interest is comparing the Politeness agent and Rudeness agent separately against the RL-rapport agent, we conducted pairwise t-tests. Scores with statistical significance are bold in Figure 4.16.

Here are some selected results: (1) The right-most column indicates that the Rudeness agent falls behind on all perspectives compared to the RL-rapport agent and Politeness agent. Generally, the Rudeness agent actually detracts from rapport. (2) Both the Politeness agent and Rudeness agent perform worse in the attentiveness dimension, perhaps because the agents did not vary their conversational strategy according to the dialogue context. (3) Users perceived less rapport in both the Politeness and Rudeness agents, which answers our key questions that building rapport is not a process of just simply being polite or teasing like a friend would.

4.5.7 Conclusions and Future Work

In this section, we presented a novel framework to blend social and task goals in conversational agents. As part of the development and fine-tuning process of our framework, we developed a self-play scenario, where both user and agent simulators played against each other, to generate a synthetic human-agent negotiation dataset. We further leveraged that dataset to build a rapport estimator that automatically recognizes the social reward in dialogue. Next, we leveraged the power of neural networks along with human supervision to develop a rapport estimator that can assess social dynamics in conversation in service of improving the systems ability to support the users social goals while enhancing considerably performance of the users task goals. Using these infrastructures as scaffolding, we constructed a neural socially-aware conversational agent model trained to optimize task and social goals simultaneously, which represents (along with the rapport estimator) the core contributions of our work. Our experiments demonstrate that good rapport leads to better task performance and cooperation, which motivates us to incorporate social strategies into a negotiation system. With the help of the Wizard-of-Oz study, our proposed fully-automated system outperforms strong baseline systems in achieving better task and social goals. Furthermore, we discover that building rapport is a process of strategic discourse planning. The limitation of the current work is that the performance of the user simulator is moderate, which can be explained by two main reasons: (a) since we have small number of users to participate in our experiments, even a few inconsistent behaviors between them will affect our training, and (b) we model very sparse discourse phenomena in negotiation. Some special situations might only appear in the test set which reduces the effectiveness of training in our case. To improve our user simulator, we believe that modeling uncertainty in human decision-making process is a promising approach, which could increase data efficiency on learning human behaviors. Meanwhile, in the future, our framework could be broadly generalized to other task domains (e.g., counseling, tutoring) and model different types of social intelligence. Our findings imply a promising research direction on developing a socially-aware dialogue system. By incorporating social goals with traditional task-oriented dialogue, we upgrade an agent from acting as a mediator to fully-fledged conversational partner. The future challenges will be rewarding as we believe improving the human-like social capability of a system will greatly impact human-agent collaboration.

Chapter 5

Conclusions and Future Work

5.1 Thesis Contributions

This dissertation contributes to incorporating social intelligence into a dialogue system. We developed a theoretical framework and practical computational models for different modules in a socially-aware dialogue system.

- **Computational model of rapport:** Our 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, Ran** et al., 2014; Sinha et al., 2015).
- **Techniques for recognizing conversational strategy:** We have implemented a conversational strategy classifier to automatically recognize a 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 the speaker's and interlocutor's verbal, visual and vocal modalities in current and previous conversational turns, we successfully recognize these dialogue phenomena with an accuracy of over 80% and with a kappa of over 60% (**Zhao, Ran** et al., 2016a; Zhao, Tiancheng et al., 2016).
- **Techniques for assessing the dynamics of rapport:** 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 (e.g., eye gaze and smiles) and verbal conversational strategies (e.g., self-disclosure, shared experience, social norm violation, praise and back-channels). We developed a forecasting model involving a two-step fusion of learned temporal associated rules. The estimation of rapport comprises two steps: first, 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); second, the intuition is to learn the weight corresponding to each binary classifier for the rapport states 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, Ran** et al., 2016b).

- **Techniques for reasoning conversational strategy given interactional goals:** In collaboration with Oscar J. Romero, I designed a social reasoner that determines the appropriate conversational style and strategy that the dialogue system uses to describe the users desired information. This strengthens the relationship between user and system (rapport). I suggested we use the spreading activation model, a behavior network that consists of activation rules that govern the next conversation strategy adopted by the system, and designed the pre- and post-conditions of each conversational strategy. Oscar J. Romero implemented the spreading activation model. We conducted several experiments to validate the effectiveness of our social reasoner. The social reasoner was inspired by empirical data analysis of friends and stranger dyads engaged in a task, and by prior literature in fields including 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% compared to cases when no social reasoner is used (Romero et al., 2017).
- **Techniques for blending task and social goals in conversation:** We present a novel framework for developing a dialogue system that leverages expert supervision to blend task and social goals in negotiation. Within the framework, our socially-aware dialogue model operates two phases in parallel: a task phase and a social phase. It incorporates a dialogue encoder, which combines a social decoder and a task decoder that learns the dialogue context. In the task phase, our task decoder determines the next high-level coarse dialogue act based on its negotiation policy, which has been trained via maximizing the derived task reward. In the social phase, our social decoder deciphers an appropriate conversational strategy by maximizing the social reward. In that study, we employ a theory-driven, template-based natural language generator to realize the task intention as a genre of social conversational strategy. During the training process, our socially-aware dialogue model initially imitates the actions of human users and improves against task and social goals, respectively. We conducted comprehensive experiments to validate our system via human evaluation. Compared to baseline systems, our approach exhibits enhanced competence in achieving better task and social outcomes (**Zhao,Ran** et al., 2018; **Zhao, Ran** et al., 2019).

5.2 Future Work

In the future, we plan to further operationalize and scale up our established socially-aware framework by augmenting human-like capabilities for the dialogue system and extend our framework to other relationships.

5.2.1 Extensive Human-level Capabilities for Socially-aware Intelligence

Learn to memorize and refer to shared experience In this thesis, we claimed that indexing commonality and shared experience strengthens connections between humans and agents and further improves coordination during the interaction. Compared to other strategies, invoking shared experience proactively marks affiliation with a conversational partner and helps maintain (Clark, 1996) a long-term relationship. Therefore, identifying and employing the appropriate shared experience will allow an agent to manage interpersonal relationships broadly, from single interactions to cross-interaction (Gratch et al., 2007). In our past studies, the types of shared experience invoked were pre-defined and rigid. Thus, we can easily adopt a template-based approach to realize the appropriate strategy with slot-filling. To generalize this strategy to a daily talk, technically, we must endow a computational model with capabilities to read and write segments of potentially extensive memory components, and to combine this seamlessly with inference. We expect the class of advanced memory-network models (Weston et al., 2014) could chain the information from multiple interactions to reason congruous past experiences for an agent.

Learn to construct and exchange personal profiles Within our framework, self-disclosure of personal information or opinions is a compelling strategy signaling attentiveness and learning about an individual at the initial stage of the relationship. To effectively employ a self-disclosure strategy, a dialogue system should be capable of formulating and dynamically refining its own and its interlocutor's personal profiles. Based on our prior study, it is infeasible to hard-code all the agent's opinions or interests, which diminishes the value of our socially-aware framework in open-domain dialogues. To alleviate this problem, we suggest developing a learning paradigm to allow a dialogue system to extract opinions or interests from large-scale knowledge bases (e.g., news) or through interaction. Recently, groundbreaking research has made some progress (Li et al., 2014; Hancock et al., 2019; Li et al., 2016a; Tigunova et al., 2019). Fundamentally, in terms of detection, we should empower natural language understanding components to construct personal knowledge graphs or databases. Then, conditioned on the learned profile, a dialogue system can respond to its interlocutor while incorporating its unique opinions based on knowledge about the user, which can lead to feelings of liking and caring (Moon, 2000).

Learn to induce social and interpersonal norms A key aspect of our theory is that behavioral expectations are allied with sociocultural norms early in a relationship, and become more interpersonally determined as the relationship proceeds. Consequently, our dialogue system behaves politely in the first few interactions and applies teasing or generally-accepted rudeness only after it has demonstrated a tight bond with the interlocutor. To more easily implement norms, we define politeness as evidence of sociocultural norms and teasing as evidence of interpersonal norms. However, norms as the informal rules that govern behavior in groups and societies exist in more forms and evolve often. For instance, extensive studies (Yu and Zhang, 2013; Beheshti, 2015) have discovered a way of modeling legal norms, moral norms, and so on. Drawing on category theory, they created mathematical representations of the norms with various abstraction mechanisms. These abstractions made it possible to show relations among objects and became a necessary infrastructure to incorporating norms into an agent. To build on this work, then, one potential path is to leverage the power of cognitive architecture and design an agent that can infer, comply, and even occasionally violate social norms in conversation with respect to its own task and interpersonal goals. As a result, the human and agent will smoothly coordinate with each other in collaborative tasks.

Learn to converse with empathy As dialogue systems increasingly become part of daily life, it is important that they move beyond simply understanding the topics being discussed to also recognizing what their interlocutors are feeling. Social science literature has found that empathy contributes significantly to establishing rapport in different species (Norfolk et al., 2007). However, since the social function of empathy partially overlapped with face management and coordination strategies, we did not explicitly model for empathy in this thesis. Nonetheless, from an evolutionary perspective, empathy is considered to create or enhance bonds between individuals (Plutchik, 1987), and recent studies have identified the importance of modeling empathy on agents to facilitate social competence in light of these human principles (Paiva et al., 2017). They suggest that empathy encompasses many other related phenomena such as emotional contagion. Therefore, we believe that endowing a dialogue system with empathetic capabilities will be another bright path to managing interpersonal closeness in human-agent interaction.

5.2.2 Generalization across Various Relationships

In this thesis, we apply our socially-aware framework to peer tutoring and negotiation conversations. We chose peer tutoring to represent fully-cooperative conversations and negotiation to represent semi-cooperative conversations. With slights adjustments, we validate the effectiveness of our framework on managing these two completely different interpersonal relationships. Further, we want to expand the applications of the framework to other interpersonal relationships, such as advisorship, therapeutic, companionship, coworkers etc. For instance,in an advisorship, we can examine if social strategy could facilitate the course discussion process between a student and an academic advisor (Kummerfeld et al., 2019). In a therapeutic relationship, social science literature (Egbert et al., 1963; Bakić-Mirić and Bakić, 2008; Leach, 2005) highlighted the importance of rapport on patient satisfaction, treatment compliance and client outcomes. Recent advances in artificial intelligence studies employ conversational agents with unconstrained natural language input capabilities in health and medical care (Laranjo et al., 2018), which demonstrates potential benefits to health across a broad range of application domains, like assisting clinicians during consultations, supporting consumers with challenges to behavioral change, or even assisting patients and elderly individuals in their living environments (Wolters et al., 2016; Bickmore et al., 2005). Since these scenarios require human and agent collaboration—where social interaction greatly impacts the outcome—we believe modeling interpersonal dynamics with our socially-aware framework will foster the next generation of interoperability in healthcare.

5.2.3 Closing

Our future work evokes core AI challenges, such as representing many aspects of user profiles. Nevertheless, we believe that this thesis has made the first step towards a realistic socially-aware dialogue system. We expect future challenges to be substantial but rewarding, as we begin to model those aspects of human-human interaction that promote human-agent collaboration and protect what we cherish most about being human.

Appendices

Appendix A

Coding Conversational Strategies

- **Self-Disclosure (SD):** Self-disclosure refers to the conversational act of revealing aspects of oneself (personal private information) that otherwise would not be seen or known by the person being disclosed to (or would be difficult to see or know). Psychological literature often discusses the ways people reveal facts about themselves to build relationships, but we are the first to look at the role of self-disclosure during social and task interactions by the same dyad, particularly for adolescents engaged in reciprocal peer tutoring. We coded for two sub-categories: (1) revealing the long-term aspects of oneself that one may feel are deep and true ("I love my pets"), (2) revealing one's transgressive (forbidden or socially-unacceptable) behaviors or actions, which may be a way of attempting to make the interlocutor feel better by disclosing one's flaws ("I suck at linear equations").
- **Referring to Shared Experience (SE):** We differentiate between shared experience - an experience that two interlocutors engage in or share with one another at the same time ("that facebook post Cecily posted last week was wild!") - from shared interests ("you like Xbox games too?"). Shared experiences may index a shared community membership (even in a community of two), which can in turn build rapport. We coded for shared experiences (e.g., going to the mall together last week).
- **Praise (PR):** We annotated both labeled praise (an expression of a positive evaluation of a specific attribute, behavior or product of the other; e.g., "great job with those negative numbers"), and unlabeled praise (a generic expression of positive evaluation, without a specific target; e.g., "Perfect").
- **Violation of Social Norms (VSN):** Social norm violations are behaviors or actions that go against generally socially acceptable and stereotypical behaviors. In a first pass, we coded whether an utterance was a social norm violation. In a second pass, if a social norm violation, we differentiated: (1) breaking the conversational rules of the experiment (e.g., off-task talk during tutoring session, insulting the experimenter or the experiment, etc.); (2) face threatening acts (e.g., criticizing, teasing, or insulting, etc.); (3) referring to one's own or the other person's social norm violations or general social norm violations (e.g., referring to the need to get back to focusing on work, or to the other person being verbally annoying, etc.). Social norms are culturally-specific, and so we judged a social norm violation by the impact it had on the listener (e.g., shock, specific reference to the behavior as a violation,

etc.). Social norm violations may signal that a dyad is becoming closer, and no longer feels the need to adhere to the norms of the larger community.

Appendix B

Coding Visual Behaviors

- **Eye Gaze:** Gaze for each participant was annotated individually. Front facing video for the individual participant was supplemented with a side camera view. Audio was turned off so that words didn't influence the annotation. We coded (1) Gaze at the partner (gP), (2) Gaze at one's own worksheet (gO), (3) Gaze at partner's worksheet (gN), (4) Gaze elsewhere (gE).
- **Smile:** A smile is defined by the elongation of the participant's lips and rising of their cheeks (smiles will often be asymmetric). It is often accompanied by creases at the corner of the eyes. Smiles have three parameters: rise, sustain, and decay (Hoque et al., 2011). We annotated a smile from the beginning of the rise to the end of the decay.
- **Head Nod:** We coded temporal intervals of head nod rather than individual nods - the beginning of the head moving up and down until the moment the head came to rest.

Appendix C

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

Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *Proceedings of the First nordic symposium on multimodal communication*. Copenhagen, pages 7–22.

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.

Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin* 23(4):363–377.

Bernard J Baars. 2003. The global brainweb: An update on global workspace theory. *Science and Consciousness Review* 2.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Nataša M Bakić-Mirić and Nikola M Bakić. 2008. Successful doctor-patient communication and rapport building as the key skills of medical practice. *Facta Univers* 15(2):74–9.

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.

Rahmatollah Beheshti. 2015. Modeling social norms in real-world agent-based simulations. *Thesis dissertation* .

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 *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 262–269.

Timothy Bickmore and Julie Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. In *Proceedings of the AAAI fall symposium on narrative intelligence*. pages 87–92.

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, Laura Pfeifer, and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. In *Proceedings of the International Conference on 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 *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 390–403.

Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. It's just like you talk to a friend relational 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. *Applied Linguistics* pages 196–213.

Margaret A Boden. 2010. Conversationalists and confidants. *Artificial Companions in Society: Perspectives on the Present and Future* page 5.

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

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, 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.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 152–166.

Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. 2014. From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 120–133.

Herbert H. Clark. 1996. *Using Language*. Cambridge Universtiy Press.

William Coon, Charles Rich, and Candace L Sidner. 2013. Activity planning for long-term relationships. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, volume 8108, page 425.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-computer dialogue simulation using hidden markov models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*. IEEE, pages 290–295.

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.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, pages 80–87.

Lawrence D Egbert, George E Battit, Herman Turndorf, and Henry K Beecher. 1963. The value of the preoperative visit by an anesthetist: A study of doctor-patient rapport. *Jama* .

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.

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.

Angela Fabregues and Carles Sierra. 2011. Dipgame: a challenging negotiation testbed. *Engineering Applications of Artificial Intelligence* 24(7):1137–1146.

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*.

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 *Proceedings of the 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 *Proceedings of the International Conference on 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 *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 125–138.

Agustín Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of the INTERSPEECH*. pages 1019–1022.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research* 5(Nov):1471–1530.

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*. pages 159–174.

Mathieu Guillame-Bert and Artur Dubrawski. 2014. Learning temporal rules to forecast events in multivariate time sequences. In *Proceedings of the 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.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415* .

Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now: Introducing the virtual human toolkit. In *Int. Conf. on Intelligent Virtual Humans*. Edinburgh, UK.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2333–2343.

Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. 2007. Searching for prototypical facial feedback signals. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, pages 147–153.

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.

Mohammed Hoque, Louis-Philippe Morency, and Rosalind W Picard. 2011. Are you friendly or just polite?–analysis of smiles in spontaneous face-to-face interactions. In *Proceedings of the Affective Computing and Intelligent Interaction*, Springer, pages 135–144.

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

Emmanuel Johnson, Jonathan Gratch, and David DeVault. 2017. Towards an autonomous agent that provides automated feedback on students' negotiation skills. In *Proceedings of the 16th*

Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pages 410–418.

Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* 33(2):125–143.

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 L Ferrin. 2014. Interpersonal trust within negotiations: Meta-analytic evidence, critical contingencies, and directions for future research. *Academy of Management Journal* 57(5):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.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S. Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25(9):1248–1258.

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.

Matthew J Leach. 2005. Rapport: A key to treatment success. *Complementary Therapies in Clinical Practice* 11(4):262–265.

Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* .

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* .

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* .

Xiang Li, Gökhan Tür, Dilek Z. Hakkani-Tür, and Qi Li. 2014. Personal knowledge graph population from user utterances in conversational understanding. *2014 IEEE Spoken Language Technology Workshop (SLT)* pages 224–229.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016b. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688* .

Gale Lucas, Giota Stratou, Shari Lieblich, and Jonathan Gratch. 2016. Trust me: multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, pages 5–12.

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 *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Yoichi Matsuyama, Arjun Bhardwaj, **Zhao,Ran**, Oscar J. Romero, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL'16)* ..

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

Johnathan Mell and Jonathan Gratch. 2016. IAGO: Interactive Arbitration Guide Online. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Singapore, pages 1510–1512.

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.

Mario Neururer, Stephan Schlgl, Luisa Brinkschulte, and Aleksander Groth. 2018. Perceptions on authenticity in chat bots. *Multimodal Technologies and Interaction* 2(3).

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.

Tim Norfolk, Kamal Birdi, and Deirdre Walsh. 2007. The role of empathy in establishing rapport in the consultation: a new model. *Medical education* .

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. 2012. Rudeness and rapport: Insults and learning gains in peer tutoring. In *International Conference on Intelligent Tutoring Systems*. Springer, pages 11–21.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* .

Alexandros Papangelis, **Zhao, Ran**, and Justine Cassell. 2014. Towards a computational architecture of dyadic rapport management for virtual agents. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'14)* ..

Taylor Pederson. 2013. Establishing rapport to increase compliance. *Psychology Student Work* 12.

James W Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Robert Plutchik. 1987. Evolutionary bases of empathy. *Empathy and its development* .

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

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*.

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 *Proceedings of the 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. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)* ..

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

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.

Tanmay Sinha and Justine Cassell. 2015a. Fine-grained analyses of interpersonal processes and their effect on learning. In *Proceedings of the 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, **Zhao, Ran**, and Justine Cassell. 2015. Exploring socio-cognitive effects of conversational strategy congruence in peer tutoring. In *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI'15)* ..

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. 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.

Fan-Yun Sun, Yen-Yu Chang, Yueh-Hua Wu, and Shou-De Lin. 2018. Designing non-greedy reinforcement learning agents with diminishing reward shaping pages 297–302.

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):74.

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

Zhao, Ran, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'14)* ..

Zhao, Ran, Oscar J Romero, and Alex Rudnicky. 2019. Learning to blend social and task goals in conversational agents through self-play.

Zhao, Ran, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL'16)* ..

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

Zhao,Ran, Oscar J Romero, and Alex Rudnicky. 2018. SOGO: A social intelligent negotiation dialogue system. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'18)* ..

Leigh L Thompson and Reid Hastie. 1987. *Judgement tasks and biases in negotiation*. National Institute for Dispute Resolution.

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

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. *The World Wide Web Conference on - WWW 19* .

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 *Proceedings of the International Conference on 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 *Proceedings of the 13th SIGDial*. 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.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* .

Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274* .

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2):270–280.

Maria Klara Wolters, Fiona Kelly, and Jonathan Kilgour. 2016. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. *Health informatics journal* .

Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal public speaking performance assessment. In

Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, pages 43–50.

Tongkui Yu and Li Zhang. 2013. An agent-based model of social norm and cooperationp. *Procedia Computer Science* .

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 W Black, and Justine Cassell. 2013. Automatic prediction of friendship via multi-model dyadic features. In *Proceedings of the 14th SIGdial*.

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* .

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

Zhao, Tiancheng, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. *arXiv preprint arXiv:1902.08858* .

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