Trust in Human and Autonomous Agent Interaction

Nuno Xu
nuno.xu@tecnico.ulisboa.pt

Supervisor: Rui Prada
Co-Supervisor: Ana Paiva
Técnico Lisboa (Taguspark)
Universidade de Lisboa
Av. Prof. Dr. Aníbal Cavaco Silva
Porto Salvo, Portugal
http://tecnico.ulisboa.pt/en/

Abstract What thesis is about and contributions

Key words: rapport, Human Robot Interaction (HRI)

Contents

1 Introduction

When you engage in a good conversation with someone, that feeling of flow and connection is what formally known as rapport.

These stud- ies demonstrate the necessity of robots displaying appropriate so- cial behaviors to effectively achieve social, cognitive, and task out- comes in human-robot interaction

Robots are increasingly becoming part of our society and its presence has been proven to impact our lives. But does any of us remember a remarkable interaction with a robot as we are able to recall one with a person? What makes one conversation memorable? How can we design robots that can achieve something that has so much impact and yet, people do so easily?

In order to ask these questions, the HRI research community has been exploring agents capable of responding emotionally and more humanly in dyadic interactions to create empathy and positivity. Formally, this phenomenon is known in psychology as building rapport which will be more detailed in Section **??**.

For that purpose, several studies were conducted to identify how people can be manipulated using different verbal and non-verbal strategies. There is evidence in these studies that rapport agents can make people feel: more connected [**?**], less tension [**?**], less embarrassed [**?**], mutual trustworthiness [**?**].

Rapport agents can be applied in social robots in several domains such as negotiation [**?**], child care [**?**], therapeutic sessions [**?**], family companions [**?**, **?**], hospitals [**?**] [**?**], weight loss [**?**], children with autism [**?**, **?**], companions at home [**?**, **?**] and several other examples.

But, despite present efforts there are still important issues to be addressed, namely:

- Lack of coordination during interactions due to poor timing predictability;
- Most of the work in the area is focused on short-term rapport and therefore does not take into account how relationships evolve and how rapport strategies are molded by it;
- Lack of rapport datasets;
- There is not a common framework nor model to manage rapport;

Lastly, we will develop a rapport component using SERA frameworkd (XX) and EMotive headY System (EMYS) and demonstrate our results in a negotiation game scenario how the developed agent can affect the emotional state and affect the trust level of the opponent. The chosen game is Split Or Steal during the negotiation phase scenarios and in trust such as Split Or Steal (YY).

1.1 Objectives

The main goals of the proposed project are:

- Create a rapport model to be integrated with the SERA unified framework;
- Create a rapport model based on machine learning algorithms;

- Create a rapport model using Wizard of Oz
  Integrate a rapport agent in the internal SERA unified framework;
- Conduct a study in a negotiation scenario to study the impact of cooperation.

In section X we introduce the background research to understand the concepts of rapport, social robotics and machine learning. In section Y, it will be described the state-of-art rapports agents developed by X, Y and Z. The proposed solution and how its quality will be evaluated will be described in section W. Lastly, the final conclusions and the main concerns will be address in section U.

The remainder of this paper is organized as follows. In the next section, we discuss related work on learning social behavior synthesis models. We introduce the IPL approach in Section III. In Sections IV and V, we describe, respectively, the setup and the results of an experiment on the synthesis of backchannel timings. We conclude with Section VI.

2 Background

For the following sections, it will be described the main concepts that drive the development of the current project: rapport, social robotics and machine learning in section **??**.

3 Related Work

Computational Trust research has been very focused on modelling trust in Multi-Agent Systems (MASs), specially on open e-commerce environments [1–5], with at least 106 models created [1], since the formalization of trust as a measurable property in 1994 [6].

The work I will be comparing my work with (direct influence, biggest part of this document, has to make sense, it has to show the relevance of mine work in a international way)

In data-driven timing generation for social behaviors, using Machine Learning (ML) techniques, it is retrieved two types of samples: positive samples representing the moments (or timings) in the interaction that are socially acceptable to generate a backchannel gesture (e.g., a head nod or a vocalization "hmm hmm") and negative samples as the moments when it is unacceptable. In previous approaches, corpus based, positive samples are taken directly from the annotated dataset and the negative randomly as long as they do not overlap with a positive sample.

3.1 Discussion

Based on everything and what everyone is doing or done, these are the problems, these are the advantageous and disavantageous of the aproach.

4 Proposed Solution

- Architecture/Model: does not need to be very detailed but enough to understand what it is trying to be done. Tools being used, etc
- Evaluation: How the work will be evaluated.
- Planning (GANTT maybe)

Aproach - Makes sense because you talked previously

As [paper do iterative perceptual learning], using hand crafted rules [7, 12] despite being intuitive use very shallow features and the development of these rules is not trivial. E.g.

4.1 Subsection Heading Here

Subsection text here. A figure can be inserted like the example of Figure **??**.

According to previous studies [33 dont stare at me], during dyadic interactions, the listener usually maintains long gazes at the speaker and only interrupts briefly from time to time.

In fact, [9 toward dyadic...] found that in a negotiation setting not reciprocating negative self-disclosure led to decreased feelings of rapport. [**?**]

mutual gaze in determine turn-taking turn-taking [8] [12] [14] [56] [47] [todos do dont stare at me].

One of the most notable non-verbal behaviors to build rapport is gaze because it is a clear signal of mutual attention (as our parents said, look to people eyes when talking to them), acts as an invitation to interaction, increases dynamism, likelability and believability [4 do dont stare at me]

its impact in a wide range of interpersonal domains includ- ing social engagement [52], classroom learning [22], suc- cess in negotiations [20], improving worker compliance [18], psychotherapeutic effectiveness [59], and improved quality of child care [11].

Gaze as object of interest [8] [37] [55]. , effects on the way communication proceeds [54] [60] [23] [19] [28].

In fact, previous work demonstrated that there is evidence that in health domains, high rapport doctors engaged in less extensive eye-contact than low rapport doctors , 85% and 70% of the interaction time respectively . However, the impact of the gaze depends if the interacts are in a helping context (e.g. meetings with a doctor) or in a non-helping context (e.g. interviewing) [dont stare at me]. On the latter, directed gaze is correlated positively with participant's evaluative impression [Tickle-Degnan and Rosenthal]. In interviweing ocntexts, Goldberg, Kiesler, and Collins [25] found that people who spent more time gazing at an interviewer received higher socio- emotional evaluations.

Argyle [1] found that in dyadic conversa- tions, the listener spent an average of about 75

Kendon [33] reported that a typical pattern of interaction when two people converse with each other consists of the listener maintaining fairly long gazes at the speaker, interrupted only by short glances away.

In short, gaze can also have negative impact if not dosed correctly.

To assess if negative arousal played some role, we asked the partici- pants to evaluate how uncomfortable they were when inte- racting with the agent (the embarrassment scale) in the post-questionnaire packet. ANOVA

5 Evaluation

Evaluation

6 Conclusion

The conclusion goes here. This is more of the conclusion.

Acknowledgment

The author would like to thank...

References

1. Granatyr, J., Botelho, V., Lessing, O.R., Scalabrin, E.E., Barthès, J.P., Enembreck, F.: Trust and Reputation Models for Multiagent Systems. ACM Computing Surveys **48**(2) (oct 2015) 1–42
2. Han Yu, Zhiqi Shen, Leung, C., Chunyan Miao, Lesser, V.R.: A Survey of Multi-Agent Trust Management Systems. IEEE Access **1** (2013) 35–50
3. Pinyol, I., Sabater-Mir, J.: Computational trust and reputation models for open multi-agent systems: a review. Artificial Intelligence Review **40**(1) (jun 2013) 1–25
4. Noorian, Z., Ulieru, M.: The State of the Art in Trust and Reputation Systems: A Framework for Comparison. Journal of theoretical and applied electronic commerce research **5**(2) (aug 2010) 97–117
5. Huang, H., Zhu, G., Jin, S.: Revisiting Trust and Reputation in Multi-agent Systems. Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on **1** (2008) 424–429
6. Marsh, S.P.: Formalising Trust as a Computational Concept. PhD thesis (apr 1994)

Appendices

A Variants of Single-subject designs

B Educational software

landscape page

Dirk Haylen et al., developed an iterative model of developing a data-driven model for generating timings for backchannels behaviors in a dyadic conversational setting. On the one hand, it is data-driven because it uses ML algorithms. On the other hand, it is iterative because the learning phase is done through several iterations, each one is more refined than the previous one contributing to the overall growth of the rapport agent's quality.

It will be first described the system developed by the authors and the issues they focused on. Following, it will be described how the system was evaluated with sufficient detail. To conclude, it will be discussed the innovation introduced in this system comparing with previous models followed by referring the important positive and negative aspects as well how the work developed by Dirk Haylen et al. is relevant to the proposed solution section 4.

System description In data-driven timing generation for social behaviors, using ML techniques, it is retrieved two types of samples: positive samples representing the moments (or timings) in the interaction that are socially acceptable to generate a backchannel gesture (e.g., a head nod or a vocalization "hmm hmm") and negative samples as the moments when it is unacceptable. As described in section XX, in previous approaches, corpus based, positive samples are taken directly from the annotated dataset and the negative randomly as long as they do not overlap with a positive sample.

Taken this into account, the author mentions that this approach potentially leads to a great number of false negative by not taking into account that social signals are optional and a reflection of the listener's personality and therefore behavior different from the corpus can also be socially appropriate. As such, in order to tackle this problem they identified which moments in the interaction are seen as socially inappropriate using subjective rating in each iteration and used this information to improve the quality of the model.

In the proposed iterative approach, both positive and negative samples are collected to further refine the classifier using a sequence of a bootstrap and iterations of three procedures: generation, evaluation and learning.

During the generation (pink areas on Figure 1), three steps occur:

1. **Extraction of features**: Using the available sensors (for example camera and microphone), it is created a feature vector with the representation of dialog's partner's behavior at a given instance;
2. **Classification of features**: Each feature vector is assigned a score, using the model trained in previous iterations, representing the probability of generating a backchannel in a given instance is socially appropriate;
3. **Stimulation generation**: the scores are transformed to a sequence of social behavior timings and with application of heuristics (e.g., top N scores, minimum score value or maximum number of timings per minute) the listener's social behavior animation is computer generated and synchronized with the speaker's video to be evaluated in the evaluation procedure.

During the evaluation (blue areas on Figure 1), using **PCS!** (**PCS!**) (PÔR NO BACKGROUND), multiple subjects rate the quality of the timings gener-

Figure 1: Taken from [**?**]. Schematic representation of the Iterative Perceptual Learning framework. The generation, evaluation and learning stage are shown in pink, blue and green, respectively.

ated by the model by pressing a button whenever they would rate the agent's behavior as socially inappropriate (the *yuck* button, introduced by Poppe et al. [15]). Filters such as minimum number of *yuck*s per social behavior or mispresses can be applied. Additionally, it is taken into account the typical delay when pressing the button.

During the final procedure, learning (green areas on Figure 1), the retrieved positive and negatives samples are used to train the classifier. Each iteration contributes with more samples to model and therefore increasing the quality of the identified timings for generation of the agent's social behaviors. As the author describes, this can be seen as a form of reinforcement learning by focusing on the relevant moments.

During the bootstrap (yellow area on Figure 1): Since, it is not possible to have non-random negative samples before using subjective evaluation the classifier will be trained using the corpus based approach: the positive samples are taken directly from the corpus and the negative samples are taken at random instance that do not overlap with the first samples. After generating and evaluating the generated timings (using the *yuck* button), these negative samples are discarded and replaced by the ones obtained form the subjective evaluation.

This approach allows the model to be refined in each iteration and focus on the negative samples subjectively evaluated by multiple users using the **PCL!** (**PCL!**) approach. Moreover, the model is refined in each iteration and has more understanding about proper timings for social behaviors.

Experiment The developed **IPM!** (**IPM!**) model was compared with the traditional corpus-based on a face-to-face conversation with a virtual agent as listener and a human subject as speaker. It was used Suport Vector Machine (SVM) as classifier with the default values: RBF kernel with $c = 1$ and $\gamma = 1/x$, where $[|x|$ is the dimensionality of the input vector. The corpus used to train and evaluate the model is the Dutch-spoken *MultiLis* corpus [21] with 131 minutes of dyadic conversations between two subjects in cubicles interacting with one another using video conferencing with the camera positioned behind an interrogation mirror with the other user projected to recreate the looking into the eye. From the audio and the video of the *MultiLis* corpus it was extracted three types of features: prosody (112 features), speaking (1 feature) and looking (1 feature based on manual annotations). The pre-processing details are dependent on the used corpus and are detailed in the paper [X].

During the creation of the baseline (bootstrap), the training was done using a frequency of 100 Hz. The positive samples were the first frame in the annotated corpus. However to make the classifier less dependent on such single frame it was selected 4 frames around it such that they were sampled using a normalized

Gaussian distribution with a ϵ with 95% of the samples within 250ms of the positive sample. Finally, it was selected an equal number of negative samples and smoothed the output of the SVM to create curves that represent the appropriateness to provide a backchannel. The local peaks were used to identify the moments in the interactions adequate to produce a backchannel.

During the generation procedure it was generated 25% more backchannels than the determined mean backchannel rate over all interactions in the MultiLis corpus to collect more negative samples to be used in subsequent iterations. During the evaluation process, participants had to press the *yuck* button whenever they perceived an individual backchannel from the virtual listener as inappropriate. The presses were matched to the last preceding backchannels if within 5000 ms of the onset. Finally, it was measured the number of *yucks* for each synthesized backchannel. In the last step, learning, the negative samples that were taken randomly are discarded and substituted by those collected during the subjective evaluation. The number of positive and negative samples was balanced using Gaussian models, sampling factor and number of negative samples. The only restriction applied was that two backchannels could not be within 2 seconds from each other.

For the experiment it was used one bootstrap phase and 4 iterations using the three procedures mentioned. The stimuli was a speaker's video from the corpus with a animated listener synchronized. The virtual agent only nodded his head while making an utterance whenever appropriate. As the author refers that head movement, posture shifts, facial expressions and eye blinks were not animated to not impact the focus of the subject. For each iteration, all positive and negative samples obtained from all previous iterations were used to learn the IPL model. Only the negative samples from the bootstrap iteration were discarded since they were evaluated subjectively by the users as inappropriate.

Evaluation Both models, corpus based and **IPL!** (**IPL!**) were trained using the same interactions to make a fair comparison. It is important to note that the evaluation results for the IPL double as negative samples for model learning in the subsequent iteration.

Participants of a 30 minutes experiment were shown stimuli through a webpage and that they were explained that they would had to evaluate the quality of the synthesized listening behavior. The subjects has to press the spacebar each time the virtual listener perfomed a backchannel they judged as inapproriate and they replay and change or discard their choices. It was shown both the IPL model and the corpus based model allowing the comparison between them pair-wise.

To measure the results, it was used two measurements:

- Objective measure: Compare the predicted timing of the backchannel with those performed by the actual listener in the MultiLis corpus. The precision and recall was combined using weighted harmonic mean into $F_1 = \frac{2pr}{p+r}$;

– Subjective measure: Use the *yucks* collected during the perceptual evaluation. It was calculated the percentage of backchannels that did not receive any yucks and the average of *yucks* per backchannel;

After iteration the corpus based model was compared with the **IPL!** model.

Discussion