A Method to Evaluate Response Models

Merijn Bruijnes, Sjoerd Wapperom, Rieks op den Akker, and Dirk Heylen

Human Media Interaction, University of Twente,
P.O. Box 217, 7500 AE, Enschede, The Netherlands
m.bruijnes@utwente.nl

Abstract. We are working towards computational models of mind of virtual characters that act as suspects in interview (interrogation) training of police officers. We implemented a model that calculates the responses of the virtual suspect based on theory and observation. We evaluated it by means of our test, the "Guess who you are talking to?" test. We show that this test can contribute to building response models for believable virtual agents.

Keywords: Response Model, Evaluation, Virtual Agent.

1 Introduction

We work towards a virtual agent that can play a suspect in a serious game that can be used by police students to hone their skills in police interviewing. A virtual agent needs three main components to be able to have a meaningful interaction. The actions of the user have to be sensed and interpreted (e.g. the user says "Confess, criminal!" which is dominant and aggressive behaviour). This interpretation provides the input to a response model that provides the reasoning of the agent (e.g. the user is dominant and aggressive which makes me sad and angry). A response model should take into account the specific role that the agent plays. In this case that is a suspect with all the tactics and psychological manoeuvring that is involved. A response model based on human behaviour can be used to make the behaviour of a virtual agent more believable to humans [5]. Based on the state of the response model the agent can select the most appropriate behaviour in its repertoire (e.g. make a sad face and say "You're not nice!"). The human responds to the agent and the cycle continues.

In this paper we discuss a method to evaluate response models. We focus on the consistency with which a response model (and thus an agent using this response model) can portray a personality. We present a way to evaluate only the response model, in an abstract interaction without actual linguistic content. We report the evaluation of a suspect response model based on the work in [3].

2 Method for Evaluation of Response Models

In this Section we present our method for evaluating response models and we show the viability of this method by evaluating a response model. The response

T. Bickmore et al. (Eds.): IVA 2014, LNAI 8637, pp. 67–70, 2014.

model is based on the work in [3] where we analysed the DPIT-corpus [1] to get insight into the social behaviour of police officers and suspects in the police interview setting. We collected terms that people use to describe the interactions in the corpus. A factor analysis revealed factors that could be interpreted as relating to the theories of *interpersonal stance* [4], *face* [2], and *rapport* [6] and the meta-concepts *information* and *strategy*. Our response model can portray a persona based on settings in the response model that are based on these theories.

We want to know whether a response model can portray a persona in a recognizable and consistent way using our "Guess who you are talking to?" test. Participants interact with the response model and have to guess which of a selection of personas is portrayed by the system. In our method, we evaluate the response model in an abstract manner, without the ambiguity of specific utterances that stem from the semantics of the utterances rather than the emotional and pragmatic variables that the model is intended to account for. Evaluating a response model using utterances that have a subjective quality introduces two sources of ambiguity related to the experiment: during the creation of the utterances (e.g. by the virtual agent) and during the interpretation of the utterance (e.g. by the user). The following examples show an interaction of two utterances (`1u` and `2u`) that are ambiguous and the 'intended' interpretation of these utterances in terms of the response model (`1i` and `2i`):

`1u` Police: "Why did you hide the body?"
`1i` Intention of the Police in terms of the response model: "Open Question, Dominant Stance, Politeness is Direct, Indication of Guilt, ..., Case Related Frame"
`2u` Suspect: "None of your business!"
`2i` Intention: "Aggressive Stance, Short Answer, Strategy Avoiding, ..., Unfriendly"

Some utterances leave room for interpretation and the reader might interpret these sentences different from how they should be interpreted according to the writer. In our method, participants interact with a response model in an abstract manner. This means the interaction takes place in the terms of the response model: the user is his own wizard of Oz. This way there is no confusion between what a writer meant and what he wrote down, and what the participant read and what he thought the writer meant. However, this comes at a cost. The participants need to be instructed on the abstract factors that the model uses and the personas that are portrayed by the model.

The participants have at least two sessions of interactions with the response model, once with one of the personas and once with a random response generator (not based on a persona or response model). During each session they are asked to indicate with which of the personas they think they are interacting. In addition, the participants are asked how confident they are about their choice, how realistic they found the interaction, and how familiar they are with the concepts and terms used in the response model. Finally, after each session they are asked about their experiences during the interaction.

Our Response Model Tester consists of two graphical frames that users see and use during interactions with the 'suspect agent'. These frames handle all input from and output generation to the user. The input the user gives in the police frame is the police contribution to the interaction. This input is given in the

terms of the response model, see example `1i` above. The input is passed to the response model that calculates the suspect behaviour. This suspect behaviour is depicted in the suspect frame again in terms of the response model, see example `2i` above. All response model input and output, and the participant's choices, confidence, and realism ratings are logged.

2.1 Participants and Evaluation

For our evaluation, 48 participants (42 male, mean age 24.8 with SD 3.7) volunteered to take part in the study.

Three personas were created, based on personas from the DPIT-corpus [1,3]. Each persona was introduced in a short text. Participants received elaborate explanation of the factors in the response model (e.g. stance) and the aspects of the contributions of both the police and suspect (e.g. an aggressive stance). Each factor was explained and illustrated with several examples. Participants were encouraged to ask questions if something was unclear to them. Once everything was clear, they could start playing with the response model.

2.2 Results and Discussion

A total of 39 (81.25%) participants guessed correctly with which persona they were interacting after eight interactions. Participants who were correct were (significantly: $Z = -2.001, p < 0.1$) more confident (4.41) compared to the participants who were incorrect (3.67) (rated on a 5-point Likert scale (1=strongly disagree, 5=strongly agree)). The realism rating was similar: 3.90 for correct compared to 3.89 for incorrect. In the interactions where the responses of the system were random we might expect that each of the personas would be chosen an equal number of times (33%). However, the distribution of choices for the personas was 62.5%, 20.8%, and 16.7%. The average confidence level for interactions with personas was significantly higher 4.27 (SD = 0.76) compared to 3.46 (SD = 0.77) for the random interactions ($Z = -4.2, p < 0.00$). The average level of realism for personas was significantly higher 3.90 (SD = 0.52) compared to 3.35 for random rounds (SD = 0.89) ($Z = -3.7, p = 0.001$).

After the experiment, we informally asked participants about their experiences during the experiment. People who interviewed the *random generator first* reported that they started doubting their decision on the first persona after they had interacted with the second persona. They felt more confident about choice for the second persona. They also felt the first to be more random after they had interviewed the second. They reported the second persona met their expectations of one of the three personas. Some participants struggled with the feeling that when they had chosen a persona for the random output they felt they could not pick that persona again at their second run. They felt this way because the output was different from the first and they did feel some sort of confidence about their first choice. This led to some people mistakenly choosing a different persona from the one they chose earlier. People tended to base their decision on parts of the output generated by the persona, they did not always

look at all the output. They tried to rationalize 'weird random output' and actively tried to find reasons to consider it as correct and realistic. Also, we asked on which aspects of the suspect response they based their decision. Most participants based their output only on parts of the suspect response. However, the part they focussed on differed and across all participants all of the suspect response output was used.

3 Conclusion

The results of this "Guess who you are talking to" test give an indication that our response model generates responses to user actions in such a way that the user is able to recognize a persona. This gives evidence of the validity of the response model and it promises that the model can be used in the implementation of believable virtual suspect characters with various personal characteristics as we encountered in our police interview corpus.

The method of evaluation of response models gives insight into the consistency with which a response model can portray a personality. It provides hints for improvements of the response model. Investigating which aspects of the model's response participants that 'guess wrong' focus on can provide hints for improvements of the model on these aspects. It is possible to investigate how each part of the response model's response contributes to a 'correct guess' of participants by showing only some parts to different participants and comparing their 'correct guess-scores'. In addition, when comparing several settings for a persona our evaluation method can show which setting is recognized most consistently as this persona, thus showing the 'optimal settings of the persona' in the response model.

Acknowledgements. This publication was supported by the Dutch national program COMMIT.

References

1. op den Akker, R., Bruijnes, M., Peters, R., Krikke, T.: Interpersonal stance in police interviews: content analysis. Computational Linguistics in the Netherlands Journal 3, 193–216 (2013)
2. Brown, P., Levinson, S.C.: Politeness: Some universals in language usage. Cambridge University Press, Cambridge (1987)
3. Bruijnes, M., Linssen, J., op den Akker, R., Theune, M., Wapperom, S., Broekema, C., Heylen, D.: Social behaviour in police interviews: Relating data to theories. In: Conflict and Negotiation: Social Research and Machine Intelligence (2014)
4. Leary, T.: Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation. Ronald Press, New York (1957)
5. Steunebrink, B.R., Dastani, M., Meyer, J.J.C.: A formal model of emotion triggers: an approach for BDI agents. Synthese 185(1), 83–129 (2012)
6. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychological Inquiry 1(4), 285–293 (1990)