Deceptive Agents and Language[*]

(Extended Abstract)

Mark Dras Debbie Richards Meredith Taylor Mary Gardiner
Department of Computing, Macquarie University, Australia
{madras,richards,mtaylor,gardiner}@science.mq.edu.au

ABSTRACT

The use of virtual agents in training requires them to have several human-like characteristics; one of these is the ability to appear deceptive. We take work from the psychology literature on cues to deception, with a focus on language-related cues, and examine whether it is possible to use resources from the field of Language Technology to construct scenarios with agents showing cues to deception detectable by human judges, a task that has been shown in a text-only context to be difficult. We show that this detection is in fact possible in the context of virtual agents, and that there are interesting results for individual cues, in particular for dialogue- versus lexical-level cues, and a 'placebo' effect.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents*

General Terms

Human Factors, Experimentation

Keywords

Agents in virtual environments, language technology

1. INTRODUCTION

Although there is a reasonable amount of work on deception in human-like agents [3] and generating language by these agents [4], no work has as yet linked language and deception in agents. Our context for being interested in this is in a training environment connected to security. Our prototype system seeks to improve the decision making skills of border security offers who need to answer questions such as "is the passenger suspicious?". There is extensive literature about the characteristics of deceptive behaviour: a meta-study that conveniently summarises much of this material is [1]. Some of the characteristics or cues that have been found to correlate with deception are physical, such as the amount of fidgeting; many are language-related, including negativity of language and verbal immediacy.

A question that then arises is whether we can construct agents that display these kinds of language behaviours in a way that is recognisable by humans as indicating deception. Research in a purely text-based context [2] has suggested that humans are fairly poor at identifying deception based on linguistic cues, although this work only asked human judges to classify text as deceptive or genuine, rather than try to teach them to identify the cues. Related to this question, there is thus also the issue of producing these cues for agents to display. Our aim in the work described in this paper, then, is to conduct a first study into whether the field of language technology allows us to construct linguistically deceptive agents. To investigate this we carry out an experiment to determine whether human judges are able to detect the cues to deception displayed by the agents.

2. METHOD

Our scenarios were implemented in a prototype system we have developed, BOrder Security System (BOSS), in which the user views a scenario in an immersive 3D virtual environment populated with avatars. The dialogues of the scenarios were designed in pairs, with one of a pair to be the default one, and the other to be the deceptive variant. Both scenario pairs were in the context of a border security interview between an immigration officer and a passenger just disembarked from a plane. In one pair of scenarios the passenger had (purportedly) arrived for a wedding, and in the other the passenger had (purportedly) arrived on a business trip. To create the variants, we use the General Inquirer (GI) wordlist [5], which attaches to words a number of tags from a set of around 200, indicating features such as whether the word has positive or negative associations.

Scenario Construction From the characteristics described in [1], we chose to embed the following in the deceptive variant: *Fewer details in the dialogue*, measured in terms of number of propositions, referred to as DET; *Taking up a smaller proportion of the talking time*, measured as proportion of words spoken (PROPN); *Less verbal immediacy*, measured by number of instances of passive voice (greater in deceptive than default) as per the GI tags (PSV); *Repetition of words and phrases*, repeating in the deceptive variant a number of sequences of words, such as a verb phrase (REP); *More negative statements and complaints*, measured by proportions of words that were positive versus negative as per GI, and their relationship (POS, NEG); *Lack of cooperation*, manually adding statements to indicate an uncooperative attitude (COOP); *Fewer spontaneous corrections*, adding a number of these to the default (CORR); *Less likely to indicate remembering*, adding a number of indications of inaccurate memory (MEM); and *Fewer extreme descriptions*, as measured by GI tags QUAL, covering intensifiers like *very*, and

[*]The work was supported by ARC Discovery grant DP0558852.

Cite as: Deceptive Agents and Language (Extended Abstract), Mark Dras, Debbie Richards, Meredith Taylor and Mary Gardiner, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1539-1540

	Scenario 1		Scenario 2	
	default	deceptive	default	deceptive
DET	29	14	22	13
PROPN	76.4%	67.9%	75.7%	74.9%
PSV	0	3	0	2
REP	0	2	0	2
POS	5.54%	2.05%	7.13%	1.86%
NEG	0%	1.54%	0%	1.86%
COOP	0	2	0	2
CORR	2	0	2	0
MEM	1	0	1	0
EXTR	2.22%	0%	1.83%	0.47%

Table 1: Profiles for dialogues

PLEASUR, covering verbs like *excited* (EXTR). In Table 1, we give the specific values of these measures for each scenario dialogue variant. Percentages are proportions of words in the text; integers are absolute frequencies of occurrences.

Experimental Setup Our experiment involved 31 participants, and began with the participant reading the online information sheet. The information sheet contained a description of the purpose of the experiment, and an explanation of some typical factors that are found in deceptive language. These factors included all of those implemented above *except for* the last one (fewer extreme descriptions), along with one other not included in the study (more fidgeting). Each participant then received the default and deception versions of the same scenario. Versions and scenarios were presented in alternation to avoid order effects.

To determine whether the cues to deception were identified by the participant, after each scenario variant we gave the participant ten statements related to the cues and asked the participant to indicate the extent to which they felt that statement was true using a 5-point Likert scale; these are listed below. While responding to these statements, the participants also had showing on the screen as a reminder the advice from the information sheet about the typical factors indicating deception. The link from the statements to the implementation of the cues as described above will be clear; we also added gestures as a control, as we do not alter these between scenario variants. The statements are that the passenger: [Q1] included a lot of detail in their dialogue; [Q2] took up a large proportion of the time during the dialogue; [Q3] used a lot of gestures; [Q4] used a lot of verbal immediacy; [Q5] repeated words and phrases; [Q6] made negative statements and complaints; [Q7] was uncooperative; [Q8] made spontaneous corrections in their speech; [Q9] indicated that they were unable to remember something; [Q10] used extreme descriptions.

3. RESULTS AND DISCUSSION

The results are summarised in Table 2, presenting: *Exp*, the expected direction of change from the default scenario to the deceptive one (e.g. for Q1 regarding level of detail, we would expect participants to observe a decrease in comparing the default to the deceptive variant); *Def*, mean of the Likert scale scores for default; *Dec*, same for deceptive; δ, difference in Likert means; *s.d.*, standard deviation of differences; *sig*, significance.

The direction of responses, as measured both by the difference between mean Likert score scales and by the comparison question, is in general the expected one. This indicates that when given guidelines about what constitutes deceptive behaviour, people are able to identify an implementation of that behaviour enacted by a virtual agent where they can compare this with an implementation of non-deceptive behaviour. There are, in addition, a number of interesting aspects to the results related to specific cues.

	Exp	Def	Dec	δ	s.d.	sig
Q1	-	4.77	3.65	-1.13	1.157	$p < 0.0005$
Q2	-	4.74	3.74	-1.00	1.078	$p < 0.0005$
Q3	-	3.55	2.74	-0.81	1.203	$p < 0.0005$
Q4	-	4.00	3.39	-0.61	0.973	$p < 0.005$
Q5	+	2.32	3.19	+0.87	1.641	
Q6	+	1.81	4.26	+2.45	1.364	$p < 0.005$
Q7	+	1.35	3.32	+1.97	1.282	$p < 0.005$
Q8	-	3.90	2.13	-1.77	1.313	$p < 0.005$
Q9	-	3.32	2.06	-1.26	1.390	$p < 0.005$
Q10	-	3.10	2.94	-0.16	1.416	

Table 2: Summary of participant responses

Strong dialogue-level properties Dialogue-level properties — ones that apply to the dialogue as a whole, such as in Q1 and Q2 — are very well recognised, more so than the mixed responses for lexical-level properties. Presumably the participant does not have to be concentrating at the right moment, but keeps an overall impression of the dialogue.

Mixed lexical-level properties Verbal immediacy (Q4) had a strong response: this is somewhat surprising, as the example given is for passive voice, which is the kind of grammatical knowledge that the average participant is likely to be hazy on. In addition, recognition of negative statements and complaints (Q6) had the largest difference in mean Likert scale responses ($\delta = 2.45$ from Table 2), indicating the applicability of the language technology resources noted above. On the other hand, recognition of repeated words and phrases (Q5), which appeared to be a much easier task, was not achieved at rates better than chance. This is also in spite of the frequencies of occurrence for repeated words and for verbal immediacy being approximately the same (see Table 1).

'Placebo' effect There was a strong belief by the human judges about a difference in gesture usage (Q3) where in fact one did not exist, which is difficult to explain. A major difference with the text-based deception detection work described in [2] is that a comparison between a default variant and a deceptive variant was given, although it was not indicated which was which. It is likely therefore that the participants decided that one variant was deceptive and the other default (which was in fact the case), rather than that deceptive cues were distributed across the two variants. However, there is no uniform placebo effect: if there were, all cues would have had the expected responses, and this was not the case for Q5 and Q10.

4. REFERENCES

[1] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003.

[2] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29:665–675, 2003.

[3] M. Rehm and E. André. Catch me if you can: exploring lying agents in social settings. In *AAMAS '05*, pages 937–944, 2005.

[4] M. Stone and D. DeCarlo. Crafting the illusion of meaning: Template-based specification of embodied conversational behavior. In *CASA*, pages 11–16, 2003.

[5] P. Stone, D. Dunphy, M. Smith, and D. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.