Ascribing and Weighting Beliefs in Deceptive Information Exchanges

Valeria Carofiglio and Fiorella de Rosis
Department of Informatics, University of Bari
{carofiglio, derosis}@di.uniba.it
WWW home page:*http://aos2.uniba.it:8080/IntInt.html*

Cristiano Castelfranchi
Department of Communication Sciences, University of Siena
castel@www.ip.rm.cnr.it

Abstract. Humans apply a large variety of deception forms in their communicative acts; they are not necessarily 'uncooperative' in doing so, as they may even deceive for the benefit of their interlocutors [2, 3]. Deception may be active or passive, according to whether the Speaker does something or not, to achieve his goal. It may be applied directly to the deception object p or may indirectly influence it through some 'deception medium' q that may be a cause, an effect or a diverting cause or effect of p. In this short paper, we examine how deception may be simulated if mental states are represented as belief networks and various weights are attached to beliefs.
keywords: belief ascription, belief networks, dialog simulation, deception.

1 A Deception Simulator: MOUTH OF TRUTH

MOUTH OF TRUTH enables simulating the way that a deceptive communicative act may be planned. The System plays the role of the Speaker S (the potential 'deceiver') and the User the role of the Addressee A. The problem considered may be stated as follows.

Given: (i) a domain; (ii) a mental state of S, as a set of beliefs about the domain represented in a bayesian network OWN-S; (iii) a set T of domain facts that A is presumed to know and (iv) a domain fact p that is taken as the 'deception object', plan a communicative act that involves a deception medium q and enables S to achieve a given goal about A's belief in p: $\mathbf{G_S} \equiv (G_S B_A p)$ *or* $G_S(\neg BW_A p)$.

The User interacts with the System through a Web interface; she may select a domain and declare what she knows about it; S builds a image of the mental state of A (IM-S-A), consistently with his own mental state. This image is an 'overlay model' of OWN-S with a reduced number of nodes. The two models are probabilistically compatible, as the prior probabilities of the nodes they share are the same. We assume that, to avoid arising suspicion, S is very prudent in evoking facts that are not 'in focus' in A's mind: his communicative acts then try to keep as close as possible to the beliefs that he considers as 'active', to A, in the given phase of interaction. Obviously, the set of active beliefs is inherently dynamic: some of them may become inactive after some time, other may activate as far as the dialog goes on. S keeps a memory of the facts that were active 'not too much long ago', in order to insure the safety of his communicative acts. Once IM-S-A has been built, the truth values of the facts in T are propagated and the System looks at the probability that A assigns to p, in her present state of knowledge. It then selects a candidate deception medium q, by considering the following factors:

a. *efficacy* in achieving the goal $\mathbf{G_S}$, as *impact* of the deception medium q on the deception object p; two different notions of impact are implemented:
   - degree of achievement of S's goal about A's belief in p, as a function of the difference between the goal-value and the posterior probability of p, to A, after she comes to believe in q: $Imp^1(q,p) = 1 - |G_S B_A P(p) - B_S B_A P(p|q)|$;
   - degree of change in A's belief about p that is induced by believing in q, as a function of the difference between the prior and posterior probability of A's belief in p: $Imp^2(q,p) = B_S B_A P(p|q) - B_S B_A P(p)$;

b. *plausibility* of q to A, as a function of the likelihood that A will get convinced about q after S's communication. Again, we implemented two different notions of plausibility:
   - *local plausibility*, as a function of the difference between what S presumes to be A's belief about q and the communication she receives from S: $Plau^1_{S,A}(q) = 1 - |B_S B_A P(q) - B_S P(q)|$;
   - *global plausibility*, as a function of the same difference, extended to all facts q_l that are 'connected to q', in the context of 'active beliefs': $Plau^2_{S,A}(q) = 1 - 1/w(\sum_{l=1,\dots w} |B_S B_A P(q_l) - B_S P(q_l)|)$. In this case, the plausibility of a fact is seen as the overall compatibility of new knowledge induced by q with the previous one, in the mentioned context.

c. *safety* of the communicative act: this is linked to the risk, to S, of being discovered, by A, in a deception attempt. This risk depends, in its turn, on various factors: (i) how plausible might be, to A, that S believes in what he says and (ii) whether A may come to understand that her believing in that fact is in S's interest and advantage;

d. *reliability*, to A, of the information source eventually evoked, by S, to support his statement. The credibility of a fact q referred by an information source IS may be defined, probabilistically, as: $Cred(q, IS){=}P(q|Say(IS,q))$ (and alike for $\neg q$);it is a function of the reliability of IS: $Rel(IS,q){=}P(Say(IS,q)|q)$ and of the prior probability of q (and alike for $\neg q$). S may combine information sources in various ways to support his communication: he may cite several reliable sources or may combine reliable with unreliable ones, to try to confound A's ideas.

2 Related work

Although the purpose of our research is different, the methods we apply have several analogies with those employed in other dialog systems (see [5, 7]). Bayesian networks seem to us a very promising approach to deception modeling: from a logical viewpoint, the problem involves a very complex reasoning [1]: to Taylor and colleagues, modeling these dialogs would require reasoning at 'deeply-nested belief' level (higher that the third one) while, on the contrary, deeply-nested beliefs are known to be 'awkward and contrary to intuitions in human-human dialogs' [6]. The approach we are here describing enables avoiding deeply-nested belief reasoning through the following assumptions: (*i*) beyond intentionally insincere assertions, there is always some interest of the Speaker, the need to achieve some domain-goal; (*ii*) deception results from a conflict (between S and A) about this goal and (*iii*) S is interested to do his best to hide his deception attempts. From these assumptions comes that deceiving is not the same as lying, so that discovering contradictions in the interlocutor enlights only a subset of the large variety of deception attempts: this entails the need of tackling the problem from an 'uncertain' viewpoint and to attach, to beliefs, a system of weights formalizing the mentioned aspects. Besides representing the inherently uncertain process of 'suspicion', belief networks offer the possibility of applying various forms of reasoning (from causes to effects and vice-versa): this enables simulating the deception forms mentioned in the Abstract. The assumption of 'probabilistic compatibility' between OWN-S and IM-S-A reduces the range of deceptions we may simulate at present: S cannot exploit, in his deceptive attempts, differences in the strength that A attaches to belief relations; this is one of the aspects we plan to investigate in the future. In general, we wish to assess whether our deception theory is effective in enabling the generation of the most common deception forms that we describe in [4]).

References

1. Ballim, A., Wilks, Y.: Beliefs, stereotypes ad dynamic agent modeling. User Modeling and User-Adapted Interaction, 1, 1, 1991.
2. Castelfranchi, C., Poggi, I.: Lying as pretending to give information. Pretending to Communicate, H. Parret (Ed), Springer Verlag, 1993.
3. Castelfranchi, C., de Rosis, F., Grasso, F.: Deception and suspicion in medical interactions; towards the simulation of believable dialogues. Machine Conversations, Y Wilks (Ed), Kluwer Series in Engineering and Computer Science, **511**, 1999.
4. de Rosis, F., Castelfranchi, C., Carofiglio, V.: Can computers deliberately deceive? A simulation attempt of Turing's Imitation Game. Sumbitted for publication.
5. Ndiaye, A. and Jameson, A.: Predictive role taking in dialog: global anticipation feedback based on transmutability. In Proc. 5th Int. Conf. on User Modeling, Kailua-Kona, Hawaii,137-144, 1996
6. Taylor, J.A., Carletta, J., Mellish, C.: Requirements for belief models in Cooperative dialogue. User Modeling and User-Adapted Interaction, 6, 1, 1996.
7. Zukerman, I., Jinath, N., McConachy, R., George, S.: Recognising intentions from rejoinders in a bayesian interactive argumentation system. In PRICAI 2000 Proceedings, Melbourne. 8.