ELIZA, but cleverer: Designing persuasive artificial agents

Debora Field (debora@csc.liv.ac.uk) and **Floriana Grasso** (f.grasso@csc.liv.ac.uk)
Dept. of Computer Science, University of Liverpool, Liverpool L69 3BX, UK

Introduction

This extended abstract introduces research into persuasive dialogue modelling which is to be carried out as part of *PIPS*,[1] one of the five Integrated Projects funded from the first call of the EU 6th Framework Programme in the field of e-Health. One of PIPS' aims is to design and build a personal computational advisory system to support European Citizens in taking informed health-related decisions about lifestyle. PIPS' advice will be based on an extensive knowledge engineering effort, aimed at integrating many heterogeneous data sources.

The PIPS artificial advisory agent can be simplistically viewed as a kind of online health advisor whom a user might typically chat with when he needs practical, health-related advice concerning some particular change of lifestyle.

It is envisaged that the artificial advisory agent will be used largely for the purposes of preventative medicine. The agent will be able to offer individually personalised counselling on nutrition, exercise, and other health-related practices, such as stopping smoking, and practising hygiene in the kitchen.

Research in health promotion has long proved that persuasive skills are needed for such advisor–advisee scenarios, as they involve changing well rooted attitudes and behaviours (*e.g.*, (Prochaska et al., 1992, 1994, 1997)). Unlike ELIZA, then, the PIPS advisory agent will not only be extremely knowledgeable (thanks to its comprehensive knowledge base, medical ontologies, and knowledge extraction techniques), but it will be skilled in the art of persuasion. This paper concerns the 'art of persuasion' part of this undeniably large problem.

[1]PIPS: Personalised Information Platform for Health and Life Services - FP6/IST No. 507019.

Rationale

Preliminary efforts are being concentrated on designing a formal model for the **generation of utterances that motivate a person to change his/her behaviour**.

Consider a user who has been told by his doctor that he is at high risk of becoming a diabetic, and that he needs to change his eating habits—but his motivation to do so is very low, and his knowledge of which changes he would need to make is poor. In order to work out how to advise a person like this, an advisory agent would need to have an understanding of the user's actual situation (gender, weight, age, diet, food likes and dislikes), and would have to possess an opinion about his cognitive state (his beliefs, emotions, desires, fears, ...). The agent would also have to be skilled in recognising different motivational states, and in techniques of persuasion. The aim of this research is to build an artificial persuasive agent that has understanding and opinions like these, and knows how to use them to influence people's behaviour.

We intend to develop a formal model of how to persuade a person to change their behaviour through conversation. Taking the above considerations in mind, we believe that this model should be strongly based on a behavioural model, and a promising candidate is Prochaska et al.'s 'stages of change' model (1992; 1994; 1997). The key principle in the 'stages of change' model is that behaviour change is a process during which one's beliefs about and attitudes towards the behaviour change evolve. At any point during that process an individual may have one of a number of different attitudes towards the predicated change in behaviour: precontemplation, contemplation, preparation, action, maintenance, or relapse. The techniques one uses to advise a person should then vary,

according to which stage of change one believes that person is experiencing. Thus the dialogue model will be strongly psychologically motivated. However, this alone will not be sufficient: if are going to succeed in influencing behaviour through NL conversation, the need for persuasive NL dialogue will also have to be addressed.

A recent testimony to this comes from an evaluation of the STOP project (Reiter et al., 2003). STOP produced tailored letters to encourage people to stop smoking. The letters were founded on the 'stages of change' model, and were produced by means of a standard natural language generation system. The system was thoroughly evaluated with a clinical trial in which more than 2500 smokers participated. The evaluation, however, produced the counter-intuitive result that tailored messages were not significantly effective in helping people to stop smoking. One of the explanations for this negative result, we believe, is a lack of recognition that explicit persuasive techniques were needed for this problem.

To address this issue, our approach is to draw insights from theories coming from the philosophy of argument, rather than creating *ad hoc* strategies. For example, Gilbert's 'coalescent argumentation' theory (Gilbert, 1997, 2001; Gilbert et al., 2003), emphasises the importance of attitudes, beliefs, feelings, and intuitions in argument, in contrast to a diminished role for logic. Alternatively, the classical New Rhetoric theory (Perelman and Olbrechts-Tyteca, 1969) describes discoursive techniques which influence the persuasiveness of the presented arguments. Our approach is in line with recent trends of work in argumentation in a computational context (Grasso et al., 2004).

Strategies

The starting point for our formal model will be a pre-existing characterisation of rhetorical argumentation (Grasso, 2002). This work formalises the above mentioned New Rhetoric by means of the notion of *rhetorical schema*, defined as a 6-tuple:

$$R_S =< N, C, O_c, A_c, R_c, S_c > \quad (1)$$

where:

- N is the name of the schema.
- C is the claim the schema supports.
- O_c are the ontological constraints the schema is based on. They represent what has to be true in the world for the argument to be put forward. These can be mathematical/logical relations, or semantic relationships.
- A_c are the acceptability constraints. They represent what the audience should believe for the argument to work.
- R_c are the relevance constraints. They represent what the audience need in order to reach the conclusion. They can be defined in terms of "effort" to process the argument, by using notions such as salience or focus of attention.
- S_c are the sufficiency constraints. They embed the notion of "fair play" of the speaker when putting the argument forward, and represent all elements that should be known to the audience in order to decide whether to accept or reject the conclusion.

A number of argumentative strategies described in the New Rhetoric have been formalised using the schema above, and the characterisation has been successfully used to analyse the structure of argumentative discourse, appealing to a theory from philosophy in the same way as much research in discourse modelling does to theories from linguistics. The most preferred of such theories is the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). As pointed out by (Reed and Long, 1998), RST is, however, unsuitable for *generating* arguments, for several reasons. First, as also identified by (Moore and Pollack, 1992), RST does not adequately handle intentions, an important shortcoming for rhetorical argumentation.[2] Secondly, RST considers as exceptional the cases in which nuclearity breaks down to give place to parallel structures, structures which are the rule rather than the exception in argumentation. Finally, RST does not account for argumentative relations as such, nor for high level organisations such as Modus Ponens, or,

[2]This problem has been partially solved by (Marcu, 2000), by coupling RST with intentions.

most importantly, for structured combinations of higher level units.

A drawback of our characterisation, as it is at the moment, is that it does not propose any 'constructive' theory for evaluating the coherence of argumentative dialogues, as, for instance, RST does for general (especially explanatory) discourse. The theory does not, therefore, fully address the criticisms that RST attracts when applied to argumentative discourse. A more complete theory would capture how arguments can be built upon other arguments, in order to evaluate the quality of the whole argument, in addition to the coherence of the dialogue. We believe that such a theory can be built on grounds like those we propose, and we set ourselves this theoretical goal in the PIPS project.

Tactics

The development of a persuasive dialogue model is to be carried out by experimentation with and development of a computational utterance planner by (Field and Ramsay, 2004). This planner is particularly suitable for experimentation with persuasive dialogue, as it has explicit representation of the belief states of agents, and its mechanisms focus on the problem of how to get the hearer to respond in the way desired by the speaker. The input to the planner is specified as a standard two-state planning problem (Newell and Simon, 1963): the world as it is now, and the world as the planning agent would like it to be. These states are specified as belief states—the planner/speaker has beliefs about what others believe, and plans its utterances on the basis of those beliefs. Figure 1 shows a simplified English paraphrase of a task (expressed in the model in standard epistemic logic), and its solutions.[3]

The planner has only one linguistic action operator for all contexts. This single act contrasts strongly with a selection of different 'speech acts' (Austin, 1962; Searle, 1965) from which a speaker chooses one that suits his particular beliefs and intentions in a given situation. Using the single act, the planner can model problematic conversational situations, including felicitous and infelicitous instances of bluffing, lying, sarcasm, and stating the obvious.

Initial state	John has gone bird-watching with Sally. John likes Sally a lot, and is enjoying telling her about birds. He thinks Sally is new to bird-watching. He also thinks that she looks cold. Just now, a huge bird has landed in a nearby tree. John doesn't know what species the bird is.
Goal condition	John wants Sally to be impressed by him.
Solutions	To impress Sally, John first **thinks of offering Sally his coat to wear**, to warm her up, but then he decides to try and impress Sally with his bird expertise, and plans to say, **"There's a dodo!"**

Figure 1: Paraphrase of a task and its output

The linguistic act has a disjunctive precondition that can always be proved, yet whose verification by the planning agent has the significant side-effect of making the agent aware of something new. The linguistic act has no direct effects, but a side-effect: a new entry is recorded in the conversational minutes (a version of the 'conversational record' (Stalnaker, 1972; Lewis, 1979; Thomason, 1990)). In the model, conversants use Gricean maxims (Grice, 1975) as a standard they expect others to abide by, as well as more practical maxims which enable them to generate and recognise violations of Grice's Cooperative Principle. Planning is carried out purely from the planner's (John's) point of view, and John remains ignorant of the consequences that his utterances would have. However, the effects of John's utterances on his hearers are modelled to enable the study of "infelicities" (Austin, 1962, p. 14).[4]

The planner is of a 'reasoning-centred' design, in contrast to the popular 'search-centred' designs (Blum and Furst, 1995; Bonet and Geffner, 1998; Hoffmann and Nebel, 2001; Nguyen and

[3] A dodo is a very large flightless bird that is famously extinct.

[4] Figure 1 is in fact only half a task: the hearer's (Sally's) beliefs are omitted from the example, for the sake of clarity. Several of the model's tasks are designed around Figure 1. In each task, John has the same belief state, but Sally's belief state varies, and so for some tasks, John succeeds in impressing Sally, whereas for others, he does not.

Kambhampati, 2001). It can achieve *entailed* goals by *reasoning about* actions: state-space planning is interwoven with theorem proving in such a way that a theorem prover uses the effects of actions as hypotheses. For example, the planner can achieve blocks-world goals of the form 'above(X,Y)' by knowing that 'above' is the transitive closure of 'on', and using that knowledge to reason over the effects of simple 'stack' and 'unstack' STRIPS (Fikes and Nilsson, 1971) operators.

To enable this, a theorem prover for first-order logic is incorporated into a STRIPS-style planner based on foundational work in classical planning (Newell et al., 1957; Newell and Simon, 1963; Green, 1969; McCarthy and Hayes, 1969; Fikes and Nilsson, 1971). The search strategy is enhanced with two specially designed heuristics which enable it to overcome the (well-documented) difficulties presented by planning for conjunctive goals using a backwards-searching state-space planner. Its LAN ('look away now') heuristic enables the planner to, in certain highly restricted contexts, achieve a goal that it can prove true in the current state. The ThA ('think ahead') heuristic takes into account the preconditions of a chronologically later goal during planning for an earlier goal, by exploiting the information held in the antecedents of recursive rules.

The epistemic theorem prover embodies a constructive/intuitionist logic, it proves theorems by natural deduction, and it embodies a deduction theory of belief (Konolige, 1986). These design features are chosen in preference to classical logic, refutation, and the 'possible worlds' model of belief (Hintikka, 1962; Kripke, 1963) respectively, because it is considered that they enable human reasoning processes to be better represented.[5]

Practicalities

Practical matters to be addressed in order to develop the utterance planner into a model of persuasive dialogue include the following.

The general idea is to develop new 'persuasive' conversational maxims, bringing together ideas from the 'stages of change' model, and argumentation theory, by means of experimenting with the said utterance planner.

Currently the planner takes the formulation of the speaker's goal as a given—each problem already contains a speaker goal (to cheer Sally up, to impress Sally, to discourage Andy,...), but there is no modelling of the decision process which has led to this goal. A layer of higher-level goal planning therefore needs to be added.

A range of personalities, or 'virtual egos' is to be designed. During development of the model, these virtual egos will play the part of typical users of the system.

The planner currently plans and models the performance of a single utterance by a single agent. Structures will need to be added to enable two or more agents to enter into a dialogue, each keeping track of what he/she believes to be the discourse history thus far.

The planner currently plans the *semantics* of utterances, but eventually, surface form will also be taken into consideration. This will be with a view to enabling recognition of clues concerning a person's emotional state from the linguistic structures and lexical items they use during the dialogue, as well as enabling the generation of surface form.

Evaluation

In a project of this sort, evaluation is crucial for success. We are fortunate to be in a position where we can draw valuable insights from other partners in the PIPS project who have specific expertise in health, nutrition, hygiene, and counselling. In particular, public health professionals working on nutrition advice will help us build a corpus of actual records of advisor-advisee conversations. This will be used both as guidelines during development of the model, and as a benchmark for the preliminary evaluation of the model. The PIPS project also plans to produce a pilot system, to be deployed and evaluated in *San Raffaele* Hospital, in Italy, while two further demonstrators are planned for similar environments in Spain and China. It is therefore our hope that by the end of the project(in 2008), we will be able to perform a thorough evaluation of our theory, with real users from a wide variety of backgrounds.

*****

[5]Relevant issues include the meaning of the implication operator in human reasoning versus classical logic, the interest of human reasoners in propositional content rather than merely whether a proposition is true, and the problem of logical omniscience.

This speculative extended abstract has described an ambitious project which, it is hoped, will advance the frontiers of knowledge in the field of persuasive dialogue modelling in some small way at least. We hope to make good progress by starting small, with inspired personnel, and a robust implementation, and building carefully.

References

J. Allen, J. Hendler, and A. Tate. Editors. *Readings in planning*, 1990. San Mateo, California: Morgan Kaufmann.

J. L. Austin. *How to do things with words*, 1962. Oxford: Oxford University Press, 2nd edition.

A. L. Blum and M. L. Furst. Fast planning through planning graph analysis, 1995. In *Artificial Intelligence* 90: 281–300, 1997.

B. Bonet and H. Geffner. Hsp: Heuristic Search Planner, 1998. Entry at Artificial Intelligence Planning Systems (AIPS)-98 Planning Competition. *AI Magazine* 21(2), 2000.

P. R. Cohen, J. Morgan, and M. E. Pollack. Editors. *Intentions in communication*, 1990. Cambridge, Massachusetts: MIT.

E. A. Feigenbaum and J. Feldman. Editors. *Computers and thought*, 1995. Cambridge, Massachusetts: MIT Press. First published in 1963 by McGraw-Hill Book Company.

D. G. Field and A. Ramsay. Sarcasm, deception, and stating the obvious: Planning dialogue without speech acts, 2004. *AI Review*, to appear.

R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving, 1971. *Artificial Intelligence* 2: 189–208. Reprinted in Allen et al. (1990), pp. 88–97.

M. A. Gilbert. *Coalescent Argumentation*, 1997. New Jersey: Lawrence Erlbaum Associates.

M. A. Gilbert. Getting good value. Facts, values, and goals in computational linguistics, 2001. In *Proceedings of the International Conference on Computational Sciences–Part I*, Springer-Verlag, pp. 989–998.

M. A. Gilbert, F. Grasso, L. Groarke, C. Gurr, and J. M. Gerlofs. The persuasion machine: An exercise in argumentation and computational linguistics, 2003. In (Reed and Norman, 2003), Ch. 5, pp. 121–174.

F. Grasso. Towards computational rhetoric, 2002. *Informal Logic* 22(3): 225-259.

F. Grasso, C. Reed, and G. Carenini. Editors. *CMNA IV - Proceedings of the 4th workshop on Computational Models of Natural Argument*, 2004. ECAI 2004.

C. Green. Application of theorem proving to problem solving, 1969. In *Proc. 1st IJCAI*, pp. 219–39. Reprinted in Allen et al. (1990), pp. 67–87.

H. P. Grice. Logic and conversation, 1975. In P. Cole and J. Morgan. Editors. *Syntax and semantics, Vol. 3: Speech acts*, 1975, pp. 41–58. New York: Academic Press.

J. Hintikka. *Knowledge and belief: An introduction to the two notions*, 1962. New York: Cornell University Press.

J. Hoffmann and B. Nebel. The FF planning system: Fast plan generation through heuristic search, 2001. *Journal of Artificial Intelligence Research* 14: 253–302.

K. Konolige. *A deduction model of belief*, 1986. London: Pitman.

S. Kripke. Semantical considerations on modal logic, 1963. In *Acta Philosophica Fennica* 16: 83–94. Also in L. Linsky. Editor. *Reference and modality (Oxford readings in philosophy)*, 1971, pp. 63–72. London: Oxford University Press.

D. Lewis. Scorekeeping in a language game, 1979. *Journal of Philosophical Logic* 8: 339–59. Reprinted in D. Lewis *Philosophical papers Volume I*, 1983, pp. 233–249. New York and Oxford: Oxford University Press.

W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization, 1987. *Text* 8(3): 243–281.

D. Marcu. Extending a formal and computational model of rhetorical structure theory with intentional structures à la grosz and sidner. In *The 18th International Conference*

on Computational Linguistics COLING'2000, 2000.

J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence, 1969. *Machine Intelligence* 4: 463–502. Reprinted in Allen et al. (1990), pp. 393–435.

J. Moore and M. Pollack. A Problem for RST: The Need for Multi-Level Discourse Analysis. *Computational Linguistics*, 18(4):537–544, 1992.

A. Newell, J. C. Shaw, and H. A. Simon. Empirical explorations with the logic theory machine, 1957. In *Proceedings of the Western Joint Computer Conference*, 15: 218–239, 1957. Reprinted in Feigenbaum and Feldman (1995), pp. 109–133.

A. Newell and H. A. Simon. GPS, a program that simulates human thought, 1963. In (Feigenbaum and Feldman, 1995), pp. 279–93.

X. Nguyen and S. Kambhampati. Reviving partial order planning, 2001. In *Proc. IJCAI*, pp. 459–66.

C. Perelman and L. Olbrechts-Tyteca. *The New Rhetoric: a treatise on argumentation*, 1969. Notre Dame, Indiana: University of Notre Dame Press.

J. O. Prochaska, C. C. DiClemente, and J. C. Norcross. In search of how people change: Applications to addictive behaviors, 1992. *American Psychologist* 47(9): 1102–1114.

J. O. Prochaska, J. C. Norcross, and C. C. DiClemente. *Changing for good: A revolutionary six-stage program for overcoming bad habits and moving your life positively forward*, 1994. New York: Avon Books.

J. O. Prochaska, C. A. Redding, and K. E. Evers. The transtheoretical model and stages of change, 1997. In K. Glanz, F. M. Lewis, and B. K. Rimer. Editors. *Health behavior and health education: theory, research, and practice*, 2nd ed., pp. 60–84. San Francisco: Jossey-Bass Publishers.

C. A. Reed and T. J. Norman. Editors. *Argumentation Machines – New Frontiers in Argument and Computation*, 2003. Kluwer.

C.A. Reed and D.P. Long. Generating the structure of argument, 1998. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pp. 1091–1097.

E. Reiter, R. Robertson, and L. Osman. Lessons from a failure: Generating tailored smoking cessation letters, 2003. *Artificial Intelligence* 144: 41–58.

J. R. Searle. What is a speech act?, 1965. In M. Black. Editor. *Philosophy in America*, pp. 221–39. Allen and Unwin. Reprinted in J. R. Searle. Editor. *The philosophy of language*, 1991, pp. 39–53. Oxford: Oxford University Press.

R. Stalnaker. Pragmatics, 1972. In D. Davidson and G. Harman. Editors. *Semantics of natural language (Synthese Library, Vol. 40)*, pp. 380–97. Dordrecht, Holland: D. Reidel.

R. H. Thomason. Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics, 1990. In (Cohen et al., 1990), pp. 325–63.