E-motion: Moving Toward the Utilization of Artificial Emotion

MICHAEL A. GILBERT [1] *York University*
CHRIS REED *University of Dundee*

Abstract: During human-human interaction, emotion plays a vital role in structuring dialogue. Emotional content drives features such as topic shift, lexicalisation change and timing; it affects the delicate balance between goals related to the task at hand and those of social interaction; and it represents one type of feedback on the effect that utterances are having. These various facets are so central to most real-world interaction, that it is reasonable to suppose that emotion should also play an important role in human-computer interaction. To that end, techniques for detecting, modelling, and responding appropriately to emotion are explored, and an architecture for bringing these techniques together into a coherent system is presented.

Résumé: Les èmotions jouent un rôle vital dans la structuration des dialogues humains. Le contenu affectif influence, par exemple, les changements de sujet et de vocabulaire, et la présence ou le manque d'à-propos; agit sur l'équilibre délicat entre les buts reliés aux tâches immédiates et sociales; et représente un type de rétroaction sur l'effet des énoncés. Ces divers aspects sont si considérables dans l'interaction humaine qu'il est raisonnable de supposer que les émotions devraient aussi jouer un rôle important dans l'interaction entre les humains et les ordinateurs. À cette fin, on explore des techniques pour déceler, modeler et réagir convenablement aux émotions, et on présente une architecture pour coordonner ces techniques dans un sytème cohérent.

Keywords: atificial intelligence, discourse analysis, emotion, plan recognition, natural language understanding

1. Computers with emotional users

Emotion plays a central role in human communication, and especially in those communicative contexts in which there is disagreement. Dispute partners, consciously or not, stay finely tuned to each other's emotional reactions, constantly weighing their partner's reactions against possible outcomes, relationship maintenance, and goal achievement. The emotional clues observed in the normal course of events provide crucial information and guidance to the participants, and very often dictate subsequent moves within the dissensual context. Emotional cues provide such vital information as whether or not a line of argument is working well, giving offence, or otherwise creating damage rather than persuasion. A dispute partner's facial reactions, body movements, and other physical cues provide the feedback we need in order to direct the argumentative process. Like traffic signals, emotional cues as evidenced in visceral and outwardly emotional reac-

tions, guide us toward good routes and away from unsafe ones. Indeed, the frustrations we often feel when communicating from afar, or in contexts where we lack direct contact and, hence, direct feedback, are indicators of the importance of those cues. Even in email, our newest and overwhelmingly popular communicative form, aids like emoticons are intended to help in understanding the emotional intent of a message.

Our aim is to begin an exploration of how emotional cues might be processed in an human–computer situation. We are supposing that the natural cues available to human agents are not typically available to a computer, viz., those physical indicators we consciously or unconsciously process to guide our way. Human agents are most likely subliminally more aware of, say, changes in facial colouration than we imagine, thus providing us with crucial hints regarding such core phenomenon as attraction and anger. While there are ways that a machine can process some, or even a great deal, of this information, our interest lies elsewhere. Our focus is in the realm of language and the linguistic cues it provides (though we do include silence under the rubric of language.) This is important, since no one arena provides sufficient information to conclude with certainty that, say, an emotional response is forthcoming. Rather, we rely on a matrix of physical and linguistic cues to cross-verify each other. So the challenge we face is to try and locate those cues that occur within language that can aid in identifying an emotional state.

The underlying model on which we will primarily rely is that of Discourse Analysis, that area of Communication Theory that analyzes ordinary conversation. In that area one key concept concerns the idea that, given certain utterances, a particular range of replies are considered appropriate, and among those some are considered preferred. The classic example is a request which can be granted or refused. The preferred response [PR] is the granting of the request, while the dispreferred response [DPR] is the refusal of the request. In addition to the simple grant-refusal responses other possibilities for DPRs are available; these include responding with a question, changing the topic, or ignoring the interchange. Any DPR can be an indication of an emotional response which, unfortunately, leaves a very large category, but does provide us with an *uber*-concept that is a beginning.

Emotional reactions are invariably multi-faceted, and certainly must be if we are going to draw conclusions about them. Someone who blushes might be doing so because they have been caught in a fib, but they might also have had a free association that has resurrected some much earlier embarrassment. The package of cues we use is intertwined so as to avoid making errors based on one source of information.[2] These packages involve body, speech, and intellectual reactions. If the computer can identify a package that signals when there is emotional response, then it can respond accordingly.

Some of the sorts of indicators we use in a day-to-day way are listed below. These are things that are noticed, perhaps unconsciously, by listeners, though the skill level in identifying such factors can vary widely depending on the partici-

pants. Some people are very aware of the mood and states of others, while some are far more literal and do not heed subliminal messages (See, B. O'Keefe 1988) or, at least, require far more blatant cues. Since our concern is specifically with the arena of argumentation, we have tried to isolate a set of cues that *may* indicate an emotional response to a query or rejoinder. These are the indicators that *can possibly* create a suspicion that an interlocutor is responding to an argument in an emotional way. The list is by no means complete and is meant to serve as an exemplar with which we might begin to work.

Table 1. Indicators of possible emotional user reaction

1. Delay in response or Rapid response

   I.e., some deviation from normal response time. Delay in response is typically an indication of thoughtfulness.[3] This could mean that the argument was not begetting an emotional response, but showed up a weakness in User's position. However, such moments are often tinged with at least a minimum of emotion, and rapid response perhaps even more so.

2. Equivocation

   This is a sign of positional instability, of an inability or lack of desire to answer the question or rejoinder directly, and as such can indicate emotional increase.

3. Claim reiteration

   Repeating one's claim rather than providing an answer or new argument is a strong indicator that a position is in trouble. The accompanying cognitive dissonance will entail some significant emotion.

4. Lexical alteration [aggressive language]

   A change in the style of language being used, in particular, from standard to aggressive, is a very reliable sign of the presence of emotion.

5. Response avoidance [changing subject]

   Like claim reiteration, response avoidance can indicate difficulties for a position. It may mean that the respondent has no available answers, or that the agent's interjections are reaching their mark.

6. Textual metrics [sentence length, word choice]

   Sometimes the responses forthcoming might not avoid or become aggressive, but may still involve certain changes. In ordinary parlance, we may note, for example, that a respondent has become "cooler" or "distant." Insofar as an agent can identify such metrics as the type of language and its structure, so the agent can also identify changes therein.

Table 2. Likely emotional reactions

1. Defensiveness

   Defensiveness is a very common reaction in argumentation. When a respondent begins to lose ground and feels that her position is under attack, but cannot readily come up with adequate ripostes, a defensiveness indicated by aggression or annoyance is liable to appear.

2. Indignation

   This reaction may occur under a variety of circumstances, including the very idea that a position is being attacked or a premiss questioned.

3. Frustration

   When a position is seen as very clear to a user, but the respondent continues to disagree, frustration may occur, as it may, when the inability to defend a position becomes acute.

4. Anger

   Anger frequently is a reaction to the necessity for forced change. The systemic reluctance users have to the alteration of belief systems often results in anger prior to the necessary adjustment.

5. Regret

   When a user does begin to alter a view, regret or reluctance might occur either as indicators of the desire not to change, or as a sign that the abandoned view ought not have been held.

6. Guilt

   Like regret, guilt may occur when a position is abandoned and the respondent believes that, in truth, he ought not have held it, or when the user is maintaining a position he knows is fatally flawed.

7. Enthusiasm

   Not all emotions are negative. A user may respond with enthusiasm or excitement when a strong point is made, or when an insight has occurred.

When we consider argument, we need to consider the core motivations for entering into the process in the first place. While some arguments may be undertaken for their own sake or in order to determine the truth of a position or the best course of action, most seem to involve fairly personal aims. That is, people enter into arguments when they want to attain something, get someone to act in a certain manner, or otherwise achieve one goal or another. One or more of the emotions mentioned above, then, frequently enters when a goal is challenged, undermined or, for that matter, achieved. Goals, broadly speaking, determine a range of actions or plans that are construed to be means of satisfying those goals. Blockage of goals, disruption in plans, unexpected turns, realizations that goals are different

than what was thought may lead to emotional reaction. And, since many users are only vaguely aware of their goals or, at least, have not fully considered them, such hiccups are quite common.

Goals fall into different categories. The broadest goals we have we may call "motives," and they are the ones that define our broadest behavioural parameters (Dillard 1990, Craig 1986, Gilbert 2001). It is one's broad motives that preclude theft as a means of achieving the goal of acquiring a new auto. The next category, and closest to the commonsense notion of goal, is the "task goal." These are goals that involve achieving a particular result such as obtaining the aforementioned auto, or getting to work on time, or obtaining tickets to the ballet. The third set, often neglected in Argumentation Theory, are "face goals," and they concern the relationships existing between the several participants in the interaction. The maintenance of that relationship has a great effect on how the argumentation will proceed, and must not be ignored. Thae satisfaction of these three sorts of goals define a set of options that may be used, and the chosen approach may be labeled the "apparent strategic goal," or ASG. It is the ASG that one typically cites in answer to a question such as, "What are you trying to do?" or, "What do you want?"

All of the above will, in the arena on which we are focused, determine a further set of "discourse goals" that will, allowing for the introduction of a set of procedures, translate into specific text. These discourse goals correspond closely to those that drive the process of language generation in Artificial Intelligence systems (Reiter & Dale 2001). The relationships between these types of goals are illustrated in Figure 1.

There are views of argumentation that seem to delimit an arena in which emotion plays little or no role. (See Gilbert 1995, 1995a.) In reality, however, emotions play a significant role in interpersonal argumentation, and, leaving aside the question of whether that is good or bad, a computer system interacting with a user in more than a very basic way must take that into account. Emotion explains why a user maintains a position (Damasio 1994), refuses to give it up, and frequently provides essential insight into understanding a respondent's goals and position.[4] If a machine is to do more than accept answers to questions, if it is to go beyond a simple branching tree model of argumentation, then it must be able to deal on the playing field that real users inhabit, and that is one where emotions play a considerable role. Real argumentation is *always* more than straightforwardly logical; it is rhetorical as well. Argument aims to persuade not only by appeals to pure dialectical reasoning, but to sympathy, concern, fear, joy, self-interest, and a myriad of other singularly human needs and wants.

So goals define positions, and emotions arise around both goals and their consequent positions. In those situations where a user is interacting with a machine in a persuasive context, we need the machine to be able to identify when an emotional reaction has occurred so that the machine can respond accordingly. Conse-

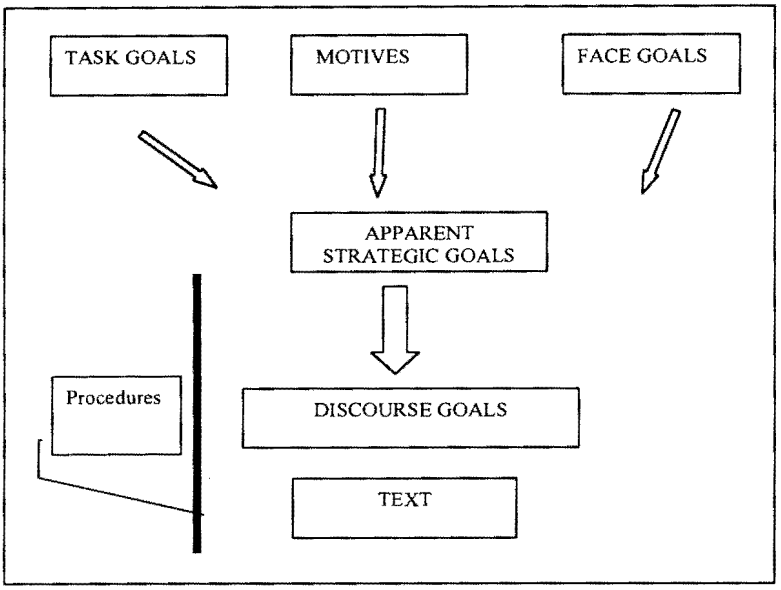

Figure 1. Goals

quently, there are two crucial questions. First, how can a machine identify an emotional reaction? Second, how can the machine respond to that reaction *qua* an emotional reaction? We suggested above in Table 1 the sorts of cues that a computer system might pick up on. The following section explores the way in which the suggested input might evolve into a system's conclusion that an emotional reaction has occurred.

2. Computer reaction to user emotion

Human respondents in an argument regularly notice emotional reactions to statements, non-verbal communications, argument rebuttals, and so on.[5] That does not mean, however, that the respondent's assessment will always be correct. Mistakes can be made regarding emotional messages and cues as easily as with logical or verbal messages (Gilbert 2002). Consequently, in what follows we must take an approach that involves the concept of *hypothesizing an emotional reaction on the user's behalf.* That is, when there are grounds for suspecting a user is responding emotionally, the system will take certain steps, but will not take the assumption to be fact.

When confronted with a (possible) emotional reaction, a respondent has to decide how to react. To a great extent, the reaction will depend on a number of variables, including the confidence the respondent has in the inference that an emotional reaction has occurred, the relationship between the respondent and protagonist, and the subject under discussion. In fact, there are an enormous range of

possibilities that can occur in human-to-human interactions, and we must, of course, greatly simplify them for the purposes of a human–computer model. Consequently, we suggest three core courses of action a system might take when confronted with an emotional reaction.

Table 3. Computer reactions to emotion

[1] Ignore, proceed as before
[2] Acknowledge directly
[3] Acknowledge indirectly

Let us examine each of these reactions in turn.

In case [1] the system has for a variety of reasons (explored below) determined that there has been an emotional reaction, but chooses to ignore it. Previous questions may be reiterated, previous processes reviewed, or the next question may be asked with a flag to return later. The emotional reaction may, for example, be to something the system deems as peripheral to the core discussion, so rather than pursue it, it may be wise to choose a different tack. If, however, there are continued emotional reactions, then the system can always come back to the issue.

In case [2] the system notes that an emotional reaction has occurred, and inquires of the user if this is correct. The system might ask why there has been an emotional reaction, or speculate in some way or other as to its significance. The point is that the system deals directly by putting, as it were, the emotional cards on the table.

In case [3] the system notes that an emotional reaction has occurred, and changes strategy accordingly in order to investigate the degree, nature, or source of the reaction. In this case, the system is assuming, for example, that the reaction holds a clue to the underlying grounds for the user's position.

If we assume that emotional responses, or at least, negative emotional responses, are signaled by DPRs, (dispreferred responses) then one possible sequence of system reactions, albeit the most elementary, is exemplified by the following.

The role of emotion for the system is to trigger ways of finding information about the intertwined categories of goals and values. When a user reacts strongly —emotionally—to something, the reaction can indicate a goal conflict, goal blockage, frustration, or other ways of goals being disturbed, denied or denigrated. Such situations are indicative of at least some degree of dissonance and, consequently, potential opportunities for change. Consequently, the system should, at least, note the occurrence; but it might, under a variety of circumstances, go further. The decision to use method [1], [2] or [3] will be a function of how much information about the goals and values the system software has, how much it

thinks it can get, and how much it thinks it needs. Emotional reaction also provides a way for the system to test hypotheses regarding goals and values, which can be done with methods [2] or [3]. Becoming (negatively) emotional in an argument indicates frustration—often, perhaps always, because one's goals are being blocked. The system can then try to use the information acquired about goals to move the argument further toward agreement.

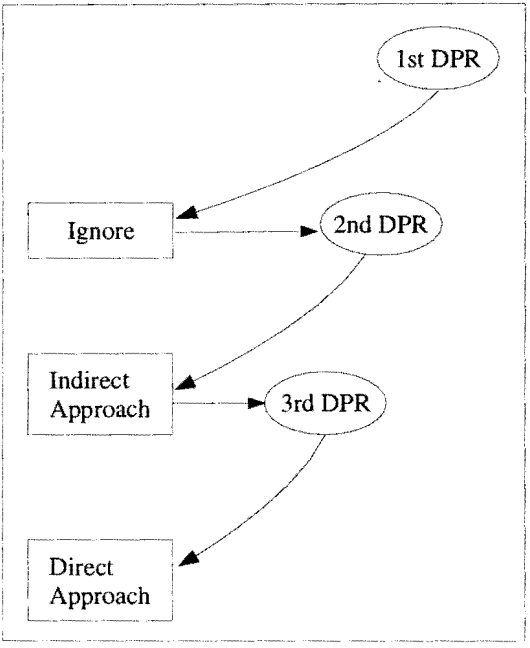

Figure 2. Elementary System Response

It makes sense to suppose that a computer is going to tread quite lightly when dealing with an emotional human. There is, or are, after all, one or more goals and objectives that the system has, and diving head first (or chip first, if you prefer) into an emotional morass will achieve little. On the other hand, emotions frequently do not lead to any sort of morass at all, and the resulting information, goal assessments, and alternatives can be extremely helpful. That is why it would be recommended that the machine begin by minimizing emotional reaction through, say, first ignoring the DPR, and then, if it persists, using an indirect approach. In other words, we see the moves outlined in Figure 2 as moving along a continuum of potential emotional increase, and believe that a computer ought not easily go where humans fear to tread.

We want now to turn to an example. In this illustration, we are imagining a machine designed to persuade a user to alter some behaviour not considered ben-

eficial to the user. A very abstract model of such a machine was developed in Gilbert *et al.* (forthcoming). The idea is that a user would choose to interact with the computer system on a subject of her or his choosing. Such subjects might include cigarette smoking, seatbelt use, diet quality, birth control, and other social *cum* personal issues. The following example concerns seatbelt use. Note that dispreferred responses are indicated as DPR*n*, where '*n*' is the sequence number of the response. In addition, 'delay +*n*' means that there has been an *n* second delay longer than the user's norm. The system is indicated as 'S' and the user in this example as 'U1.' The system reaction to each DPR is indicated immediately below the utterance.

Example 1 Seatbelts

1. S: Why won't you use seatbelts?
2. U1: I want to be able to get to my children.
3. S: But if you're hurt, then you could not help them.
4. U1: [delay +2] I'm a very careful driver. [DPR1]

   Ignore
5. S: Not everyone on the road is. Someone could run into you.
6. U1: I don't like seatbelts. [DPR2]

   Direct
7. S: I know that. Why?
8. U1: None of your business. [DPR3]

   Ignore
9. S: But surely, you do many things you don't want to. Right?
10. U1: *[U1 does not respond.]*

    Direct
11. S: Do you want to stop now?
12. U1: [delay +4] No.

    Indirect
13. S: It's unpleasant having to do things we don't like, isn't it?

The first thing to notice is that such competence in natural language understanding is currently beyond the capabilities of even the very best artificial intelligence systems, such as the end-to-end (i.e., speech-input to speech-output) TRAINS system (Allen *et al.* 1995). Natural language understanding, however, is not the focus here. What is interesting is that assessment of emotional content can be performed to a large extent independently of the semantic content.

There are three DPRs in this excerpt, one each at lines 4, 6, and 8. In 4, the User does not react to the actual statement of the system, but instead attempts to

shore up her defense. There is thus a slight shift in topic. On its own, such minor change might not be grounds for supposing that a response has been emotionally charged. When accompanied by an unusual pause, however, it might reasonably be concluded that a DPR has occurred at 4.

Rather than respond to the DPR explicitly and point up the avoidance, the system follows up on it, incorporating the comment into the system's argument. 4 is accepted as an hypothesis, and the argument continued on its basis in 5. 6 results in the second DPR, which demonstrates a clear failure of formal relevance, and to this the system responds directly with a reiteration of the initial premiss to the discussion. 8 is an easily identified emotional response, because, given the previous system utterance, it does not even fit into the right category of an answer to the specific question of 7. The system ignores the statement itself, i.e., the system does not argue about whose business it is, but rather goes to a deeper level of argument on which User's objection might rest. User does not respond in 10, and since this is the second consecutive strong emotional response, the system responds directly to the emotion and inquires if the interaction should be stopped. After a moment's hesitation, User agrees to continue, which is, of course, a very good sign.

The aim here is not to construct from scratch a natural language understander and responder—that is the focus of no small part of artificial intelligence. Instead, the aim is to arrange extant AI techniques and provide the algorithmic glue in order to explore how to deal with the emotional structure of text and dialogue.

In the same way that AI has concentrated upon syntactic, semantic and pragmatic coherence in natural language understanding and responding, so we hope to concentrate upon 'emotional coherence.' In 1950, Alan Turing described a test (now, the "Turing Test") by which artificial imitation of intelligence is tested against true human intelligence by means of computer-mediated linguistic interaction (Turing, 1950). If the imitation is indistinguishable from the human, then, Turing's argument runs, the imitation should itself be classified as intelligent. In the AI community, the Turing Test still drives research into syntactic, semantic and pragmatic linguistic interaction. In the same way that the Turing Test can be made specific to these aspects: "Does the imitation produce syntax as well as a human?", etc. so too can it be applied to emotional coherence: "Does the imitation produce responses that are emotionally as appropriate and coherent as a human would produce?" Furthermore, tackling the full Turing Test necessitates handling the emotional components. Forever ignoring dispreferred responses, or trampling over clear emotional clues from an interlocutor, would at best render an artificial intelligence impoverished (as, for example, an autistic intelligence), and at worst, a dialogue participant so frustrating that human interlocutors would swiftly terminate interaction.

A dispreferred response to an utterance represents a conflict between the goal of the utterer and the goal evinced by the response. That is, the primary means of

recognizing a DPR is to reason about the goals manifest in the respondent's utterance. From an AI perspective, this requires integrating three aspects: the goals and plans of the system; the goals and plans of the human at a large scale; and how the human's most recent utterance fits in to her or his wider goals and plans. This plan recognition process has been demonstrated to be extremely tough (Carberry and Pope, 1993). In specific—and typically task-oriented—dialogues, limited plan recognition has been reasonably successful (Allen *et al.* 1995), but more generic techniques are still elusive. In particular, recent plan recognition work has focused upon identifying discourse goals (that is, goals associated with quite specific perlocutionary effects), rather than on inferring ASGs. Though plan recognition may ultimately have an important role to play, this section focuses on other, simpler, means of detecting DPRs, whilst laying out a framework that would admit plan recognition components as that technology matures.

We work here on the evidence that an utterance that is loaded with emotion may manifest a constellation of emotion indicators, such as those listed in Table 1 (Planalp 1998). This constellation effect can be exploited by arranging simple detection modules in a weighted network. Each module is triggered by a specific feature of the user's utterance (delay, lexicalization choice, sentence length, etc.), and then the pattern of units that fire for a given utterance can be used to determine both the presence and extent of emotional content being expressed. Figure 3 shows this arrangement.

Figure 3 shows just three triggers, response rate deviation, lexicalization deviation and response avoidance, though others from Table 1 and those described in the literature could be added to this list. Each trigger can avail itself not only of data from the current user utterance, but also of a range of other information represented in the system, including current contextual information (such as norms for this interaction, topical shifts, etc.) and more generic background information (such as attitudes and beliefs of the user). In addition, the previous system utterance may also effect the decision making within each trigger. Consider, for example, the response avoidance trigger. A representation of the current topic (such as is provided in Reed 1999, for example) would suffice to detect a change in focus, but only when combined with information about the structure of the system's previous utterance can response avoidance be inferred: the phrasing of the system's utterance as a yes/no question would expect to elicit a yes or no response, for example.

Once a trigger module has detected a particular clue to an emotional DPR, it fires, indicating a match. The firings of the different modules are weighted differently (indicated by the width of the arrows in Figure 3). Response avoidance, for example, is a substantially more reliable cue than lexical deviation, and as such, the firing of the response avoidance module counts for more when deciding whether or not a DPR has occurred.

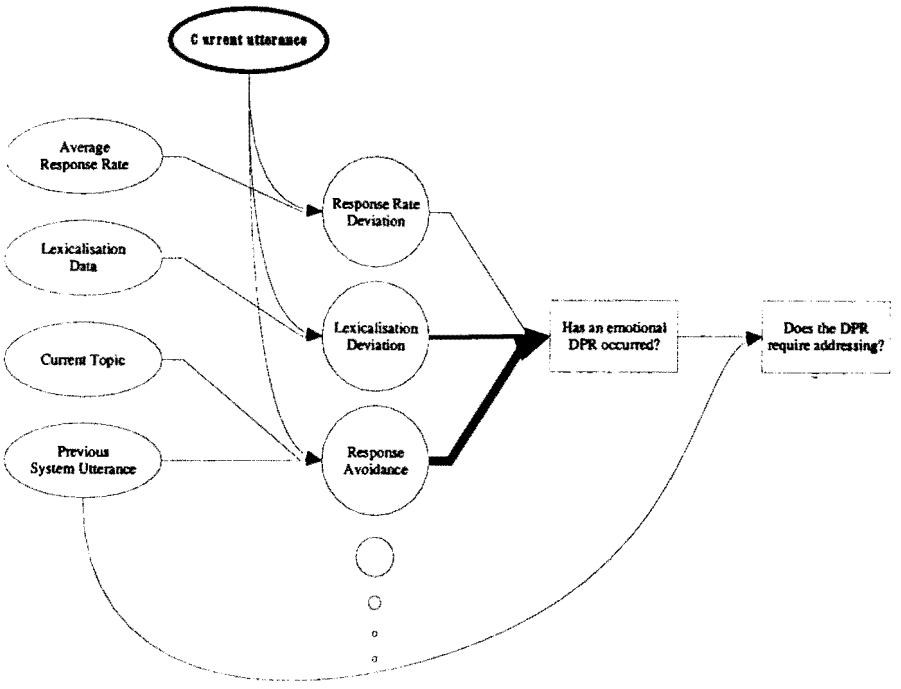

Figure 3. Emotional trigger network

By aggregating the weighted firings, the system can determine if a DPR has occurred, and also the degree of that DPR. In a final step, this information can be combined with the emotional status of the system's previous utterance to determine if something should be done about the DPR. An emotional question from the system is liable to elicit an emotional response; the key is to locate *unusually strong* emotional responses.

The picture is further enriched by the cumulative effect of DPRs. By including the results of previous DPR analysis, i.e., by noting that some previous user utterances involved emotional DPRs, or even "near-DPRs," it is possible to aggregate successive ratings. This can be achieved by a simple decay function, such that a recent near-DPR increases the likelihood of a fully triggered DPR. That is, the model can account for and then detect emotional escalation in a dialogue.

The rich literature on emotion in verbal and non-verbal communication (see Anderson and Guerrero 1998 for one overview) could be integrated in a similar fashion. Although the system here focuses, for simplicity, upon written (i.e., typed)

text, the broad range of linguistic, nonlinguistic and paralinguistic emotion indicators in spoken language could each be detected by a simple, well-defined modular unit. These units could then be included in a large network and combined with those responsible for text features.

With a battery of triggers looking for clues that a DPR has occurred, the aggregated result then forms the basis of the response mechanism introduced in Figure 3, above.

We would like to present a second example to specifically indicate the use of triggers and how they might be used to both identify and handle DPRs. In the following illustration, as in the previous one, the computer is attempting to persuade the user to adopt a mode of behaviour that is beneficial to his well-being (in this case wearing a motorcycle helmet.)

Example 2. Helmets

1. U2: Man, helmets are a drag.
2. S: But they're required by law. Aren't they?
3. U2: So what? I mean who cares? [DPR1]
4. S: But they do protect you.
5. U2: Hey, you ride, you take chances. [DPR2]
6. S: But you do take some precautions. You look before you cross a street, right?
7. U2: [delay +3] Yeah, so what?
8. S: So sometimes you don't take chances, right?
9. U2: So I'm not stupid, so what?
10. S: So you take the chance of spending the rest of your life in a wheelchair, right?
11. U2: [delay –2] Hey! [delay =10] [DPR3]
12. S: Not too attractive is it?
13. U2: [delay +6] That only happens to idiots. [DPR4]
14. S: It can't happen to you?
15. U2: I'm a careful rider.
16. S: But wearing a helmet is being careful.
17. U2: You don't get it, man. You ride because it's exciting, because it's a rush. You wear a helmet and it ruins it. [DPR5]
18. S: Riding a wheelchair for the rest of your life wouldn't be very exciting, would it?
19. U2: [delay +6] No, it wouldn't.
20. S: Something to think about, eh?

The system's utterance at 2 sets up a yes/no question. The fact that 3 does not exhibit a response to that question fires the response-avoidance trigger, which, as modelled in Figure 4, has a strong influence on (that is, it is a reliable indicator of) emotional response. The system concludes that a DPR has occurred. The final step is to determine an appropriate response. In this case, as this is the first DPR encountered, the system decides to ignore the DPR and press on.

At 5, the User verges on irrelevance—the combination of the established current topic, and the User's shift from it, lead to the relevance trigger firing. With only weak support for concluding that a DPR has occurred, this borderline DPR is then also ignored. 7 involves a delay, which could cause the response rate deviation trigger to fire. However, the weak support offered for a DPR by response rate deviation, coupled with the format of a preferred response (that is, of directly answering a yes/no question) leads to a conclusion that 7 does not constitute a DPR.

11 then demonstrates a clear DPR, with response rate deviation, lexicalization deviation, and response avoidance. In addition the 'near-miss' DPR of 7 is still recent in the discourse history. Note that the [Delay +10] gives the system a clear field to answer freely, i.e., User2 is not responding, suggesting that User2 is "at a loss." With this clear evidence of an emotional escalation, the system decides at 12 to respond directly to the emotional content. Since the idea put forward at 10 engendered a strong response, the system will pursue the emotion-inducing comment by what is essentially a reiteration.

13 again represents a DPR, as evinced in particular by the lexical choice. (In fact, 13 might also be analysed as a restatement of DPR2 at 5, though such an analysis may place too great a burden on the assumptions of the system's natural language competence). Though somewhat mitigated by the system's direct, emotional response at the previous step, it is still, with the response rate deviation, classed as a DPR. Then, at 14, the system offers an indirect response to DPR4, in an attempt to de-escalate the emotional content broached directly at 12.

At 17, the response avoidance (the system's previous utterance is opening an argument about being careful) and lexicalization deviation suggest a DPR. The cumulative effect of several DPRs in recent discourse history leads the system to adopt a gambit of direct, emotional response, at 18, and it turns out that the gambit is successful.

Of course, it is also possible that at 19 U2 would say something like, "Well the hell with this," and just storm away. But it must be kept in mind that the effects of a persuasive interaction may not materialize directly at the time. They can simmer and emerge at a later time or in a different context. In any case, this example illustrates how the triggers discussed can be watched to determine how the argument is proceeding. By careful monitoring of the emotional status of the interaction, using a battery of relatively simple techniques, it becomes possible to develop a much more sophisticated computational model.

It warrants repeating that the attempt to incorporate emotional content and interaction into communication between human and computer systems is a very difficult and complex one. The ability to model the role of emotion in argumentation in such a way as to create a sufficiently regular pattern suitable for systematization calls for a deeper understanding of the place of emotion in human to human interaction than we likely have. Simplifying is, therefore, a necessity, and it is hoped that the degree of simplification has not obviated the point of the exercise.

Our aim has been to demonstrate the central role played by emotion in determining how to contribute to dialogic argumentation. The focus in many computational models has been upon the fulfillment of specific task goals, but the two examples given here show that these goals represent only one constraint on the progression of the dialogue. The monitoring and maintenance of the emotional component of the dialogue has an equally substantial impact on the direction and success of a dialogue.

We have presented a model for the automated detection of emotional DPRs, based on a network of modules encapsulating single trigger responses, the outputs from which are weighted, collected together and used as input to a time-cumulative thresholding function. This function gives a binary response to the question, "Does the system need to attend to a DPR?"

By answering this question at each turn, a computational dialogue participant can be aware of the emotional structure of the dialogue, and take that into account in forming responses. This holds the potential for avoiding irretrievable conflict and break-down, and for improving the chances that a given dialogue succeeds in meeting the aims of it participants.

Notes

[1] Michael A. Gilbert would like to acknowledge the support of the Canadian Social Sciences and Humanities Research Council, Grant # 410-2000-1340.

[2] It must be remembered that mistakes regarding the imputation of emotional states can always be made by both machines and human agents. There is no avoiding this.

[3] Of course, it may also be an indication that the user has taken a bite of an apple, but we beg a large *ceteris paribus*.

[4] We will not provide a defense of the role of emotion here. Detailed arguments can be found in Gilbert 1995, 2001, 2001a.

[5] The inability to notice, or appropriately handle, emotional response is often considered a clinical condition, e.g., autism.

References

Allen, James F., Schubert, Lenhart K., *et al.* 1995. "The TRAINS Project: A case study in building a conversational planning agent," *Journal of Experimental and Theoretical AI*, 7: 7-48.

Anderson, P.A. and Guerrero, L.K. (Eds.). 1998. "Principles of Communication and Emotion in Social Interaction" in Anderson and Guerrero 1998b, 49-89.

—. 1998b. *Handbook of Communication and Emotion.* London: Academic Press.

Carberry, Sandra and Pope, W. Alan. 1993. "Plan Recognition Strategies for Language Understanding," *International Journal of Man-Machine Studies*, 39: 529-577.

Craig, Robert. 1986. "Goals In Discourse," in D.G. Ellis and W.A. Donohue, (Eds.), *Contemporary Issues In Language and Discourse Processes,* pp. 257-273. Hillsdale, NJ: Erlbaum.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason and the Human Brain.* New York: Avon Books.

Dillard, James P. 1990. "The Nature and Substance of Goals in Tactical Communications" in M.J. Cody and M.L. McLaughlin, (Eds.), *The Psychology of Tactical Communication,* pp. 69-90. Avon, UK: Multilingual Matters Ltd.

Gilbert, Michael A. 1995. "Emotional Argumentation, or, Why Do Argumentation Theorists Argue with their Mates?" *Analysis and Evaluation: Proceedings of the Third ISSA Conference on Argumentation, Vol. II,* F.H. van Eemeren, R. Grootendorst, J.A. Blair, and C.A. Willard, (Eds.). Amsterdam: Sic Sat.

—. 1995a. "The Delimitation of 'Argument'," *Inquiry,* 15: 1: 63-75.

—. 2001. "Getting Good Value: Facts, Values, and Goals In Computational Linguistics," *Proceedings of the International Conference on Computational Science*, San Francisco, May 2001.

—. 2002. "Effing the Ineffable: The Logocentric Fallacy in Argumentation," *Argumentation,* 16:1: 21-32.

Gilbert, M.A., F. Grasso, L. Groarke, C. Gurr and J.M. Gerlofs. (Eds.) (Forthcoming). "The Persuasion Machine: An Exercise in Argumentation and Computational Linguistics." *Argumentation.*

O'Keefe, Barbara J. 1988. "The Logic of Message Design: Differences in Reasoning About Communication," *Communication Monographs,* 55: 80-103.

Planalp, S. 1998. "Communicating Emotion in Everyday Life: Cues, Channels and Processes" in Anderson and Guerrero 1998b, 30-45.

Reed, Chris A. 1999. "The Role of Saliency in Generating Natural Language Arguments" in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, 876-881. Stockholm: Morgan Kaufmann.

Reiter, Ehud & Dale, Robert. 2000. *Building Natural Language Generation Systems.* Cambridge: Cambridge University Press.

Turing, Alan M. 1950. "Computing Machinery and Intelligence," *Mind* 59: 433-460.

Michael A. Gilbert
Department of Philosophy
York University
4700 Keele Street
Toronto, Ontario
CANADA M3J 1P3

Gilbert@YorkU.ca

Chris Reed
Applied Computing
University of Dundee
Dundee
SCOTLAND DD1 4HN

chris@computing.dundee.ac.uk