An Argumentation Agent Models Evaluative Criteria

John Debenham

Abstract An approach to argumentation attempts to model the partner's evaluative criteria, and by attempting to work with it rather than against it. To this end, the utterances generated aim to influence the partner to believe what we believe to be in his best interests — although it may not be in fact. The utterances aim to convey what is so, and not to point out "where the partner is wrong". This behaviour is intended to lead to the development of lasting relationships between agents.

1 Introduction

This paper is based in rhetorical argumentation [1] and is in the area labelled: *information-based agency* [3]. An information-based agent has an identity, values, needs, plans and strategies all of which are expressed using a fixed ontology in probabilistic logic for internal representation. All of the forgoing is represented in the agent's deliberative machinery. [2] describes a rhetorical argumentation framework that supports argumentative negotiation. It does this by taking into account: the relative information gain of a new utterance and the relative semantic distance between an utterance and the dialogue history. Then [4] considered the effect that argumentative dialogues have on the on-going *relationship* between a pair of negotiating agents.

This paper is written from the point of view of an agent α that is engaged in argumentative interaction with agent β. The history of all argumentative exchanges is the agents' *relationship*. We assume that their utterances, u, can be organised into distinct dialogues, Ψ^t. We assume that α and β are negotiating with the mutual aim of signing a contract, where the contract will be an instantiation of the mutually-understood object $o(\Psi^t)$. An argumentation agent has to perform two key functions: to understand incoming utterances and to generate responses.

John Debenham
University of Technology, Sydney, Australia, e-mail: debenham@it.uts.edu.au

2 Assessing a Contract

No matter what interaction strategy an agent uses, and no matter whether the communication language is that of simple bargaining or rich argumentation, a negotiation agent will have to decide whether or not to sign each contract on the table. An agent's preferences may be uncertain. In which case, we ask the question: "how certain am I that $\delta = (\phi, \varphi)$ is a good contract to sign?" — under realistic conditions this may be easy to estimate. $\mathbb{P}^t(\text{sign}(\alpha, \beta, \chi, \delta))$ estimates the certainty, expressed as a probability, that α should sign proposal δ in satisfaction of her need χ, where in (ϕ, φ) ϕ is α's commitment and φ is β's. α will accept δ if: $\mathbb{P}^t(\text{sign}(\alpha, \beta, \chi, \delta)) > c$, for some level of certainty c.

To estimate $\mathbb{P}^t(\text{sign}(\alpha, \beta, \chi, \delta))$, α will be concerned about what will occur if contract δ is signed. If agent α receives a commitment from β, α will be interested in any variation between β's commitment, φ, and what is actually observed, as the enactment, φ'. We denote the relationship between commitment and enactment:

$$\mathbb{P}^t(\text{Observe}(\alpha, \varphi') | \text{Commit}(\beta, \alpha, \varphi))$$

simply as $\mathbb{P}^t(\varphi' | \varphi) \in \mathscr{M}^t$, and now α has to estimate her belief in the acceptability of each possible outcome $\delta' = (\phi', \varphi')$. Let $\mathbb{P}^t(\text{acc}(\alpha, \chi, \delta'))$ denote α's estimate of her belief that the outcome δ' will be acceptable in satisfaction of her need χ, then we have:

$$\mathbb{P}^t(\text{sign}(\alpha, \beta, \chi, \delta)) = f(\mathbb{P}^t(\delta' | \delta), \mathbb{P}^t(\text{acc}(\alpha, \chi, \delta'))) \quad (1)$$

for some function f; if f is the arithmetic product then this expression is mathematical expectation. f may be more sensitive; for example, it may be defined to ensure that no contract is signed if there is a significant probability for a catastrophic outcome.

There is no prescriptive way in which α should define $\mathbb{P}^t(\text{acc}(\alpha, \chi, \delta'))$; the following three components at least will be required. $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta'))$ represents α's belief that enactment δ' will satisfy her need χ. $\mathbb{P}^t(\text{obj}(\delta'))$ represents α's belief that δ' is a fair deal against the open marketplace — it represents α's *objective* valuation. $\mathbb{P}^t(\text{sub}(\alpha, \chi, \delta'))$ represents α's belief that δ' is acceptable in her own terms taking account of her ability to meet her commitment ϕ [2] [3], and any way in which δ' has value to her personally — it represents α's *subjective* valuation. That is:

$$\mathbb{P}^t(\text{acc}(\alpha, \chi, \delta')) = g(\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta')), \mathbb{P}^t(\text{obj}(\delta')), \mathbb{P}^t(\text{sub}(\alpha, \chi, \delta'))) \quad (2)$$

for some function g.

Suppose that an agent is able to estimate: $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta'))$, $\mathbb{P}^t(\text{obj}(\delta'))$ and $\mathbb{P}^t(\text{sub}(\alpha, \chi, \delta'))$. The specification of the aggregating g function will then be a strictly subjective decision. A highly cautious agent may choose to define:

$$\mathbb{P}^t(\text{acc}(\alpha,\chi,\delta')) = \begin{cases} 1 & \text{if: } \mathbb{P}^t(\text{satisfy}(\alpha,\chi,\delta')) > \eta_1 \\ & \wedge\ \mathbb{P}^t(\text{obj}(\delta')) > \eta_2 \ \wedge\ \mathbb{P}^t(\text{sub}(\alpha,\chi,\delta')) > \eta_3 \\ 0 & \text{otherwise.} \end{cases}$$

for some threshold constants η_i.

First β must give meaning to $\mathbb{P}^t(\text{satisfy}(\beta,\chi,\delta))$ by defining suitable criteria and the way that the belief should be aggregated across those criteria. Suppose the information acquisition process is managed by a plan π. Let random variable X represent $\mathbb{P}^t(\text{ease-of-use}(\beta,\delta) = e_i)$ where the e_i are values from an evaluation space that could be $\mathscr{E}$ ={fantastic, acceptable, just OK, shocking}. Then given a sequence s that was supposed to achieve task τ, suppose that β's tame human rates s as evidence for ease-of-use as $e \in \mathscr{E}$ with probability z. Suppose that β attaches a weighting $\mathbb{R}^t(\pi,\tau,s)$ to s, $0 < \mathbb{R} < 1$, which is β's estimate of the *significance* of the observation of sequence s within plan π as an indicator of the true value of X. For example, the on the basis of the observation alone β might rate ease-of-use as $e =$ acceptable with probability $z = 0.8$, and separately give a weighting of $\mathbb{R}^t(\pi,\tau,s) = 0.9$ to the sequence s as an indicator of ease-of-use. For an information-based agent each plan π has associated *update functions*, $J_\pi(\cdot)$, such that $J^X_\pi(s)$ is a set of linear constraints on the posterior distribution for X. In this example, the posterior value of 'acceptable' would simply be constrained to 0.8.

Denote the prior distribution $\mathbb{P}^t(X)$ by $\mathbf{p}$, and let $\mathbf{p}_{(s)}$ be the distribution with minimum relative entropy with respect to $\mathbf{p}$: $\mathbf{p}_{(s)} = \arg\min_{\mathbf{r}} \sum_j r_j \log \frac{r_j}{p_j}$ that satisfies the constraints $J^X_s(s)$. Then let $\mathbf{q}_{(s)}$ be the distribution:

$$\mathbf{q}_{(s)} = \mathbb{R}^t(\pi,\tau,s) \times \mathbf{p}_{(s)} + (1 - \mathbb{R}^t(\pi,\tau,s)) \times \mathbf{p} \tag{3}$$

and then let:

$$\mathbb{P}^t(X_{(s)}) = \begin{cases} \mathbf{q}_{(s)} & \text{if } \mathbf{q}_{(s)} \text{ is more interesting than } \mathbf{p} \\ \mathbf{p} & \text{otherwise} \end{cases} \tag{4}$$

A general measure of whether $\mathbf{q}_{(s)}$ is more interesting than $\mathbf{p}$ is: $\mathbb{K}(\mathbf{q}_{(s)} \| \mathbb{D}(X)) > \mathbb{K}(\mathbf{p} \| \mathbb{D}(X))$, where $\mathbb{K}(\mathbf{x} \| \mathbf{y}) = \sum_j x_j \log \frac{x_j}{y_j}$ is the Kullback-Leibler distance between two probability distributions $\mathbf{x}$ and $\mathbf{y}$, and $\mathbb{D}(X)$ is the expected distribution in the absence of any observations — $\mathbb{D}(X)$ could be the maximum entropy distribution. Finally, $\mathbb{P}^{t+1}(X) = \mathbb{P}^t(X_{(s)})$. This procedure deals with integrity decay, and with two probabilities: first, the probability z in the rating of the sequence s that was intended to achieve τ, and second β's weighting $\mathbb{R}^t(\pi,\tau,s)$ of the significance of τ as an indicator of the true value of X. Equation 4 is intended to prevent weak information from decreasing the certainty of $\mathbb{P}^{t+1}(X)$. For example if the current distribution is $(0.1, 0.7, 0.1, 0.1)$, indicating an "acceptable" rating, then weak evidence $\mathbb{P}(X =$ acceptable$) = 0.25$ is discarded.

3 Modelling the Argumentation Partner

In this Section we consider how the agent models its partner's contract acceptance logic in an argumentative context. In Section 2 we discussed modelling contract acceptance, but there is much more to be done.

Estimating β's evaluative criteria. α's world model, $\mathscr{M}^t$, contains probability distributions that model the agent's belief in the world, including the state of β. In particular, for every criterion $c \in \mathscr{C}$ α associates a random variable C with probability mass function $\mathbb{P}^t(C = e_i)$.

The distributions that relate object to criteria may be learned from prior experience. If $\mathbb{P}^t(C = e|O = o)$ is the prior distribution for criteria C over an evaluation space given that the object is o, then given evidence from a completed negotiation with object o we use the standard update procedure described in Section 2. For example, given evidence that α believes with probability p that $C = e_i$ in a negotiation with object o then $\mathbb{P}^{t+1}(C = e|O = o)$ is the result of applying the constraint $\mathbb{P}(C = e_i|O = o) = p$ with minimum relative entropy inference as described previously, where the result of the process is protected by Equation 4 to ensure that weak evidence does not override prior estimates. In the absence of evidence of the form described above, the distributions, $\mathbb{P}^t(C = e|O = o)$, should gradually tend to ignorance. If a decay-limit distribution [2] is known they should tend to it otherwise they should tend to the maximum entropy distribution.

In a multiagent system, this approach can be strengthened in repeated negotiations by including the agent's identity, $\mathbb{P}^t(C = e|(O = o, Agent = \beta))$ and exploiting a similarity measure across the ontology. Two methods for propagating estimates across the world model by exploiting the $\mathrm{Sim}(\cdot)$ measure are described in [2]. An extension of the $\mathrm{Sim}(\cdot)$ measure to sets of concepts is straightforward, we will note it as $\mathrm{Sim}*(\cdot)$.

Disposition: shaping the stance. Agent β's *disposition* is the underlying rationale that he has for a dialogue. α will be concerned with the confidence in α's beliefs of β's disposition as this will affect the certainty with which α believes she knows β's key criteria. Gauging disposition in human discourse is not easy, but is certainly not impossible. We form expectations about what will be said next; when those expectations are challenged we may well believe that there is a shift in the rationale.

α's model of β's *disposition* is $D_C = \mathbb{P}^t(C = e|O = o)$ for *every* criterion in the ontology, where o is the object of the negotiation. α's confidence in β's disposition is the confidence he has in these distributions. Given a negotiation object o, confidence will be aggregated from $\mathbb{H}(C = e|O = o)$ for *every* criterion in the ontology.

4 Strategies

This section describes the components of an argumentation strategy starting with tools for valuing information revelation that are used to model the fairness of a negotiation dialogue.

Information Revelation: computing counter proposals. Everything that an agent communicates gives away information. *Illocutionary categories* and an *ontology* together form a framework in which the value of information exchanged can be categorised. The LOGIC framework for argumentative negotiation [4] is based on five illocutionary categories: Legitimacy of the arguments, Options i.e. deals that are acceptable, Goals i.e. motivation for the negotiation, Independence i.e: outside options, and Commitments that the agent has including its assets. In general, α has a set of illocutionary categories $\mathscr{Y}$ and a categorising function $\kappa : \mathscr{L} \rightarrow \mathscr{P}(\mathscr{Y})$. The power set, $\mathscr{P}(\mathscr{Y})$, is required as some utterances belong to multiple categories. For example, in the LOGIC framework the utterance "I will not pay more for a bottle of Beaujolais than the price that John charges" is categorised as both Option (what I will accept) and Independence (what I will do if this negotiation fails).

Then two central concepts describe relationships and dialogues between a pair of agents. These are *intimacy* — degree of closeness, and *balance* — degree of fairness. In this general model, the *intimacy* of α's relationship with β, A^t, measures the amount that α knows about β's private information and is represented as real numeric values over $\mathscr{G} = \mathscr{Y} \times V$.

Suppose α receives utterance u from β and that category $y \in \kappa(u)$. For any concept $x \in V$, define $\Delta(u,x) = \max_{x' \in concepts(u)} \mathrm{Sim}(x',x)$. Denote the value of A_i^t in position (y,x) by $A_{(y,x)}^t$ then:

$$A_{(y,x)}^t = \rho \times A_{(y,x)}^{t-1} + (1-\rho) \times \mathbb{I}(u) \times \Delta(u,x)$$

for any x, where ρ is the discount rate, and $\mathbb{I}(u)$ is the *information*[1] in u. The *balance* of α's relationship with β_i, B^t, is the element by element numeric difference of A^t and α's estimate of β's intimacy on α.

Given the needs model, υ, α's *relationship model* (Relate$(\cdot)$) determines the target *intimacy*, A_i^{*t}, and target *balance*, B_i^{*t}, for each agent i in the known set of agents *Agents*. That is, $\{(A_i^{*t}, B_{*i}^t)\}_{i=1}^{|Agents|} = \mathrm{Relate}(\upsilon, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ where, $\mathbf{X}_i$ is the trust model, $\mathbf{Y}_i$ is the honour model and $\mathbf{Z}_i$ is the reliability model as described in [2]. As noted before, the values for intimacy and balance are not simple numbers but are structured sets of values over $\mathscr{Y} \times V$.

When a need fires α first selects an agent β_i to negotiate with — the social model of trust, honour and reliability provide input to this decision, i.e. $\beta_i = \mathrm{Select}(\chi, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$. We assume that in her social model, α has medium-term intentions for the state of the relationship that she desires with each of the available agents — these intentions are represented as the target intimacy, A_i^{*t}, and target balance, B_i^{*t}, for each agent β_i. These medium-term intentions are then distilled into short-term targets for the intimacy, A_i^{**t}, and balance, B_i^{**t}, to be achieved in the current dialogue Ψ^t, i.e. $(A_i^{**t}, B_i^{**t}) = \mathrm{Set}(\chi, A_i^{*t}, B_i^{*t})$. In particular, if the balance

[1] Information is measured in the Shannon sense, if at time t, α receives an utterance u that may alter this world model then the (Shannon) *information* in u with respect to the distributions in $\mathscr{M}^t$ is: $\mathbb{I}(u) = \mathbb{H}(\mathscr{M}^t) - \mathbb{H}(\mathscr{M}^{t+1})$.

target, B_i^{**t}, is grossly exceeded by β failing to co-operate then it becomes a trigger for α to terminate the negotiation.

Computing arguments For an information-based agent, an incoming utterance is only of interest if it reduces the uncertainty (entropy) of the world model in some way. Information-based argumentation is particularly interested in the effect that an argumentative utterance has in the world model including β's disposition, and α's estimate of β's assessment of current proposals in terms of its criteria.

If u requests α to perform a task then u may modify β's disposition i.e. the set of conditional estimates of the form: $\mathbb{P}^t(C = e|O = o))$. If β rates and comments on the demonstration of a sequence then this affects α's estimate of β's likelihood to accept a contract as described in Equation 1 (this is concerned with *how* β will apply his criteria).

Suppose that u rates and comments on the performance of a sequence then that sequence will have been demonstrated in response to a request to perform a task. Given a task, τ, and a object, s, α may have estimates for $P^t(C = e|(O = o, \mathscr{T} = \tau))$ — if so then this suggests a link between the task and a set of one or more criteria C_u. The effect that u has on β's criteria (what ever they are) will be conveyed as the rating. We assume that for every criterion and object pair (C, o) α has a supply of positive argumentative statements $\mathscr{L}_{(C,o)}$. Suppose α wishes to counter the negatively rated u with a positively rated u'. Let Ψ_u be the set of all arguments exchanged between α and β prior to u in the dialogue. Let $M_u \subseteq \mathscr{L}_{(C,o)}$ for any $C \in C_\mu$. Let $N_u \subseteq M_u$ such that $\forall x \in N_u$ and $\forall u' \in \Psi_u$, $\text{Sim}*(concepts(x), concepts(u')) > \eta$ for some constant η. So N_u is a set of arguments all of which (a) have a positive effect on at least one criterion associated with the negative u, and (b) are at 'some distance' (determined by r) from arguments already exchanged. Then:

$$u' = \begin{cases} \arg\min_{u' \in N_u} \text{Sim}*(concepts(u), concepts(u')) & \text{if } N_u \neq \emptyset \\ \arg\min_{u' \in M_u} \text{Sim}*(concepts(u), concepts(u')) & \text{otherwise.} \end{cases}$$

So using only 'fresh' arguments, α prefers to choose a counter argument to u that is semantically close to u, and if that is not possible she chooses an argument that has some general positive effect on the criteria and may not have been used previously.

References

1. Rahwan, I., Ramchurn, S., Jennings, N., McBurney, P., Parsons, S., Sonenberg, E.: Argumentation-based negotiation. Knowledge Engineering Review **18**(4), 343–375 (2003)
2. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In: P. Stone, G. Weiss (eds.) Proceedings 5th International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2006, pp. 1225 – 1232. ACM Press, New York, Hakodate, Japan (2006)
3. Sierra, C., Debenham, J.: Information-based agency. In: Proceedings 12th International Joint Conference on Artificial Intelligence IJCAI-07, pp. 1513–1518. Hyderabad, India (2007)
4. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: Proceedings 6th International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2007, pp. 1026–1033. Honolulu, Hawai'i (2007)