Information-Based Argumentation

Carles Sierra[1] and John Debenham[2]

[1] Institut d'Investigació en Intel·ligència Artificial – IIIA,
Spanish Scientific Research Council, CSIC
08193 Bellaterra, Catalonia, Spain sierra@iiia.csic.es
[2] University of Technology, Sydney, Australia debenham@it.uts.edu.au

Abstract. Information-based argumentation aims to model the partner's reasoning apparatus to the extent that an agent can work with it to achieve outcomes that are mutually satisfactory and lay the foundation for continued interaction and perhaps lasting business relationships. Information-based agents take observations at face value, qualify them with a belief probability and build models solely on the basis of messages received. Using augmentative dialogue that describes *what* is *good* or *bad* about proposals, these agents observe such statements and aim to model the way their partners react, and then to generate dialogue that works in harmony with their partner's reasoning.

1 Introduction

This paper is in the area labelled: *information-based agency* [1]. 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 and in an illocutionary language [2] for communication. All of the forgoing is represented in the agent's deliberative machinery.

In line with our "Information Principle" [2], an information-based agent makes no *a priori* assumptions about the states of the world or the other agents in it — represented in a world model inferred from the messages that it receives. These agents build up their models by comparing expectation with observation — in this way we have constructed general models of trust, honour and reliability in a single framework [1].

[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 [3] considered the effect that argumentative dialogues have on the on-going *relationship* between a pair of negotiating agents. Neither of these contributions addressed the relationship between argumentative utterances or strategies for argumentation. In this paper we adress these two issues.

The basis of our approach differs from [4] who builds on the notion of one argument "attacking" another. With the exception of a logical 'attack', whether one argument attacks another or not will depend on the receiving agent's private circumstances that are unlikely to be fully articulated. Thus, the notion of attack is of little use to information-based agents that build their models on the contents of utterances. This paper considers how to *counter* the effect of the partner agent's arguments, and aims to lead a negotiation towards some desired outcome by persuasive argumentation.

This paper is based in rhetorical argumentation [5]. For example, suppose I am shopping for a new car and have cited "suitability for a family" as a criterion. The salesman says "This LandMonster is great value," and I reply "My grandmother could not climb into that." Classical argumentation may attempt to refute the matriarch's lack of gymnastic prowess or the car's inaccessibility. Taking a less confrontational and more constructively persuasive view we might note that this statement impacts negatively on the "suitability for a family" criterion, and attempt to counter that impact possibly with "It's been voted No 1 for children." Although a smarter response may look for an argument that is semantically closer: "The car's height ensures a very comfortable ride over rough terrain that is popular with old people."

Information-based agents build their world models using an expectation/observation framework; this includes a model of the negotiation partner's behaviour. Agents form an *a priori* expectation of the significance of every event that occurs, and when the effect of that event is finally observed they revise their expectations. The model of behaviour includes measures of: trust, honour, reliability, intimacy, balance and disposition — *disposition* attempts to model what the partner means which may not be what they say. These measures are summarised using: temporal criteria, the structure of the ontology, and the illocutionary category of observed utterances.

Our argumentation agent has to perform two key functions: to understand incoming utterances and to generate responses. The approach is founded on a model of contract acceptance that is described in Section 2. Section 3 details a scenario that provides the context for the discussion. Sections 4 and 5 consider the scenario from each side of the bargaining table. Reactive and proactive argumentation strategies are given in Section 6, and Section 7 concludes.

2 Contract Acceptance

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. We will argue in Section 4 that the buyer will be uncertain of his preferences in our Scenario described in Section 3. If an agent's preferences are uncertain then it may not make sense to link the agent's criterion for contract acceptance to a strategy that aims to optimise its utility. Instead, we pose the more general 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[3] 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

[3] A richer formulation is $\mathbb{P}^t(\text{eval}(\alpha, \beta, \chi, \delta) = e_i)$ where $\text{eval}(\cdot)$ is a function whose range is some descriptive evaluation space containing terms such as "unattractive in the long term".

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 $\mathbb{P}^t(\varphi'|\varphi) \in \mathcal{M}^t$, 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:

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

for some function f;[4] 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 $\mathbb{P}^t(\text{acc}(\alpha, \chi, \delta'))$, it is a matter for applied artificial intelligence to capture the essence of what matters in the application. In any real application 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. $\mathbb{P}^t(\text{obj}(\alpha, \delta'))$ 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] [1], and any way in which δ' has value to her personally[5] — it represents α's *subjective* valuation. That is:

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

for some function g.

Suppose that an agent is able to estimate: $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta'))$, $\mathbb{P}^t(\text{obj}(\alpha, \delta'))$ 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:

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

for some threshold constants η_i. Whereas an agent that was prepared to permit some propagation of confidence from one factor to compensate another could define:

$$\mathbb{P}^t(\text{acc}(\alpha, \chi, \delta')) = \mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta'))^{\eta_1} \times \mathbb{P}^t(\text{obj}(\alpha, \delta'))^{\eta_2} \times \mathbb{P}^t(\text{sub}(\alpha, \chi, \delta'))^{\eta_3}$$

where the η_i balance the influence of each factor.

The point of this is: if an agent aims to produce persuasive argumentative dialogue then in the absence of any specific information concerning the structure of g

---

[4] β influences the equation in the sense that different βs yield different $\mathbb{P}^t(\delta'|\delta)$.

[5] For example, when buying a new digital camera, α may give a high subjective valuation to a camera that uses the same memory cards as her existing camera.

the agent should ignore g and concentrate on the three categories: $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta'))$, $\mathbb{P}^t(\text{obj}(\alpha, \delta'))$ and $\mathbb{P}^t(\text{sub}(\alpha, \chi, \delta'))$.

So how then will α specify: $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta))$, $\mathbb{P}^t(\text{sub}(\alpha, \chi, \delta))$ and $\mathbb{P}^t(\text{obj}(\alpha, \delta))$? Of these three factors only $\mathbb{P}^t(\text{obj}(\alpha, \delta))$ has a clear meaning, but it may only be estimated if there is sufficient market data available. In the case of selling sardines this may well be so, but in the case of Google launching a take-over bid for Microsoft it will not[6]. Concerning $\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta))$ and $\mathbb{P}^t(\text{sub}(\alpha, \chi, \delta))$ we assume that an agent will somehow assess each of these as some combination of the confidence levels across a set of privately-known *criteria*. For example, if I am buying a camera then I may be prepared to define:

$$\mathbb{P}^t(\text{satisfy}(\alpha, \chi, \delta)) = h(\mathbb{P}^t(\text{easy-to-use}(\alpha, \delta)), \mathbb{P}^t(\text{well-built}(\alpha, \delta))) \quad (3)$$

for some function h. Any attempt to model another agent's h function will be as difficult as modelling g above. *But*, it is perfectly reasonable to suggest that by observing my argumentative dialogue an agent could form a view as to which of these two criteria above was more important.

This paper considers how an agent may observe the argumentative dialogue with the aim of modelling, within each of the three basic factors, the partner's criteria and the relative importance of those criteria. In repeated dealings between two agents, this model may be strengthened when the objects of the successive negotiations are semantically close but not necessarily identical.

3 The Scenario

Rhetorical argumentation is freed from the rigour of classical argumentation and descriptions of it can take the form of "this is how it works here" and "this is how it works there" without describing a formal basis. We attempt to improve on this level of vagary by using a general scenario and describing the behaviour of our agents within it.

In a general retail scenario there is a seller agent, α, and a buyer, β. The items for sale are abstracted from: digital cameras, mobile phones, PDAs, smart video recorders, computer software, sewing machines and kitchen mixers. The features of an item are those that are typically listed on the last few pages of an instruction booklet. For example, a camera's features could include the various shutter speeds that it is capable of, the various aperture settings, the number of years of warranty, and so on — together the *features* describe the capabilities of the item. For the purpose of comparison with other items, β will consider a particular item as a typed Boolean vector over the (possible) features of each item available, this vector shows which feature is present. The *state* of an item is then specified by identifying which of the item's features are 'on'. For example, the state of a camera could be: 'ready' with aperture set to 'f8' and shutter speed set to '1 500'th of a second'. In this scenario an *offer* is a pair (supply of a particular item, supply of some money) being α's and β's commitments respectively.

β may wish to know how well an item performs certain tasks. Software agents are not naturally endowed with the range of sensory and motor functions to enable such

[6] In this example the subjective valuation will be highly complex.

an evaluation. We imagine that the seller agent has an associated tame human who will demonstrate how the various items perform particular tasks on request, but performs no other function. We also imagine that the buyer agent has an associated tame human who can observe what is demonstrated, articulates an evaluation of it that is passed to its own agent, but performs no other function.

To simplify our set up we assume that the seller, α, is β's only source of information about what tasks each item can perform, and, as we describe below, what sequence of actions are necessary to make an item perform certain tasks[7]. That is, our multiagent system consists only of $\{\alpha, \beta\}$, and the buyer is denied access to product reviews, but *does* have access to market pricing data. This restriction simplifies the interactions and focusses the discussion on the argumentation.

For example, if the item is a camera the buyer may wish to observe how to set the camera's states so that it may be used for 'point-and-shoot' photography. If the item is a sewing machine she may wish to see how to make a button hole on a piece of cloth. If the item is graphics software she may wish to see how to draw a polygon with a two-pixel red line and to colour the polygon's interior blue. These tasks will be achieved by enacting a process that causes the item to pass though a sequence of states that will be explained to β by α. So far our model consists of: features, states, sequences and tasks.

We assume that the object of the negotiation is clear where the object is an uninstantiated statement of what both agents jointly understand as the intended outcome — e.g. I wish to exchange a quantity of eggs of certain quality for cash. We assume that each agent is negotiating with the aim of satisfying some goal or need that is private knowledge. In determining whether a negotiation outcome is acceptable in satisfaction of a need we assume that an agent will blend the factors in our acceptance model described in Section 2. We assume that for each factor an agent will articulate a set of *criteria* that together determine whether the factor is acceptable. The criteria may include private information such as deadlines.

More formally, there is a set of feature names, $\mathcal{F}$, a set of item names, $\mathcal{I}$, a feature mapping: $\mathrm{feature} : \mathcal{I} \to \times^n(\mathbb{B} : \mathcal{F})$ where there are n feature names, and $\mathbb{B}$ is a boolean variable that may be $\top$ or $\bot$. Each item name belongs to a unique concept — e.g.: "Nikon123 is-a camera". For any particular item name, ν, $\mathrm{feature}(\nu)$ will be a typed Boolean vector indicating which features that item ν possesses. Let $\mathcal{F}_\nu$ be the set of n_ν features that item ν possesses. At any particular time t, the state of an item is a mapping: $\mathrm{state}^t : \mathcal{I} \to \times^{n_\nu}(\mathbb{B} : \mathcal{F}_\nu)$ where the value $\top$ denotes that the corresponding feature of that item is 'on'. A *sequence* is an ordered set of states, $(\boldsymbol{w}_i)$, where successive states differ in one feature only being on and off. A sequence is normally seen as performing a *task* that are linked by the mapping: to-do : $\mathcal{T} \to 2^{\mathcal{S}}$ where $\mathcal{T}$ is the set of tasks and $\mathcal{S}$ the set of all possible sequences — that is, there many be several sequences that perform a task. If a sequence is *performed* on an item then, with the assistance of a human, the agent rates how well it believes the sequence performs the associated task. The evaluation space, $\mathcal{E}$, could be $\{$good, OK, bad$\}$. A criterion is a predicate: criterion(ν), meaning that the item ν satisfies criterion 'criterion'. The set of criteria is

[7] In other words, the sort of information that is normally available in the item's Instruction Booklet — we assume that α conveys this information accurately.

$\mathcal{C}$. The argumentation requirements include (where $x \in V$, $c \in \mathcal{C}$, $v = \text{feature}(x)$, $y \in \mathcal{T}$, $z \in \mathcal{S}$, and $r \in \mathcal{R}$):

- "I need an x"
- "What sort of x do you need?"
- "I need an x that satisfies criterion c"
- "What features does x have?"
- "x has features v"
- "How do you make x do y"
- "The sequence z performed on x does y"
- "Perform sequence z on x"
- "If sequence z is performed on x then how would you rate that?"
- "I rate the sequence z as performed on x as r"

4 The Buyer Assesses A Contract

In this Section we consider how the buyer might use the general framework in Section 2 to assess a contract[8]. In general an agent will be concerned about the enactment of any contract signed as described in Equation 1. In the scenario described in Section 3, enactment is not an issue, and so we focus on Equation 2. To simplify things we ignore the subjective valuation factor. Before addressing the remaining two factors we argue that the buyer will not necessarily be preference aware.

Consider a human agent with a need for a new camera who goes to a trusted camera shop. If the agent is preference aware he will be able to place the twenty to fifty cameras on offer in order of preference. If is reasonable to suggest that a normal, intelligent human agent could not achieve this with any certainty, nor could he with confidence represent his uncertainty in his preferences as a probability distribution over his preferences. This lack of awareness of preferences may be partially due to lack of information about each camera. But, what could "perfect information" realistically mean in this example? Even if the purchaser could borrow all the cameras for a day and had access to trusted, skilled users of each camera even then we suggest that our human agent would still be unable to specify a preference order with confidence. The underlying reason being the size and complexity of the issue space required to describe all of the features of every camera on offer, and the level of subjective judgement required to relate combinations of those features to meaningful criteria.

In large issue spaces, in terms of which an agent is unable to specify a preference ordering, there is a useful special case when it is possible to specify preferences on each issue individually (e.g. "I prefer to pay less than more", "I prefer to have a feature on the camera to not having it"). In this case the agent is *individual preference aware*.

4.1 Assessing $\mathbb{P}^t(\text{satisfy}(\beta, \chi, \delta))$

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 one of β's criteria is $\mathbb{P}^t(\text{ease-of-use}(\beta, \delta))$. The idea is that β will ask α to demonstrate how certain

[8] The seller will have little difficulty in deciding whether a contract is acceptable if he knows what the items cost.

tasks are performed, will observe the sequences that α performs, and will use those observations to revise this probability distribution until some clear verdict appears.

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 $\mathcal{E}$ ={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 $e \in \mathcal{E}$ 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 $J_\pi^X(s)$ 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 $\boldsymbol{p}$, and let $\boldsymbol{p}_{(s)}$ be the distribution with minimum relative entropy[9] with respect to $\boldsymbol{p}$: $\boldsymbol{p}_{(s)} = \arg\min_{\boldsymbol{r}} \sum_j r_j \log \frac{r_j}{p_j}$ that satisfies the constraints $J_s^X(s)$. Then let $\boldsymbol{q}_{(s)}$ be the distribution:

$$\boldsymbol{q}_{(s)} = \mathbb{R}^t(\pi, \tau, s) \times \boldsymbol{p}_{(s)} + (1 - \mathbb{R}^t(\pi, \tau, s)) \times \boldsymbol{p} \tag{4}$$

and then let:

$$\mathbb{P}^t(X_{(s)}) = \begin{cases} \boldsymbol{q}_{(s)} & \text{if } \boldsymbol{q}_{(s)} \text{ is more interesting than } \boldsymbol{p} \\ \boldsymbol{p} & \text{otherwise} \end{cases} \tag{5}$$

A general measure of whether $\boldsymbol{q}_{(s)}$ is more interesting than $\boldsymbol{p}$ is: $\mathbb{K}(\boldsymbol{q}_{(s)} \| \mathbb{D}(X)) > \mathbb{K}(\boldsymbol{p} \| \mathbb{D}(X))$, where $\mathbb{K}(\boldsymbol{x} \| \boldsymbol{y}) = \sum_j x_j \log \frac{x_j}{y_j}$ is the Kullback-Leibler distance between two probability distributions $\boldsymbol{x}$ and $\boldsymbol{y}$, 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 5 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 $\mathbb{P}(X = \text{acceptable}) = 0.25$ is discarded.

Equation 4 simply adds in new evidence $\boldsymbol{p}_{(s)}$ to $\boldsymbol{p}$ weighted with $\mathbb{R}^t(\pi, \tau, s)$. This is fairly crude, but the observations are unlikely to be independent and the idea is that π

[9] Given a probability distribution $\boldsymbol{q}$, the *minimum relative entropy distribution* $\boldsymbol{p} = (p_1, \ldots, p_I)$ subject to a set of J linear constraints $\boldsymbol{g} = \{g_j(\boldsymbol{p}) = \boldsymbol{a_j} \cdot \boldsymbol{p} - c_j = 0\}, j = 1, \ldots, J$ (that must include the constraint $\sum_i p_i - 1 = 0$) is: $\boldsymbol{p} = \arg\min_{\boldsymbol{r}} \sum_j r_j \log \frac{r_j}{q_j}$. This may be calculated by introducing Lagrange multipliers $\boldsymbol{\lambda}$: $L(\boldsymbol{p}, \boldsymbol{\lambda}) = \sum_j p_j \log \frac{p_j}{q_j} + \boldsymbol{\lambda} \cdot \boldsymbol{g}$. Minimising L, $\{\frac{\partial L}{\partial \lambda_j} = g_j(\boldsymbol{p}) = 0\}, j = 1, \ldots, J$ is the set of given constraints $\boldsymbol{g}$, and a solution to $\frac{\partial L}{\partial p_i} = 0, i = 1, \ldots, I$ leads eventually to $\boldsymbol{p}$. Entropy-based inference is a form of Bayesian inference that is convenient when the data is sparse [6] and encapsulates common-sense reasoning [7].

will specify a "fairly comprehensive" set of tasks aimed to determine $\mathbb{P}^t(X)$ to a level of certainty sufficient for Equation 2.

4.2 Assessing $\mathbb{P}^t(\mathrm{obj}(\alpha, \delta))$

$\mathbb{P}^t(\mathrm{obj}(\beta, \delta))$ estimates the belief that δ is acceptable in the open-market that β may observe in the scenario. Information-based agents model what they don't know with certainty as probability distributions. Suppose that X is a discrete random variable whose true value is the open-market value of an item. First, β should be able to bound X to an interval $(x_{\min}, x_{\max})$ — if this is all the evidence that β can muster then X will be the flat distribution (with maximum entropy) in this interval, and $\mathbb{P}^t(\mathrm{obj}(\beta, (\mathrm{item}, y)) = \sum_{x \geq y} \mathbb{P}(X = x)$. β may observe evidence, perhaps as observed sale prices for similar items, that enables him to revise particular values in the distribution for X. A method [2] similar to that described in Section 4.1 is used to derive the posterior distribution — it is not detailed here. An interesting aspect of this approach is that it works equally well when the valuation space has more than one dimension.

5 The Seller Models the Buyer

In this Section we consider how the seller might model the buyer's contract acceptance logic in an argumentative context. As in Section 4 we focus on Equation 2 and for reasons of economy concentrate on the factor: $\mathbb{P}^t(\mathrm{satisfy}(\alpha, \chi, \delta))$.

5.1 Modelling Contract Acceptance

Suppose that β has found an item that he wants to buy, α will be interested in how much he is prepared to pay. In a similar way to Section 4.2, α can interpret β's proposals as willingness to accept the offers proposed, and counter-offers as reluctance to accept the agent's prior offer — all of these interpretations being qualified with an epistemic belief probability. Entropy-based inference is then used to derive a complete probability distribution over the space of offers for a random variable that represents the partner's limit offers. This distribution is "the least biased estimate possible on the given information; i.e. it is maximally noncommittal with regard to missing information" [8]. If there are n-issues then the space of limit offers will be an $(n-1)$-dimensional surface through offer space. As described in [2], this method works well as long as the number of issues is not large and as long as the agent is aware of its partner's preferences along each dimension of the issue space.

5.2 Estimating β's key criteria.

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

β may present information in the form of a high-level description of what is required; e.g. "I want a camera for every-day family use". In a sense this is nothing more than a criterion, but it does not fit comfortably within the terms of Equation 2 and may have implications for all of them. We call such a statement as the *object* of the negotiation. It is realistic to assume that the object is common knowledge to some degree. We assume that there is a structured section of the ontology that describes negotiation objects. We also assume that for each object there is are prior probabilities associated with each of a set of negotiation criteria represented in another section of the ontology. For example, the object “A camera for every-day family use” may associate the prior probability distribution $(0.6, 0.4, 0.0, 0.0)$ with “ease of use for point-and-shoot” in terms of the example evaluation space given above.

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 4.1. 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 5 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 our scenario, during a dialogue β will ask α to perform certain tasks that we assume are represented in a structured section of the ontology. Following the reasoning above, if α is asked to perform task τ then this may suggest prior beliefs of what β's criteria are. For example, suppose α is asked to demonstrate how to photograph a duck when the object is a “camera suitable for photographing wildlife”. If the ontology relates ducks to water to some degree then α may believe that β rates the criterion “waterproof” as “I would prefer to have this” with some probability. Then if α is subsequently asked to to demonstrate how to photograph a duck when the object is a s' she may believe that “waterproof” is a criterion. Using the semantic-similarity-based method described in [2], this evidence would update the estimates for $\mathbb{P}^t(\text{Waterproof} = e|(O = o', T = \tau))$ in a way that is moderated by the semantic distance between s' and ‘camera suitable for photographing wildlife‘”

The discussion so far has not considered argumentative dialogue such as: “My grandmother could not climb into that (car)”. This statement would presumably follow a request to demonstrate the task “how to get into the car” that the observer would rate against the criterion “suitable for an octogenarian” as “unacceptable”. So this example at least can be accommodated in the framework as long as we can link the ‘grandmother’ with an ‘octogenarian’. Given two strings of ontological concepts, S_1 and S_2, define their *similarity* as:

$$\Pi(S_1, S_2) = \frac{\sum_{x \in S_1, y \in S_2} \text{Sim}(x, y)}{||S_1|| \times ||S_2||} \tag{6}$$

where $||S||$ is the number of concepts in S, and $\text{Sim}(\cdot)$ is measures semantic distance [9]. A semantically deeper analyses of text would be better, but we claim that this approach will link such the given statement with the criterion "suitable for an octogenarian" — particularly if the set of admissible criteria are known and agreed in advance.

5.3 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.

The bargaining literature consistently advises (see for example [10]) that an agent should change its *stance* (one dimension of stance being the 'nice guy' / 'tough guy' axis) to prevent other agents from decrypting their private information, and so we should expect some sort of "smoke screen" surrounding any dialogue between competitive agents. It would be convenient to think of disposition as the mirror-image of stance, but what matters is the agent's confidence in its model of the partner. The problem is to differentiate between a partner that is skilfully preventing us from decrypting their private information, and a partner that has either had a fundamental change of heart or has changed his mind in a way that will significantly influence the set of contracts that he will agree to. The first of these is normal behaviour, and the second means that the models of the partner may well be inaccurate.

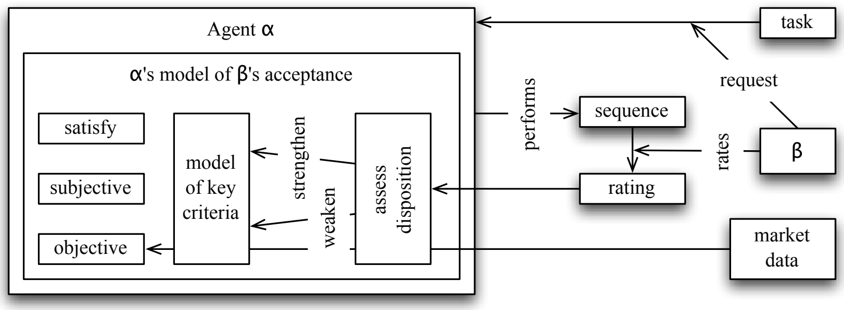
If an agent believes that her partner's disposition has altered then the entropy of her model of the partner should be increased — particularly beliefs concerning the key criteria should be relaxed to prevent the dialogue attempting to enter a "blind alley", and to permit the search for common ground to proceed on broader basis. The mechanics for achieving this are simple: if an agent believes that his partner's disposition has shifted then his certainty of belief in the structure of the model of the partner is decreased.

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. Then the idea is that if in the negotiation for a camera "for family use" α is asked to demonstrate how to photograph a drop of water falling from a tap then this would presumably cause a dramatic difference between $\mathbb{P}^t(C = e|(O = \textit{"family use"}))$ and $\mathbb{P}^t(C = e|(O = \textit{"family use"}, O' = \textit{"photograph water drops"}))$. This difference causes α to revise her belief in "family use", to revise the disposition towards distributions of higher entropy, and to approach the negotiation on a broader basis. A high-level diagram of α's model of β's acceptance criteria that includes disposition is shown in Figure 1.

6 Strategies

In this section we describe the components of an argumentation strategy starting with tools for valuing information revelation that are used to model the fairness of a negotiation dialogue. This work compares with [11], [12] and [13].

Fig. 1. The model of β's acceptance criteria that lies at the heart of the argumentation strategy.



6.1 Information Revelation: computing counter proposals

Everything that an agent communicates gives away information. The simple offer "you may purchase this wine for €3" may be intrepretd in a utilitarian sense (e.g. the profit that you could make by purchasing it), and as information (in terms of the reduction of your entropy or uncertainty in your beliefs about my limit price for the item). Information-based agents value information exchanged, and attempt to manage the associated costs and benefits.

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 [3] 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 $\mathcal{Y}$ and a categorising function $\kappa : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{Y})$. The power set, $\mathcal{P}(\mathcal{Y})$, 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 $\mathcal{G} = \mathcal{Y} \times V$.

Suppose α receives utterance u from β and that category $y \in \kappa(u)$. For any concept $x \in V$, define $\Delta(u, x) = \max_{x' \in concepts(u)} \text{Sim}(x', x)$. 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*[10] 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.

We are particularly interested in the concept of intimacy in so far as it estimates what α knows about β's criteria, and about the certainty of α's estimates of the random variables $\{C_i\}$. We are interested in balance as a measure of the 'fairness' of the dialogue. If α shows β how to take a perfect photograph of a duck then it is reasonable to expect some information at least in return.

Moreover, α acts proactively to satisfy her needs — that are organised in a hierarchy[11] of *needs*, Ξ, and a function $\omega : \Xi \rightarrow \mathcal{P}(W)$ where W is the set of perceivable states, and $\omega(\chi)$ is the set of states that satisfy need $\chi \in \Xi$. Needs turn 'on' spontaneously, and in response to *triggers*. They turn 'off' because α believes they are satisfied. When a need fires, a plan is chosen to satisfy that need (we do not describe plans here). If α is to contemplate the future she will need some idea of her future needs — this is represented in her *needs model*: $\upsilon : T \rightarrow \times^{|\Xi|}[0,1]$ where T is time, and: $\upsilon(t) = (\chi_1^t, \ldots, \chi_{|\Xi|}^t)$ where $\chi_i^t = \mathbb{P}(\text{need } \chi_i \text{ fires at time } t)$.

Given the needs model, υ, α's *relationship model* $(\text{Relate}(\cdot))$ 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, $\{(A_i^{*t}, B_{*i}^t)\}_{i=1}^{|Agents|} = \text{Relate}(\upsilon, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ where, $\boldsymbol{X}_i$ is the trust model, $\boldsymbol{Y}_i$ is the honour model and $\boldsymbol{Z}_i$ 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 $\mathcal{Y} \times V$.

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. $\beta_i = \text{Select}(\chi, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. 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. $(A_i^{**t}, B_i^{**t}) = \text{Set}(\chi, A_i^{*t}, B_i^{*t})$. In particular, if the balance target, B_i^{**t}, is grossly exceeded by β failing to co-operate then it becomes a trigger for α to terminate the negotiation.

6.2 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. In information-based argumentation we are 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.

[10] 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 $\mathcal{M}^t$ is: $\mathbb{I}(u) = \mathbb{H}(\mathcal{M}^t) - \mathbb{H}(\mathcal{M}^{t+1})$.

[11] In the sense of the well-known Maslow hierarchy [14], where the satisfaction of needs that are lower in the hierarchy take precedence over the satisfaction of needs that are higher.

Information-based argumentation attempts to counter the effect of the partner's arguments, in the simple negotiation protocol used here, an argumentative utterance, u, will either contain a justification of the proposal it accompanies, a rating and justification of one of α demonstration sequences, or a counter-justification of one of α's prior proposals or arguments. 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 $P^t(C = e|(O = o, \mathcal{T} = \tau))$ — 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. In the spirit of the scenario, we assume that for every criterion and object pair (C, o) α has a supply of positive argumentative statements $\mathcal{L}_{(C,o)}$. 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 $M_u \subseteq \mathcal{L}_{(C,o)}$ 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.

Suppose that u proposes a contract. α will either decide to accept it or to make a counter offer. We do not describe the bargaining process here, see [2].

6.3 All together

If β_i communicates u then α responds with:

$$u' = Argue(u, \mathcal{M}^t, \Psi^t, A^{**t}, B^{**t}, C_u, N_u, M_u, D_u))$$

where:

- the *negotiation* mechanisms as explained in Section 6.1 sets parameters A^{**t}, B^{**t}) (see e.g. [3] for further details);
- the *argumentation* process determines the parameters N_u, M_u needed to generate the accompanying arguments to the proposal, see Section 6.2;
- the *criteria* modeling process determines the set of criteria C_u used by our opponent to assess the proposals, see Section 5.2; and,

– the *disposition* modeling sets the distributions D_u used to interpret the stance of the opponent, see Section 5.3.

The personality of the agent will be determined by the particular f chosen to select the answer to send. The study of concrete functions is subject of ongoing research as well as their application into a eProcurement setting.

7 Discussion

We have described an approach to argumentation that aims to:

– discover what the partner's key evaluative criteria are,
– model how the partner is evaluating his key criteria given some evidence,
– influence the partner's evaluation of his key criteria,
– influence the relative importance that the partner attaches to those criteria, and
– introduce new key criteria when it is strategic to do so.

The ideas described here are an attempt to develop an approach to argumentation that may be used in the interests of both parties. It aims to achieve this by unearthing the 'top layer' of the partner's reasoning apparatus and by attempting to work with it rather than against it. To this end, the utterances produced 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". In the long term, this behaviour is intended to lead to the development of lasting relationships between agents that are underpinned both by the knowledge that their partners "treat them well" and that their partners act as they do "for the right reasons".

The ideas in this paper have been developed within a highly restrictive scenario that is deliberately asymmetric (being based on a buy / seller relationship). The structure of the analysis is far more general and applies to any scenario in which something has to be bought/made/designed that satisfies a need, and that can do various things. The agents who try to make it do things (use-cases if you like) subjectively rate what they see.

In previous work [3] we have advocated the gradual development of trust and intimacy[12] through successive argumentative exchanges as a way of building relationships between agents. The act of passing private information carries with it a sense of trust of the sender in the receiver, and having done so the sender will wish to observe that the receiver respects the information received. In this paper we have gone one step further by including a modest degree of *understanding* in the sense that an agent attempts to understand what her partner likes. This falls well short of a deep model of the partner's reasoning but we believe strikes a reasonable balance between being meaningful and being achievable. This augments the tools for building social relationships through argumentation by establishing:

– *trust* — my belief in the veracity of your *commitments*
– *intimacy* — my belief in the extent to which I know your private *information*

[12] The revelation of private information.

– *understanding* — my belief in the extent to which I know what you *like*

Acknowledgements This research has been supported by the Sabbatical Programme of the Catalan Government BE2007, the Australian Research Council Discovery Grant DP0557168, and by the Spanish Ministerio de Educación y Ciencia project "Agreement Technologies" (CONSOLIDER CSD2007-0022, INGENIO 2010).

References

1. Sierra, C., Debenham, J.: Information-based agency. In: Proceedings of Twentieth International Joint Conference on Artificial Intelligence IJCAI-07, Hyderabad, India (2007) 1513–1518
2. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In Stone, P., Weiss, G., eds.: Proceedings Fifth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2006, Hakodate, Japan, ACM Press, New York (2006) 1225 – 1232
3. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: Proceedings Sixth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2007, Honolulu, Hawai'i (2007) 1026–1033
4. Dung, P.M.: On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence **77** (1995) 321–358
5. Rahwan, I., Ramchurn, S., Jennings, N., McBurney, P., Parsons, S., Sonenberg, E.: Argumentation-based negotiation. Knowledge Engineering Review **18** (2003) 343–375
6. Cheeseman, P., Stutz, J.: On The Relationship between Bayesian and Maximum Entropy Inference. In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering. American Institute of Physics, Melville, NY, USA (2004) 445 – 461
7. Paris, J.: Common sense and maximum entropy. Synthese **117** (1999) 75 – 93
8. Jaynes, E.: Information theory and statistical mechanics: Part I. Physical Review **106** (1957) 620 – 630
9. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering **15** (2003) 871 – 882
10. Lewicki, R.J., Saunders, D.M., Minton, J.W.: Essentials of Negotiation. McGraw Hill (2001)
11. Bentahar, J., Mbarki, M., Meyer, J.J.C., Moulin, B.: Strategic agent communication: An argumentation-driven approach. In Baldoni, M., Son, T.C., van Riemsdijk, M.B., Winikoff, M., eds.: Declarative Agent Languages and Technologies V. Volume 5397 of Lecture Notes in Computer Science., Springer (2008) 233–250
12. Kakas, A., Maudet, N., Moraitis, P.: Layered strategies and protocols layered strategies and protocols for argumentation-based agent interaction. In Rahwan, I., Moraitis, P., Reed, C., eds.: Argumentation in Multi-Agent Systems ArgMAS 2004. Volume 3366., Heidelberg, Springer Verlag (2005) 64–77
13. Mbarki, M., Bentahar, J., Moulin, B.: Specification and complexity of strategic-based reasoning using argumentation. In Maudet, N., Parsons, S., Rahwan, I., eds.: ArgMAS. Volume 4766 of Lecture Notes in Computer Science., Springer Verlag (2006) 142–160
14. Maslow, A.H.: A theory of human motivation. Psychological Review **50** (1943) 370–396