

Planning clarification questions to resolve ambiguous references to objects

Jeremy L Wyatt

Intelligent Robotics Laboratory
School of Computer Science
University of Birmingham
Birmingham, UK, B15 2TT
jeremy.wyatt@britishlibrary.net

Abstract

Our aim is to design robots that can have task directed conversations with humans about objects in a table top scene. One of the pre-requisites is that the robot is able to correctly identify the object to which another speaker refers. This is not trivial as human references to objects are often ambiguous, and rely on contextual information from the scene, the task, or the dialogue to resolve the reference. This paper describes work in progress on building a robot system able to plan the content of clarifying questions that when answered provide the robot with enough information to resolve ambiguous references. It describes an algorithm that models the degree of uncertainty about the binding of a referent using a probability distribution. We use the visual salience of the object as a way to generate the prior distribution over candidate objects, which we call the belief state. Then we generate action models, for the effects of various clarifying questions, on the fly. Finally we evaluate the mean reduction in the entropy of the resulting belief states. The method can be seen as a form of prior-posterior analysis, or as one step look ahead in an information state Markov decision process. We are currently implementing the algorithm in a robot and discuss the issues we have encountered to date.

1 Introduction

Human-robot communication is an increasingly active field [Roy *et al.*, 2004; Sidner *et al.*, 2004; Sidner and Dzikovska, 2004; Oates *et al.*, 2000], with many challenging problems. One of the most basic abilities for a robot capable of conversing with a human about objects in a scene is the ability to bind the references made by a speaker to objects in the world. One of the problems of human dialogue is that references to objects are often linguistically underspecified. Because of this the robot may need to incorporate other information to resolve the reference. An example is the case below:

- H: “What is to the left of the red cup?”
- R: “Is it the large red cup?”
- H: “Yes.”



Figure 1: A scene with three red mugs, and two other objects.

- R: “There is a green ball to the left of that cup.”

Here the tutor has asked the robot about the identity of an object with a particular relationship to another object (a red cup) in the scene. First of all the question makes a reference to a red object, and indirectly (through the spatial relationship to this object) to the object of interest. Answering the question requires that the robot is capable of decoding this indirect reference, i.e. figuring out which object is the one to the left of the red cup. Here the reference involves an ambiguous reference to another object (a red cup) which is being used as a landmark. If the reference to the red cup is ambiguous (here there are two cups with significant areas of red, each with an object to their left), then the robot must take an action to resolve this ambiguity. This could involve checking to see if the human is pointing at the object, or it could involve asking a clarifying question as in the dialogue above. Where there are several possible clarifying actions we will require a system for generating clarifying questions and for deciding which one is most appropriate. In addition the system should be able to incorporate information from either language or vision.

In this paper we shall suppose that we have a vision system that is capable of producing a list of the objects in the scene, and their approximate positions on the ground plane. We will also assume that it is possible from these to generate a

scene graph, which we have already been able to do for small numbers of simple objects [Kruijff and Kelleher, 2005]. It is important to note that while this is possible for a very limited number of fixed objects, it is not possible for a wide variety of objects, or for cluttered or complex scenes. To make a real robot system scale, further mechanisms are necessary, particularly an attentional system. We return to this issue briefly in the final section. The rest of the paper is structured as follows. In section 2 we introduce a running example, in section 3 we describe the kind of clarifying actions we can take to resolve ambiguity. In section 4 we describe how we generate action models for the dialogue moves on the fly, and then in section 5 we describe the various metrics we use to evaluate them.

2 An example problem

Suppose that we have a vision system that is capable of building a list of objects together with their attributes. The attributes we can reasonably expect to get from our existing vision system are as follows:

```
Object      :category
            :colour
            :projective relations
            :proximity relations
            :position in image plane
            :position on ground plane
```

In order to extract the proximity relations we use a potential field model to capture the notion of nearness between objects [Kruijff and Kelleher, 2005]. Using this model we can extract the qualitative relationship *near* from the geometric locations of the objects in a way that is contextually sensitive. In Figure 2, for example we might say that object o_8 is near o_2 , but this will depend on how far away the objects are, their relative size, and where other objects are relative to them. To model projective relationships, such as *left of*, *right of*, *behind* and *in front of* we create simple tessellations of the space centred on the objects themselves. At the moment we only extract the relationships for the robot's own frame of reference, but given a geometric model it is possible to extend this to other frames of reference, i.e. speaker or object based.

We also have a discourse context, which consists of an information state and some additional structures. These are based around a set of logical forms, each of which represents the content of an utterance. Each logical form will contain one or more discourse referents. These include referents that are variables which can be bound to physical objects in the scene. A significant part of the task of integrating the visual and linguistically derived representations is to decide which object to bind to which discourse referent. We use a couple of running examples. In the first suppose the human utters the question: "What colour is the mug?" where there are objects arranged on the table as shown in Figure 2.

To answer this question the discourse manager needs to create a goal. This goal is to reach an information state where the human agent believes the mug to have a certain colour, and where that belief is correct. To satisfy this goal the robot will have to make an utterance stating the colour of the mug. In order to do this it needs to bind the referent to a physical

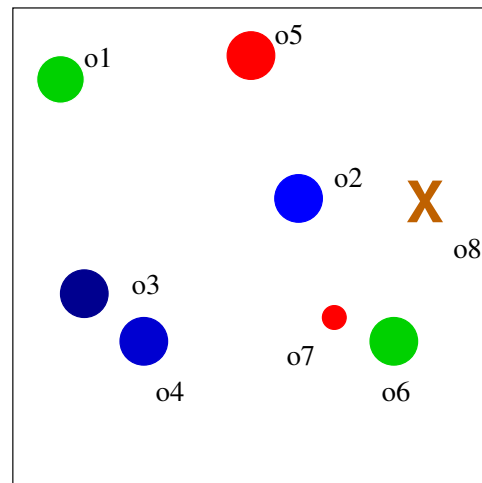


Figure 2: An overhead map of some objects on the table. Large circles represent mugs, the small circle represents a ball, the cross represents an object that has not been identified.

object. In this example there is a significant degree of referential ambiguity. In the second example, let's assume the same layout of objects on the table, but that the human instead says "Pick up the mug".

Let's also assume that when the discourse manager realises that a reference is ambiguous it makes a call to resolution routine *resolve(discourse_referent, discourse_context, scene_graph)* which will plan and make an utterance to resolve the reference. We will now consider the actions that this procedure might consider.

3 Asking clarifying questions to resolve references

What are the different actions that the robot could take in order to resolve the reference above? In the first example (where we are asked for the colour of the mug) we will need a representation that captures the rule that it is not sensible to ask a question of the form "Is it the red mug?". Obviously, in the second example ("Pick up the mug.") this question is fine. We don't discuss this problem further in this paper. Reasonable things the robot can do for example 1 include:

- Checking to see if the speaker is pointing at the object.
- Asking whether it is on the left or the right of the scene.
- Asking if it is at the back or the front of the scene.
- Asking if it has some projective or proximal relationship to some other object not in the set of objects that can be bound, e.g. "Is it next to the red ball?"
- Asking the human to point at the mug.
- Asking the human the general question "Which mug is it?"

- Asking a question about which mug it is by referring to a mug using one of a number of other attributes e.g. ‘‘Is it the big mug?’’.

Reasonable things I can do for example 2 include all of the above, and questions about colour:

- ‘‘Is it the red mug?’’
- ‘‘What colour is the mug?’’
- ‘‘Is it the green mug next to the red ball?’’
- ‘‘Which mug is it?’’

In the following sections we consider the effects of questions about the attributes or relations of an object, of general questions (“which one is it?”), and of questions involving referring expressions. What we ideally want is a model of selection that while not enforcing the sorts of utterances that humans make, settles upon those utterances for good reasons. As an example we want a model that will typically not ask questions such as “Is it blue?”, as this is not a natural response, but would prefer questions such as “Is it the blue mug?” or “Which mug is it?”. We want this preference to arise out of sensible criteria (such as the cognitive load, or the expected gain in information), rather than by excluding certain classes of question from consideration. The question we should prefer will depend on the degree of ambiguity in the reference. Systems such as Ripley [Roy *et al.*, 2004] use a simple catch all strategy of asking “Which one is it?” whenever confronted with referential ambiguity. While such a strategy is quite effective humans use a wider range of expressions dependent on context and the degree of uncertainty. We want to be able to produce qualitatively similar behaviour from our model.

Our first problem is representing the degree of uncertainty about the reference, and using this to incorporate existing information from modalities other than language, e.g. vision. The objects in the set of possible referents have varying degrees of visual salience, large objects in the foreground are highly salient, whereas small objects in the background are not. There is a simple algorithm that lets us calculate the visual salience for an object in terms of its size on the image plane, and how central it is in the current view [Kelleher and van Genabith, 2004]. In that system the visual salience alone is used to resolve references to objects that have not previously appeared in the dialogue, but which appear in the visual scene. We extend this, by assuming that picking the most visually salient object that satisfies the reference may not be enough. In addition we want a mechanism that allows us to measure the degree of ambiguity after taking the salience into account. We propose that having obtained these saliences we normalise them and interpret them as probabilities in a prior:

$$p_i = \Pr(d_1 = o_i) = \frac{\text{salience}(o_i)}{\sum_{j \in D(d_1)} \text{salience}(o_j)} \quad (1)$$

where $D(d_1)$ is the distractor set for the discourse referent d_1 . We will refer to a probability distribution over the elements in the distractor set as a belief distribution. Let’s denote the belief distribution as follows:

$$B(d_1) = \{(o_1, p_1) \dots (o_i, p_i) \dots (o_7, p_7)\} \quad (2)$$

where p_i is the likelihood that the discourse referent binds to object o_i . The prior belief distribution B' is simply the distractor set augmented by the prior probability of each binding. We will sometimes refer to a belief distribution as a state. The distractor set itself is simply the set of objects in the object list (from the visual information) that match the properties of the discourse referent. Our discourse referent has a set of attributes and relations garnered from the utterance. We obtain this set by parsing with a combinatorial categorial grammar designed for talking about objects and their spatial relationships. For the utterance “Pick up the mug.” the attributes of the discourse referent can be simply represented in the same form as our object list supporting our scene graph.

```
d1          :category=mug
```

So initially our distractor set is:

$$D(d_1) = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\} \quad (3)$$

If $D(d_1)$ had a single element then we would want `resolve` to terminate and return the only possible binding. If the set is empty, then we must return with that fact and look for a candidate object elsewhere. If there are several possible bindings, then we must ask clarifying questions to gain information in order to resolve the reference. Let’s denote asking a question as `ask(x, q)` where x is the agent asked and q is the question. For each question that we can ask we will assume that we can build an action model on the fly that effectively generates a set of possible new information states, in that it generates a set of possible new distractor sets after the question is answered.

4 Generating action models for dialogue moves on the fly

There are many algorithms for generating referring expressions that take an object and a distractor set for which we have a set of relations and properties and calculate an expression that will refer uniquely to that object if possible. So if we want to generate a question of the form ‘‘is $d_1 = o_1$?’’ then we can call such an algorithm to generate a referring expression for o_1 . In our algorithm we will, as part of the planning process, internally generate a referring expression for every object in the distractor set. We will therefore also assume that we are able to generate a call to an algorithm for generating referring expressions from the scene graph, `generate_reference(scene_graph)`. A reasonable algorithm that requires a scene graph and given this efficiently handles referring expressions involving properties and spatial relations is that of [Krahmer *et al.*, 2003]. These referring expressions will be the basis of questions of the form ‘‘Is it the ... ?’’.

Of course until we’ve called the algorithm we may not have any idea as to whether it is possible to generate such a referring expression given the information we have. Consider distinguishing between the two blue mugs on the left of Figure 2. Suppose that there is nothing in the scene graph that

enables us to generate a referring expression that uniquely identifies either of the two objects. If it is not possible to generate an unambiguous referring expression for an object then the least ambiguous referring expression can be returned, and a question can be considered which will ask about whether the attributes of the object match those returned (e.g. if we no information about spatial relations in the graph we could ask "Is it green?".)

Once we have a method to generate possible questions, we must be able to model their effects and then evaluate them. Consider example 2: "Pick up the mug". We may want to consider asking the clarification questions:

- "Is it the green mug on the left?"
- "What colour is the mug?"
- "Is it on the left?"
- "Is it blue?"

Each one of these will move us to a new belief distribution B'' (or equivalently here $B|a, q$). All we need in our action model is a way of stating the likelihood of ending up in a particular state. This is simple. Suppose we have the situation in example 2, "Pick up the mug.", and we have a uniform prior over the candidate objects. If we ask "Is it blue?" then, if p'_i is the prior for object o_i then a trivial Bayes update gives the posterior:

$$p''_{i|q,a} \propto p_{a|q,i} p'_i \quad (4)$$

Where q is the question, and a is the answer. The only thing we need to know in advance is $p_{a|q,i}$, the likelihood of the answer to the question q ("Is it blue?") given that object o_i is the correct binding. This is given by the attributes and relations of the objects as given in the scene graph. If an attribute has the value referred to in the question, or the relation asked about holds, the probability of receiving an answer confirming that attribute or relation, is 1. In our simple case of the question "Is it blue?", $p_{yes|q,i} = 1$ for any blue object, and $p_{no|q,i} = 0$ for any non-blue object. In general, however, what we actually need for the action model is to use these probabilities to estimate the likelihood of making the transition to the posterior belief distribution given by answer a to question q , conditioned over the possible objects. For this we use $p_{a|q,B'}$:

$$p_{a|q,B'} = \sum_{i \in D} p_{a|q,i} p'_i \quad (5)$$

Which is simply the likelihood given the distribution B' that the answer to q will be a . This is why we are generating the action model on the fly, because $p_{a|q}$ will depend on the situation (B). We also have to normalise over all the possible hypotheses (objects). So if we assume a uniform prior over the distractor set, ask "Is it blue?" and get back the answer "Yes" the posterior will be:

$$B(d_1)|blue?, yes = \{(o_2, \frac{1}{3}), (o_3, \frac{1}{3}), (o_4, \frac{1}{3})\} \quad (6)$$

Note that by taking this approach we are making the assumption that the only possible answers to our questions are

ones that are relevant and do not take into account any conversational implicature of the question asked. We will later incorporate the effects of a human taking into account the conversational implicature. We can therefore build a simple probability tree giving the transitions between the belief distributions given the questions I can ask. Let's assume the questions are: $q_1 =$ "Is it on the left?", $q_2 =$ "Is it blue?", $q_3 =$ "What colour is it?":

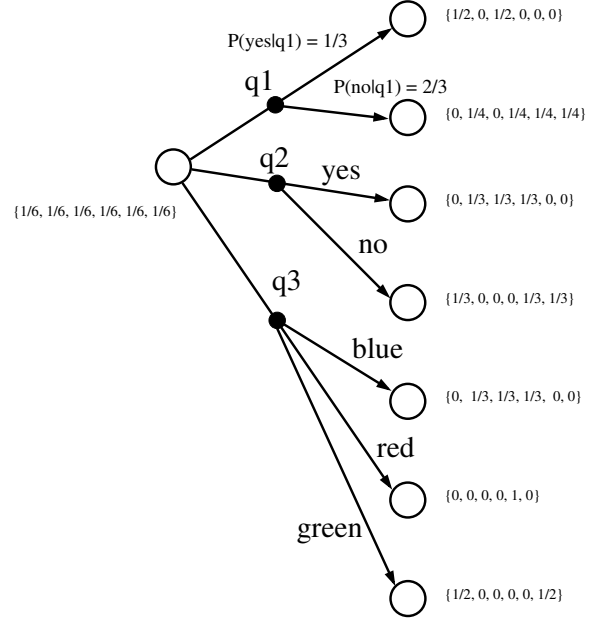


Figure 3: One step action models for three actions.

We may want to consider how to model the effects of multi-modal actions, e.g. pointing at an object while speaking about it. We will return to this question later. For now, we look at how we decide that, given the action models, we prefer one dialogue move action over another.

5 Evaluating belief distributions

There are a number of criteria we intuitively want to include in our evaluation of a dialogue move and its effects. First we simply want to evaluate the likely degree of ambiguity that remains, clearly our main priority should be to remove the ambiguity altogether if possible. Second, we want a model that will take into account the costs for the other speaker of understanding our clarifying action, and also the likely cost to ourselves of understanding their response.

With respect to the first there are two simple ways we can evaluate distributions over belief distributions. First we can use the entropy of the distribution, and second we can use the expected error rate for the Bayes classifier according to the distribution. In the second we are essentially imagining what our likelihood would be of getting it wrong if we were forced to pick an object. If we have a general distribution then we can calculate the two measures as follows:

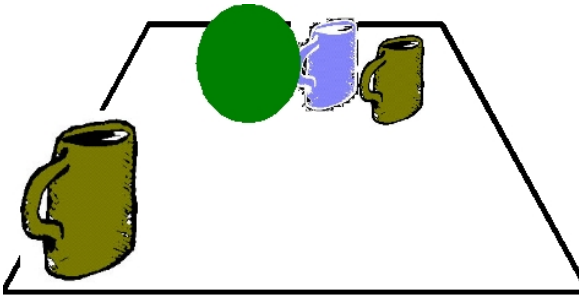


Figure 4: A simple scene. The mug on the left is green, the mug next to the ball is blue, the mug to the right of that is green.

$$Pr(error|q) = \sum_{a \in A} p_{a|q,B} (1 - \max_{k \in B|a} \{p''_{k|a,q}\}) \quad (7)$$

$$E[Entropy|q] = \sum_{a \in A} p_{a|q,B} \sum_{i \in B|a} \{p''_{i|a,q} \log_2 p''_{i|a,q}\} \quad (8)$$

Note that both of these actually calculate the expected value of the belief distributions that could result from a question q , each of which is weighted by the probability of giving the answer a that induces the new state. The inner part is the value of the new belief distribution $B|a$ itself. If we use the expected error rate we obtain very little distinction between most questions. Indeed if we start with a uniform prior as above we can quickly show that all questions with $N = |A|$ possible answers (where the answers are always relevant) will be equally ranked, and that questions are more preferable the larger N is. The error rate is rather insensitive to the type of question asked even if the prior over the distractor set is non-uniform. It transpires that the entropy makes it rather easier to distinguish between the effects of questions in reducing the ambiguity. Of course, we would expect an effective question to reduce the ambiguity, and hence the entropy to near zero.

To take into account the cost of understanding the utterances in the planned dialogue we can use the costs that are typically used in the algorithms for generating referring expressions. Imagine the scene in Figure 4, in which the robot has been given the instruction ‘‘Pick up the mug’’. The complete scene graph is given in Figure 5.

It can be clearly seen that we can attach costs to the arcs in the scene graph, and it is precisely these that are used to calculate the ranked costs of possible referring expressions when generating the possible questions. Classically object type is given the lowest cost, followed by absolute attributes, relative attributes, proximal, and then projective relations [Dale and Reiter, 1995; Krahmer *et al.*, 2003]. We use these costs, which are generated for the best referring expression for each object, to rate the difficulty to the listener of interpreting the robot’s question. We refer to this as a cognitive load model. The full table of loads is:

The total load imposed by an utterance is simply the sum of the loads involved, e.g. ‘‘Is it the green mug to the right of the ball?’’ carries a load

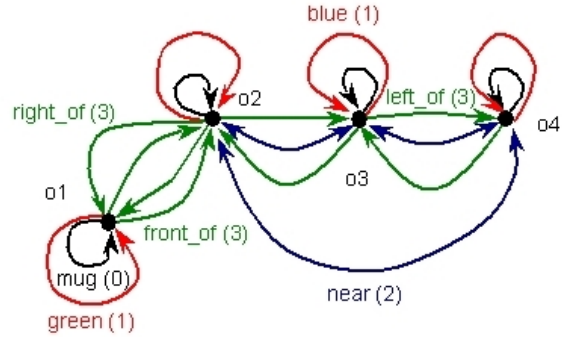


Figure 5: A simple scene graph for the scene in Figure 4. The arcs are labelled with the corresponding attributes and relations, and their associated costs in brackets. Arcs of the same colour are attributes of the same type, and carry the same costs. Not all labels and costs have been included.

Reference by	Example	Load
Type	mug	0
Absolute attribute	red	1
Proximal relation	near	2
Projective relation	left of	3

of $3 + 1 = 4$. We assume that questions that make no references, e.g. ‘‘Which mug is it?’’ carry a load of zero. This load is the cost C_1 in Tables 1–4.

Finally we estimate the expected cost to the robot of processing the response. This depends on whether we model the listener’s ability to take account of the implication that the robot has not understood the reference, and is seeking an answer to resolve the reference. We refer to the two types of speaker as helpful and unhelpful. Costs, or loads, C_2 are the expected costs of interpreting the relevant response.

Table 1: Multi-objective evaluation of three questions for a simple scene. We assume an unhelpful speaker, and a prior of 0.5, 0.25, 0.25.

Question	Evaluation		
	$E[H q, a]$	C_1	C_2
Is it the mug to the left of the ball?	0.5	3	0
Which mug is it?	0	0	2.5
Is it green?	0.69	1	0

Table 2: Multi-objective evaluation of three questions for the same scene, but with a prior of 0.8, 0.1, 0.1.

Question	Evaluation		
	$E[H q, a]$	C_1	C_2
Is it the mug to the left of the ball?	0.2	3	0
Which mug is it?	0	0	2.8
Is it green?	0.45	1	0

The resulting cost estimates for the case of the scene in Figure 4 above are shown in Table 1. In this case we start with

Table 3: Evaluation of the same questions where we assume a helpful speaker, and a prior of 0.5, 0.25,0.25.

Question	Evaluation		
	$E[H q, a]$	C_1	C_2
Is it the mug to the left of the ball?	0	3	1
Which mug is it?	0	0	2.5
Is it green?	0.69	1	.25

Table 4: Evaluation of the same questions, with a helpful speaker and a prior of 0.8, 0.1, 0.1.

Question	Evaluation		
	$E[H q, a]$	C_1	C_2
Is it the mug to the left of the ball?	0	3	0.4
Which mug is it?	0	0	2.8
Is it green?	0.45	1	0.1

a prior distribution over the distractor set of 0.5 for the mug on the left, and 0.25 for the other two mugs. We assume a non helpful speaker. In this case the speaker will provide a literal answer to a question, without giving additional information on the basis of inferring that the robot wishes to know the object in question. In other words the human doesn't give helpful answers, like "No, it's the blue one." to the question "Is it green?".

For the question q_1 : "Is it the mug to the left of the ball?" we need to generate the possible answers, and estimate their different costs. In summary these costs are: the uncertainty remaining after the question and the answer, the load on the human listener of interpreting the question, and the load on the robot of interpreting the answer. Note that we will not calculate an overall combined cost, but merely rank the costs from most to least important. The entropy given a question is the sum of the likelihoods of each answer by the entropy remaining after that answer has been made, as given in Equation 8.

We can see the likelihoods of the answers, and the resulting belief distributions after the answers in Figure 6. If the answer to q_1 is "Yes", then the posterior belief distribution is (1,0,0) the entropy of which is 0. If the answer is "No" the belief distribution is (0, 1/2, 1/2), the entropy of which is 1. So the expected entropy for q_1 is $H|q = 0.5 \times 0 + 0.5 \times 1 = 0.5$. We can calculate the expected entropy for the other questions in the same way.

We now need to calculate the expected load of the answer. If the speaker is unhelpful most of the answers are a simple "Yes", or "No" and they carry a cost of 0. In this case only answers to a question like q_2 : "Which mug is it?" will include a referring expression, and thus carry a load. We simply imagine the referring expressions the robot would choose if it were the speaker, and weight their associated costs by the prior. If the answers are: "It's the mug on the left of the ball."; "It's the mug on the right of the ball." and "It's the blue mug.", the expected load is $C_2 = 0.5 \times 3 + 0.25 \times 3 + 0.25 \times 1 = 2.5$.

Of course, if the speaker is helpful, things will be differ-

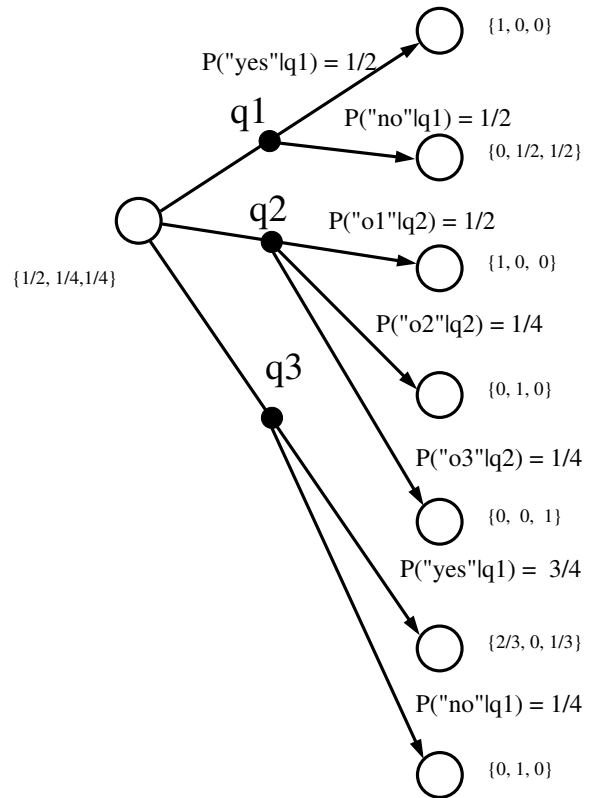


Figure 6: One step model of the three questions for the scene in Figure 4 and the prior (0.5, 0.25, 0.25). Note that here, an answer of "o1" is simply a shorthand for a referring expression to object o_1 .

ent, as shown in Table 3. Under this assumption we can assume that the entropy will usually be zero, since if from our clarifying question the speaker can infer our intent they will make an appropriate referring expression. This causes the expected entropy to be 0 for all questions and outcomes except that of answering "Yes" to the question "Is it green?". We assume here that the human would not give additional information since the intention of the speaker is marginally more ambiguous. If the answer always involves a referring expression however, such as that generated by a helpful speaker, then the expected load of processing the answer rises. This would give q_1 a new expected load for the replies to q_1 of $C_2 = 0.5 \times 0 + 0.25 \times 1 + 0.25 \times 3 = 1$.

Finally we can see from Table 1 that different orderings of the different types of costs will induce different preferences among the questions. In the unhelpful speaker model, if we wish first to reduce entropy below some threshold (say 0.3), then reduce the cognitive load on ourselves, and then reduce the load on the other speaker we will prefer q_2 : "Which mug is it?". If, however, the prior belief distribution changes enough so that we are more confident that the object is o_1 then we will ask "Is it the mug to the left of the ball?". This is shown in Table 2, where the costs are calculated for the same problem, but with a prior belief distribution of (0.8, 0.1, 0.1). The general point here is that

the prior becomes more certain about a specific object it becomes more appealing to ask the question based on the reference to that likely object. This depends on the cost of that reference. Here the cost of a reference to the likely object is high, so that the total costs always exceed those for asking the general question "Which mug is it?".

In the case of a helpful speaker (Tables 3 and 4) we have to decide whether we prefer to lower the load on the robot or the person. An argument for lowering the load on the robot (i.e. preferring questions with a low C_2 to those with a low C_1) is that the time taken for the robot to process an utterance by reference to the visual scene will be significantly higher than for the human. Thus to improve speed of response we will typically prefer the specific question involving a referring expression. It is also worth noting that the model will never prefer attribute based questions, such as "Is it blue?".

Despite this argument it is not entirely clear whether for smooth robot-human dialogue a rational robot should always prefer the higher costs to be placed on a human because of their greater cognitive ability, or whether some other ranking or weighting of the objectives will produce the right mix of behaviour. The flexibility of the cost model is considerable however, so it should not be too challenging to tune it for a reasonable qualitative match with human performance.

6 Discussion and Work in Progress

There is a long history of work on question answering systems, many of which are concerned with similar issues. In [Fleming and Cohen, 2001] a related cost model for reasoning about the benefits of different dialogue strategies, including clarification dialogues is proposed. Clarification question planning is also carried out in [Raskutti and Zukerman, 1997], although there the types of uncertainty involve more structured entities, such as more complex discourse relations. There are also similarities with recent probabilistic approaches to dialogue management [Goddeau and Pineau, 2000; Roy *et al.*, 2000]. In those approaches the problem is also posed as either a Markov decision process (MDP), or a partially observable MDP. Our approach should be viewed as myopic planning over an information state MDP, which is, for our purposes, a simplified way of looking at a POMDP. Here the true underlying state of the system is the actual object to which the human has referred. We also remove all learning from the problem, while retaining a more complex (multi-objective) cost function. Finally there are also some connections to the work on incremental production of references to objects. There is strong empirical evidence [Pechmann, 1989] that humans, particularly adults, tend to over-specify references. Pechmann argues that this is due to the incremental nature of speech production. Although referring expressions could be generated incrementally in our system, it does not attempt to be cognitively plausible in that we plan and take into account the precise nature of the references before we have made them. In this sense, our model is not cognitively plausible, although there are opportunities to explore the real impact of visual processing and visual attention in our robot system. Finally we note that because of the challenges of visual processing to obtain object location the sens-

ing problems addressed here are rather different from those of the "Put-that-there" systems of the 1970s [Bolt, 1980]. Indeed the problem of integrating spatial reasoning with language and vision in robotics in a scalable way for natural scenes is still well beyond the state of the art.

We currently have an almost complete communication system that is capable of parsing utterances about objects in a scene, as well as vision routines able to determine the pose of simple objects on a table top, and determine their visual salience. We are currently implementing the planning algorithm described in this paper. There are important issues that crop up when placing such a system in a robot. The most prominent of these is the need for an attentional system that is capable of deciding which parts of the image and scene to process in order to generate a partial scene graph. Generation of a complete scene graph is not feasible for either humans or robots for complex everyday scenes. This will therefore require a mechanism for deciding what processing to do. This is an extremely interesting question because we believe that it might provide a much more satisfactory model of the cognitive load of different referring expressions than merely ranked costs. In addition there is the opportunity for a robot system to estimate its own costs in terms of the processing times and loads for different operations.

We could ask why we would want to put such effort into generating a variety of responses when a general question will do. Indeed why do we really expect references to be commonly underspecified. The answer is that they may not be once context is taken into account, but a robot may well be incapable of spotting and processing all the relevant cues in time, e.g. pointing gestures, or eye fixations. Conversational robots are likely to need a fall back resolution mechanism that reduces the costs for both speakers by attempting to focus attention whenever possible. We believe that such an approach might both provide an interesting account of human reference resolution, as well as making human-robot dialogue more robust. We plan to present initial results from the robot system at the workshop.

Our plans include incorporating action models for watching and making pointing actions. The filters for these are typically probability distributions expressing the observer's uncertainty about the location to which the arm is pointing. We anticipate that with an appropriate model of pointing that extension to multi-modal actions should be possible.

Acknowledgments

The author would like to acknowledge the work of Geert-Jan Kruijff, and John Kelleher for the design and implementation of the communication system on which this work is based. He would also like to thank the anonymous reviewers for their useful suggestions.

References

- [Bolt, 1980] Richard A. Bolt. "Put that there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on computer graphics and interactive techniques*, pages 262–270, 1980.

- [Dale and Reiter, 1995] Robert Dale and Ehud Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.
- [Fleming and Cohen, 2001] Michael Fleming and Robin Cohen. Dialogue as decision making under uncertainty: The case of mixed-initiative AI systems. In *Proceedings of NAACL-2001 Adaptation in Dialogue Systems Workshop*, 2001.
- [Goddeau and Pineau, 2000] D. Goddeau and J. Pineau. Fast reinforcement learning of dialog strategies. In *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2000.
- [Kelleher and van Genabith, 2004] J. Kelleher and J. van Genabith. Visual salience and reference resolution in simulated 3d environments. *Artificial Intelligence Review*, 21(3):253–267, 2004.
- [Krahmer *et al.*, 2003] E. Krahmer, E.S. van Erk, and A. Verleg. Graph based generation of referring expressions. *Computational Linguistics*, 29(1), 2003.
- [Kruijff and Kelleher, 2005] G.J. Kruijff and J.D. Kelleher. A context-dependent model of proximity and regions. In *to appear in Proceedings of the IJCAI-05*, 2005.
- [Oates *et al.*, 2000] Tim Oates, Zachary Eyer-Walker, and Paul R. Cohen. Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors. In *In Proceedings of The Fourth International Conference on Autonomous Agents.*, pages 227–228, 2000.
- [Pechmann, 1989] Thomas Pechmann. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89:110, 1989.
- [Raskutti and Zukerman, 1997] Bhavani Raskutti and Ingrid Zukerman. Generating queries and replies during information seeking interactions. *IJCHS*, 47(6):689–734, 1997.
- [Roy *et al.*, 2000] N. Roy, J. Pineau, and S. Thrun. Spoken dialog management using probabilistic reasoning. In *ACL*, 2000.
- [Roy *et al.*, 2004] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(3):1374–1383, 2004.
- [Sidner and Dzikovska, 2004] C. Sidner and M. Dzikovska. A first experiment in engagement for human-robot interaction in hosting activities. Technical Report 2003-134, Mitsubishi Electric Research Labs, December 2004.
- [Sidner *et al.*, 2004] C. Sidner, C. Lee, C. Kidd, and N. Lesh. Exploration in engagement for humans and robots. Technical Report 2004-048, Mitsubishi Electric Research Labs, June 2004.