Why do you ask? Good questions provoke informative answers.

Robert X. D. Hawkins, Andreas Stuhlmüller, Judith Degen, Noah D. Goodman
{rxdh,astu,jdegen,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

What makes a question useful? What makes an answer appropriate? In this paper, we formulate a family of increasingly sophisticated models of question-answer behavior within the Rational Speech Act framework. We compare these models based on three different pieces of evidence: first, we demonstrate how our answerer models capture a classic effect in psycholinguistics showing that an answerer's level of informativeness varies with the inferred questioner goal, while keeping the question constant. Second, we jointly test the questioner and answerer components of our model based on empirical evidence from a question-answer reasoning game. Third, we examine a special case of this game to further distinguish among the questioner models. We find that sophisticated pragmatic reasoning is needed to account for some of the data. People can use questions to provide cues to the answerer about their interest, and can select answers that are informative about inferred interests.

Keywords: language understanding; pragmatics; Bayesian models; questions; answers

Introduction

Q:"Are you gonna eat that?" A:"Go ahead."

In this (real life) example, Q strategically chooses a question that differs from her true interest, avoiding an impolite question, yet manages to signal to A what her interests are; A in turn reasons beyond the overt question and provides an answer that addresses Q's interests. This subtle interplay raises two questions for formal models of language: What makes a question useful? What makes an answer appropriate?

A number of studies in psycholinguistics have provided evidence that answerers are both sensitive to a questioner's goals and attempt to be informative with respect to those goals. For instance, in the classic study of Clark (1979), researchers called liquor merchants and opened the conversation with one of two sentences to set context: "I want to buy some bourbon" (the *uninformative* condition) or "I've got $5 to spend" (the *five dollar* condition). They then asked, "Does a fifth of Jim Beam cost more than $5?" Merchants gave a literal yes/no answer significantly more often in the latter condition than the former, where an exact price was more common.

When provided with the five dollar context, the merchant inferred that the questioner's goal was literally to find out whether or not they could afford the whiskey, hence a simple 'yes' sufficed. In the uninformative context, however, the merchant inferred that the questioner's goal was just to buy whiskey, so the exact price was the most relevant response (Clark, 1979). Context and questioner goals have also been implicated in accounts of answers to identification questions like "who is X?" (Boër & Lycan, 1975), and to questions like "where are you?" that permit answers at many levels of abstraction (Potts, 2012). While most of this work has focused on *answerer* behavior, it suggests that the question itself is important in prompting a relevant answer.

Recent work on Rational Speech Act (RSA) models (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) has mathematically formalized pragmatic language understanding as a form of recursive Bayesian inference, where listeners reason about speakers who choose utterances that maximize information gained by an imagined listener. In this paper we extend the RSA framework to address simple question-answer dialogs. The immediate challenge in doing so is that the speaker utility in RSA is based on direct information provided by an utterance—since questions don't provide direct information, we must say what utility they do have.

We suggest, following Van Rooy (2003), that the value of a question is the extent to which it can be expected to elicit information relevant to the questioner later in the dialogue. More specifically, for the questioner, the value of a question is the expected information gained about her interests, given the set of likely answers it may provoke. This diverges from regular RSA in that the value of a question depends on information gained by the speaker (rather than listener), and that this information comes later in the (very short) conversation.

To fully specify this questioner we need a model of the answerer, which can serve as both the model assumed by a questioner, and as a model of answer behavior itself. We explore three, increasingly sophisticated, answerer models. The simplest answerer provides a literal answer to the question (without attempting to be informative); the explicit answerer attempts to be informative with respect to the explicit question asked (without inferring the questioner's underlying interests); the pragmatic answerer infers the most likely true interests of the questioner, and then informatively addresses those interests. The latter model extends RSA to reason about the topic of conversation, as proposed by Kao, Wu, Bergen, and Goodman (2014) to explain hyperbole; it goes beyond previous work by using the explicit question as a (potentially indirect) cue to this topic.

The rest of this paper is structured as follows. First, we lay out the details of our question-answer models. We show that the pragmatic answerer model can select different answers to a question depending on context, as in Clark (1979), described above. We then use a communication game paradigm that allows us to manipulate goals, potential questions, and potential answers, testing the predictions of the different models. We close with a brief discussion of related models and future directions.

A Rational Speech Act model of question and answer behavior

How should a questioner choose between questions? We start by assuming that the questioner aims to *learn information relevant to a private goal*. In order to choose a question that results in useful information, the questioner reasons about how the answerer would respond, given different possible states of the world; she selects a question that results in an answer that tends to provide goal-relevant information.

More formally, suppose there is a set of world states $\mathcal{W}$, a set of possible goals $\mathcal{G}$, a set of possible questions $\mathcal{Q}$, and a set of possible answers $\mathcal{A}$. These sets are taken to be in common ground between the questioner and the answerer. An informational goal $g \in \mathcal{G}$ is a projection function that maps a world state to a particular feature or set of features that the questioner cares about; this is similar to the notion of a question-under-discussion (Roberts, 1996). We will use the notation $P_g(w)$ to indicate the probability $\hat{P}(g(w))$ of the g-relevant aspect of w under the projected distribution $\hat{P}(v) = \int_{\mathcal{W}} \delta_{v=g(w)} P(w) dw$.

The **questioner** takes a goal $g \in \mathcal{G}$ as input and returns a distribution over questions $q \in \mathcal{Q}$:

$$P(q|g) \propto e^{\mathbb{E}_{P(w^*)}\left[D_{KL}(P_g(w|q,w^*) \,\|\, P_g(w))\right] - C(q)}$$

It trades off the cost of asking a question, $C(q)$, and expected information gain. The cost likely depends on question length, among other factors. Information gain is measured as the Kullback-Leibler divergence between the prior distribution over g-relevant worlds, $P_g(w)$, and the posterior distribution one would expect after asking a question q whose answer reflected true world state w^*:

$$P_g(w|q,w^*) = \sum_{a \in \mathcal{A}} P_g(w|q,a) P(a|q,w^*)$$

This distribution has two components: First, it depends on $P(a|q,w^*)$, a model of the answerer which we will explore shortly. Second, it depends on (the goal projection of) $P(w|q,a)$, an 'interpreter' that specifies the likelihood assigned to different worlds given question and answer pairs.

To define the interpreter function, which all agents use to compute the literal interpretation of a question-answer pair, we must assign questions a semantic meaning. We assume that a question is an informational goal that projects from worlds to the answer set $\mathcal{A}$. This is equivalent to the more common partition semantics of Groenendijk and Stokhof (1984), as can be seen by considering the pre-image of such a projection; an answer picks out an element of the partition via $q^{-1}(a)$. The **interpreter** constrains the prior on worlds to the subset of its support that is consistent with the semantics of a question-answer pair[1]:

$$P(w|q,a) \propto P(w) \delta_{q(w)=a}$$

[1]We should also have a semantic evaluation function that maps an answer utterance to its value in $\mathcal{A}$. For clarity we assume this is a trivial mapping and suppress it.

We next describe three different answerer models; the questioner could assume any one of them, leading to three corresponding versions of the questioner model. All answerers take a question $q \in \mathcal{Q}$ and a true world state $w^* \in \mathcal{W}$ as input and return a distribution over answers $a \in \mathcal{A}$. The **literal answerer** simply chooses answers by trading off prior answer probability and how well a question-answer pair conveys the true state of the world to an interpreter:

$$P(a|q,w^*) \propto P(a) P(w^*|q,a)$$

For a fixed question, this is equivalent to the speaker in previous RSA models. The question enters only in specifying the literal meaning of an answer. The **explicit answerer** additionally evaluates answers with respect to how well they address the explicit question q:

$$P(a|q,w^*) \propto P(a) P_q(w^*|q,a)$$

The **pragmatic answerer** also evaluates answers with respect to how well they address the informational goal, but doesn't take the question's explicit meaning at face value. Instead, the pragmatic answerer reasons about which goals g are likely given that a question q was asked, and chooses answers that are good on average:

$$P(a|q,w^*) \propto p(a) \sum_{g \in \mathcal{G}} P(g|q) P_g(w^*|q,a)$$

Reasoning backwards from questions to goals is a simple Bayesian inversion of the (explicit) questioner using a prior on goals:

$$P(g|q) \propto P(q|g) P(g)$$

For all of the questioner and answerer models, we can vary how strongly optimizing they are—that is, to what extent they are sampling from the distributions defined above, and to what extent they deterministically choose the most likely element. For any such distribution over utterances, we introduce an optimality parameter α and transform it by $P'(x) \propto P(x)^{\alpha}$.

This concludes our specification of the model space, giving a set of three answerers and three corresponding questioners that reason about them. We have implemented these models in WebPPL, a probabilistic programming language (Goodman & Stuhlmüller, electronic). The model predictions shown throughout the rest of the paper are computed using this implementation.

Whiskey pricing: a case study

Our model can provide different—sometimes over- or under-informative—answers to the same explicit question, depending on context. To illustrate, we model Clark's (1979) whiskey study. Recall that liquor merchants were more likely to give over-informative answers (specifying exact price) to the question "Does a fifth of Jim Beam cost more than $5?" in the *uninformative* context ("I want to buy some bourbon") than in the *five dollar* context ("I've got $5 to spend").

Our world state is a pair of the whiskey's price ($1, $2, . . . , $10) and a Boolean indicating whether the merchant takes credit cards. There are three possible goals: learning the price of whiskey, learning whether the price is greater than $5, and learning whether the merchant takes credit cards. Note that the credit card question was not in the original study, but reflects the important fact that there exist alternative reasons for calling a liquor store aside from price-related questions. The set of answers includes exact prices as well as "yes" and "no", with lower cost for "yes" and "no" than the price statements.

We model the context sentence as affecting the answerer's goal prior. We assume that there is a fixed 40% probability of the credit card goal, with the remaining 60% split between the two price-related goals. When the context is "I'd like to buy some whiskey," we assume that the split is even. When it is "I only have $5 to spend," we assume that it is 9:1 in favor of learning whether the price is greater than $5.

Results When the question is "Do you take credit cards?", the pragmatic answerer prefers to give the accurate Boolean answer (with probability .76 and .78, weakly depending on context), with no preferential treatment for any of the numeric answers. When the question is "Does Jim Beam cost more than $5?", the correct Boolean answer is still the most probable choice, but more weakly (at probability .44 and .49). Critically, there is a context-dependence for answers to this question: when prefaced with "I'd like to buy some whiskey.", the correct exact price answer is favored more strongly (at probability .18) than when the context is "I only have $5 to spend." (probability .11). By contrast, the explicit answerer (which has no natural way to account for context) does not make differential predictions in the two situations.

This suggests that our pragmatic *answerer* is consistent with human behavior in psychologically interesting situations, passing a first, qualitative, test. However, we have not yet shown that the *questioner* behavior matches that of humans. Indeed, the questioner has been largely neglected in studies of answering (but see, e.g., Potts, 2012), even though, as our opening example illustrates, the choice of question is important for understanding answers. In the next section we introduce an experimental paradigm that allows us to jointly explore quantitative behavior of both questioners and answerers.

Exp. 1: Hierarchical questions and answers

In order to simultaneously test how questioners choose questions when faced with a particular goal and how answerers respond under uncertainty about this goal, we used a guessing-game task played by two players: a questioner and an answerer. In this game, 4 animals (a dalmatian, a poodle, a cat, and a whale) were hidden behind 4 gates. These animals corresponded to different levels in a class hierarchy (see Fig. 1). The questioner received a private goal of finding one of the objects (e.g. 'find the poodle'), and the answerer (but not the questioner) knew the location of each object. Before choosing a gate, the questioner asked the answerer a single question, chosen from a restricted set of options, and the answerer responded by revealing the object behind a single gate. This restriction was motivated by one of the key features of our opening example: when the most direct question ("can I eat your food?") is suppressed due to politeness, utterance length, complexity, or some other intervening factor, questioners must rely instead on an indirect question.

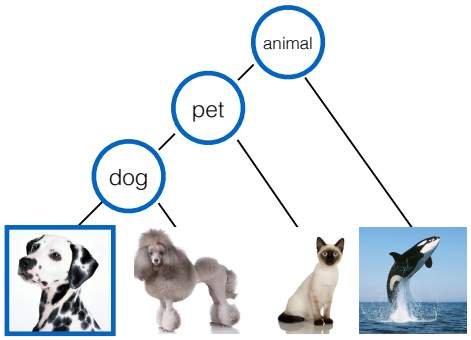

Figure 1: Stimulus hierarchy used in Exp. 1. The goal space and answer space contained the four leaves. The question space, however, was restricted to the highlighted nodes, proceeding up the hierarchy, allowing for indirect questions.

This set of restricted options was critical to distinguishing between the pragmatic and explicit variants of our model. If all questions were equally available, both our 'explicit' and 'pragmatic' questioner models would prefer the most direct one. To see how they make different predictions in the presence of restrictions, suppose 'poodle?' was not available the questioner. If the questioner asked about a 'dog?', the poodle and dalmatian would be considered equally good options by an explicit answerer because they are both dogs. However, the pragmatic answerer could reason that if the questioner was truly interested in the location of the dalmatian, he would have *asked* about the dalmatian. Because he didn't, he must be interested in the other valid response that he lacks a direct question for: the poodle.

Participants We recruited 125 participants from Amazon's Mechanical Turk to participate in this task. Eleven participants were excluded due to self-reported confusion about the task instructions or due to being non-native English speakers.

Stimuli & Procedure In terms of our model specification, the world space $\mathcal{W}$ was the set of $4! = 24$ possible assignments of four objects to four gates. The goal space $\mathcal{G}$ was the set of four objects that the questioner could be trying to find (the leaves of the tree in Fig. 1). The answer space $\mathcal{A}$ was the set of four gates that the answerer could reveal. The restricted question space $\mathcal{Q}$ contained the set of highlighted nodes in the hierarchy: 'dalmatian?', 'dog?', 'pet?', and 'animal?'.

Each participant provided responses for four trials in the role of the questioner (corresponding to the four goals), and four trials in the role of the answerer (corresponding to the four possible questions). In the questioner block, players

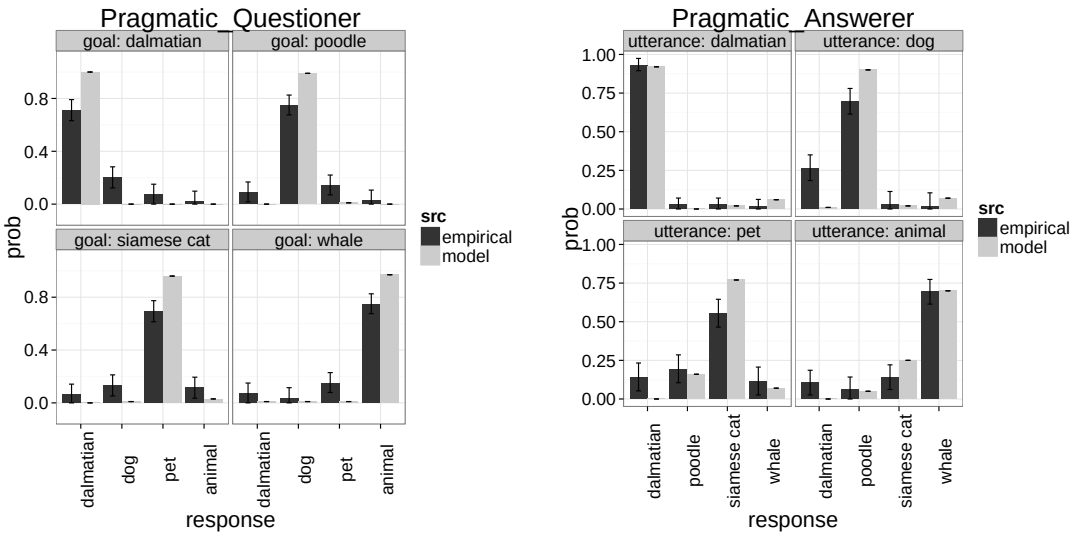

Figure 2: Exp. 1 results, compared with the predictions of the best-performing model for questioner (left) and answerer (right). The explicit and pragmatic questioner models do not make different predictions in this task, but the pragmatic answerer better accounts for the qualitative patterns in the response data than the explicit answerer.

were presented with a private goal from $\mathcal{G}$, like "find the poodle!" and were prompted to select a question from a drop-down menu containing elements of $\mathcal{Q}$ that would best help them find it. In the answerer block, players were shown which items were behind which gates and were told that the other player had asked a particular question from $\mathcal{Q}$. They were prompted to select a gate from a drop-down menu that would be most helpful for the questioner, keeping in mind his or her constraints. (To minimize learning effects, questioners did not receive answers and neither role saw the outcome of the game.) In order to collect responses for all elements of $\mathcal{G}$ and $\mathcal{Q}$, the order of the questioner and answerer blocks was randomly assigned for each participant, and the order of stimuli within these blocks was also randomized[2].

Results Results for the questioner role are shown alongside model predictions in Fig. 2 (left). We find that questioners systematically prefer to ask different questions given different goals, even as those questions become more indirect. χ^2 tests over each of the four response distributions show a significant divergence from uniform. Questioners preferentially ask about the 'dalmatian' given the dalmatian goal, $\chi^2(3) = 137, p < .001$, about the 'dog' given the poodle goal, $\chi^2(3) = 152, p < .001$, about the 'pet' given the cat goal, $\chi^2(3) = 120, p < .001$, and about the 'animal' when given the whale goal, $\chi^2(3) = 150, p < .001$.

Results for the answerer role are shown in Fig. 2 (right). Answerers are highly sensitive to the constraints of the questioner, giving information about the dalmatian when asked about a 'dalmatian', $\chi^2(3) = 281, p < .001$, about the poodle when asked about a 'dog', $\chi^2(3) = 137, p < .001$, about the cat when asked about a 'pet', $\chi^2(3) = 57, p < .001$, and about the whale when asked about an 'animal', $\chi^2(3) = 121, p < .001$. Note that, under an explicit interpretation of the question, revealing the dalmatian and the poodle would both be perfectly acceptable answers to a question about a 'dog', but answerers strongly prefer to give the location of the poodle. In the next section, we compare these results to the predictions of our family of models (Fig. 3).

Model comparison Each model was run with uniform prior probability over worlds, goals, questions, and answers, and with equal cost for all utterances. For each model, a single optimality parameter, which applied to all agents as described above, was fit to maximize correlation with the data.

We can rule out both the literal answerer and literal questioner. The *literal answerer* yields a uniform distribution over the four answers. This has consequences for the corresponding *literal questioner* model: when this questioner reasons about which question would generate the most helpful answer from the literal answerer, it finds no differences in response probabilities, and therefore has no preference for which question to ask. The predictions of these model, plotted against our empirical results, are shown in the left-hand column of Fig. 3.

The two remaining questioner models make roughly the same predictions for this task, and we are not able to distinguish them on the basis of these data. We found a model-data correlation of $r = 0.96$ for the explicit questioner and correlation of $r = 0.99$ for the pragmatic questioner. Although the pragmatic model has a slightly better fit, the two models only differ slightly in the magnitude of predictions, not in qualitatively important ways such as the rank ordering of response. The pragmatic questioner model's predictions for each response distribution are shown in Fig. 2 (left). Although the magnitude of its predictions are not in perfect alignment with

[2]The experiment is online at `http://cocolab.stanford.edu/cogsci2015/Q_and_A/experiment1/experiment1.html`

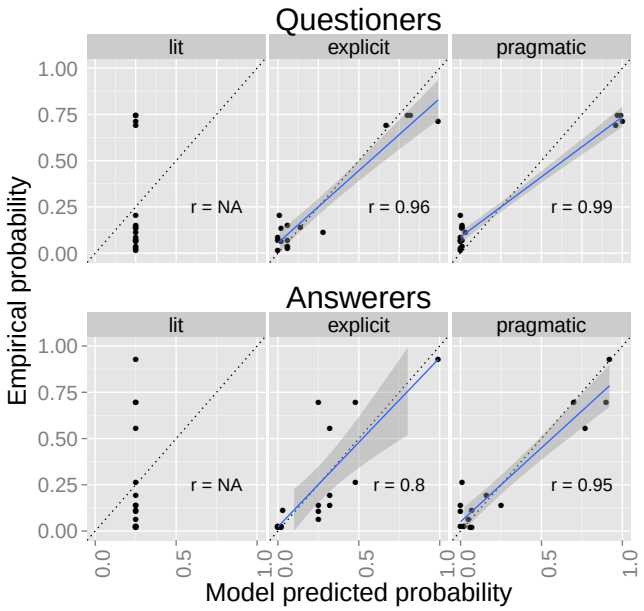

Figure 3: Full space of models, and their correlations with the data from Exp. 1. Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

the magnitude of the empirical data (because it is strongly optimizing), it captures most of the interesting qualitative patterns of the data, particularly the modal responses.

The pragmatic answerer provides a much better fit to the data than the explicit answerer: a model-data correlation of $r = 0.8$ for the explicit answerer and $r = 0.95$ for the pragmatic answerer. Only the pragmatic answerer can account for essential qualitative features of the response data. For example, the explicit answerer predicts that participants will be equally likely to show the 'dalmatian,' 'poodle,' and 'cat' when asked about a pet. Instead, the data show a significant preference for revealing the cat, leaving 'dalmatian' and 'poodle' at the same level as the other alternative. The pragmatic answerer correctly predicts this pattern (see Fig. 2 (right)). Even more dramatically, the explicit answerer predicts a uniform distribution over responses to the 'animal?' question. However, the empirical distribution was significantly different from uniform. Thus, the pragmatic answerer is necessary to account for these data.

These data provide strong evidence for a pragmatic answerer, but are more equivocal with respect to the explicit and pragmatic questioner. Because the two models did not make significantly different predictions for this experiment (and both work quite well), we ran a follow-up study on a special case of the guessing-game paradigm in which the explicit and pragmatic questioners make different predictions.

Exp. 2: A Critical Test of Questioner Models

Participants We recruited 50 participants to participate only in the questioner scenario of the guessing game presented above. Ten participants were excluded on the basis of having a non-English native language, or reporting confusion about the instructions.

Stimuli & Procedure The procedure was the same as before with some changes to the stimuli. The world space $\mathcal{W}$ consisted of possible assignments of the three pets to three gates. The possible goals $\mathcal{G}$ were the dalmatian and poodle (not the cat). The possible questions $\mathcal{G}$ were 'dalmatian?' or 'cat?'. The possible answers $\mathcal{A}$ were the three gates. Each participant was given the two goals in a random order[3].

Results When the goal was to find the dalmatian, participants were significantly more likely to ask about the dalmatian than the cat, $\chi^2(1) = 12, p < 0.001$. When the goal was to find the poodle, participants were marginally more likely to ask about the cat than the dalmatian, $\chi^2(1) = 3.6, p = 0.058$. When looking only at the first of the two trials, the dalmatian result held, $\chi^2(1) = 14.4, p < 0.001$, but participants' preference for asking about the cat disappeared, $\chi^2(1) = 0.07, p = 0.79$. These results are shown in Fig. 4.

Model comparison The explicit questioner predicts that participants should have no preference for a question given the 'poodle' goal, since an explicit answerer would be equally unlikely to give the desired answer for both. The pragmatic questioner model, however, predicts that participants should prefer to ask about the cat. This is because the (internal) pragmatic answerer would reason that if the questioner was interested in the dalmatian, they would ask about the dalmatian; if they didn't, they must be interested in the other possible goal.

It is again unclear which questioner model is best. Overall, the response distribution matches the predictions of the pragmatic model: questioners prefer to ask about the cat. However, participants don't show this behavior if we look at only the first trial. This could be due to a number of reasons. Interestingly, the pragmatic model predicts a more explicit-like response distribution if the questioner does not take into account the constraint on possible goals: if participants thought the poodle was the only goal (counter to the instructions), then asking about the dog would be consistent with the pragmatic model as well. It is possible that participants only fully-processed the alternative (dalmatian) goal if they had first done the trial where that was the goal.

General discussion

Perhaps the most important formal advance of the models considered here is to move the Rational Speech Act framework beyond interpretation of single utterances (in context), to consider the dynamics of simple dialogs (albeit consisting of a single question and its answer). Doing so requires replacing the immediate motive to convey true information with the more distant motive to provoke useful information from one's interlocutor. On the answerer side, sophisticated

[3]The experiment is online at `http://cocolab.stanford.edu/cogsci2015/Q_and_A/experiment2/experiment2.html`

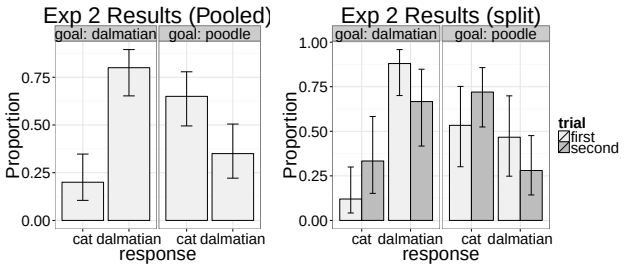

Figure 4: The overall response distribution in Exp. 2 (left) and the same distribution split into first- and second-trial data (right).

inference was required to account for the implicit interests of the questioner. This provides a useful connection to current game-theoretic and decision-theoretic models (Vogel, Bodoia, Potts, & Jurafsky, 2013; Van Rooy, 2003), which also emphasize the importance of goals and speaker beliefs in communication but emphasize less the complex interplay of inference between questioner and answerer.

We have presented evidence that answerer behavior is best described by a pragmatic model that *does* reason about questioner intentions, using the question utterance as a signal. The superiority of pragmatic answerer predictions over the other answerer models was robust. Questioner behavior in Exp. 2, however, seemed to be much more dependent on experience. In another version of Exp. 1, we did not emphasize certain aspects of the game in the instructions, such as the fact that the answerer knows about the restricted answer set, which might prompt perspective-taking. Our data in this pilot experiment appeared to contain a mixture of explicit and pragmatic answerers and questioners (though other confounds were present in this version). Overall, it will be important to explore the mixture of explicit- and pragmatic-questioning across a larger range of situations: these issues may be a product of our artificial game paradigm, or they may be reflective of real tendencies in language use, raising novel questions about audience design in question-answer behavior.

While the artificiality of our question-answer game may distance the behavior of participants from the natural use of language, there are also some benefits to this design. In particular, it is easy in this setting to control the exact space of questions, goals, and answers. While the restrictions on question space may seem peculiar, it is directly motivated by conversational scenarios in everyday usage which feature restrictions on the set of things one can ask about, due to politeness, salience, time cost, and other factors. In future work, we will explore the extent to which the proposed model can scale up to real-time, multiplayer games, extended dialogues, and other more naturalistic language settings. To deal with dialogues lasting longer than a single exchange, for instance, we must specify the way in which the contributions of questioner and answerer affect the *context* in which later utterances operate.

Humans are experts at inferring the intentions of other agents from their actions (Tomasello, Carpenter, Call, Behne, & Moll, 2005). Given simple motion cues, for example, we are able to reliably discern high-level goals such as chasing, fighting, courting, or playing (Barrett, Todd, Miller, & Blythe, 2005; Heider & Simmel, 1944). Experiments in psycholinguistics have shown that this expertise extends to speech acts. Behind every question lies a goal or intention. This could be an intention to obtain an explicit piece of information ("Where can I get a newspaper?"), signal some common ground ("Did you see the game last night?"), test the answerer's knowledge ("If I add these numbers together, what do I get?"), politely request the audience to take some action ("Could you pass the salt?"), or just to make open-ended small talk ("How was your weekend?"). These wildly different intentions seem to warrant different kinds of answers. By formalizing the computational process by which answerers infer these different intentions, our model framework provides a unifying way to accommodate this diversity.

Acknowledgements

This work was supported by ONR grants N00014-13-1-0788 and N00014-13-1-0287, and a James S. McDonnell Foundation Scholar Award to NDG. RXDH was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Boër, S. E., & Lycan, W. G. (1975). Knowing who. *Philosophical Studies*, *28*(5), 299–344.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430–477.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Groenendijk, J., & Stokhof, M. (1984). On the semantics of questions and the pragmatics of answers. *Varieties of formal semantics*, *3*, 143–170.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Potts, C. (2012). Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics* (pp. 1–20).

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91–136.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*(6), 727–763.

Vogel, A., Bodoia, M., Potts, C., & Jurafsky, D. (2013). Emergence of gricean maxims from multi-agent decision theory. In *Human language technologies: The 2013 annual conference of the north american chapter of the association for computational linguistics* (pp. 1072–1081). Stroudsburg, PA: Association for Computational Linguistics.