Coherence and Rationality in Grounding

Matthew Stone
Computer Science and Cognitive Science
Rutgers University
Piscataway NJ 08854-8019 USA
Matthew.Stone@Rutgers.edu

Alex Lascarides
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB Scotland UK
alex@inf.ed.ac.uk

Abstract

This paper analyses dialogues where understanding and agreement are problematic. We argue that pragmatic theories can account for such dialogues only by models that combine linguistic principles of discourse coherence and cognitive models of practical rationality.

1 Introduction

Interlocutors in conversation have only indirect evidence as to whether others understand and agree with them. Take the joke about the old folks in the bus shelter:

(1) a. *A*: Windy, en'it?
b. *B*: No it's not, it's Thursday.
c. *C*: So am I. Let's go and 'ave a drink!

Evidently *B* mishears *windy* as *Wednesday* and *C* mishears *Thursday* as *thirsty*. Even when somebody says they understand and agree, it's no guarantee that they do.

Our judgements about (1) depend on *principles of discourse coherence*. *B* formulates (1b) as a denial of (1a), so *B* must think that (1b) is semantically incompatible with (1a). Knowing this, *A* can infer information about *B*'s interpretation of (1a). The implicit discourse relation connecting the two halves of (1c) similarly shows that *C* thinks the salient property *B* and *C* share gives them a reason to go have a drink.

But there's more going on. After all, (1) is *not* a coherent discourse. Our judgements about (1) also rely on our knowledge of the kinds of mistakes that people can make in conversation—hearing one word for another, for example—and our presumption that people choose their utterances reasonably to fit the conversation as they understand it. Such inferences represent *cognitive modelling*.

The problem of managing understanding and agreement in conversation is known as grounding (Clark, 1996). In the formal and computational literature, previous approaches to grounding have focused either on discourse coherence or on cognitive modelling, but failed to consider the interactions between the two. In this paper, we outline how the two sets of considerations can be reconciled. We regiment utterance content so that it encapsulates the way dialogue moves are coherently or rhetorically connected to prior utterances. And we apply probabilistic reasoning to assess speakers' rationality in choosing to commit to specific contents. We analyse naturally-occurring examples involving implicit grounding, ambiguous grounding moves, misunderstandings and repair to illustrate the need for both kinds of reasoning.

Our work contributes to three different spheres of investigation. Firstly, it helps to explain how it might be possible for interlocutors to draw precise conclusions about others' mental states, despite the complexity and indirectness of the linguistic evidence. Secondly, it contributes to the Gricean programme of analysing conversation as cooperative activity, by showing how an important and independently-characterised set of conversational inferences might actually be calculated. (Of course, as in (1), these inferences need not always involve Gricean *implicature*.) And finally, we particularly hope that our work will inform the design of more robust and powerful conversational systems, by correlating the architecture, reasoning and knowledge that is realised in these systems with the grounding those systems can do.

2 Examples and Perspective

Following Clark (1996), we view grounding fundamentally as a skill rather than as an epistemic state. Interlocutors achieve grounding when they can detect misunderstanding, clarify utterances, negotiate meaning and coordinate their responses

Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue

in pursuit of successful joint activity. They may or may not thereby achieve *common ground* in the philosophical sense (Stalnaker, 1978). Our view is that the skill of grounding reflects the ability to entertain multiple hypotheses about the organisation of dialogue and to rank these hypotheses quantitatively to make strategic choices. Coherence and rationality are both essential to these calculations.

Dialogue (1) illustrates how principles of discourse coherence contribute to inferences about the nature of an implicit misunderstanding. Lascarides and Asher (2009) use dialogue (2) from Sacks et al. (1974, p.717) to illustrate how principles of coherence can also contribute to inferences about implicit agreement and understanding:

(2) a. *Mark (to Karen and Sharon):*
Karen 'n' I're having a fight,
b. after she went out with Keith and not me.
c. *Karen (to Mark and Sharon):*
Wul Mark, you never asked me out.

Intuitively, Mark and Karen agree that they had a fight, caused by Karen going out with Keith and not Mark. Thus *implicatures can be agreed upon*—that (2b) explains (2a) goes beyond compositional semantics. Furthermore, *agreement can be implicated*—Karen does not repeat (2a), (2b) or utter *OK* to indicate agreement.

As in (1), the basis for recognising Karen's implicit acceptance stems from coherence, which compels us (and Mark) to recognise the rhetorical connection between her contribution and Mark's. Here, the fact that Karen commits to (2c) *explaining why* (2b) is true should be sufficient to recognise that Karen accepts Mark's utterance (2b). Karen's implicit endorsement of (2b) also seems sufficient to conclude that she (implicitly) accepts its *illocutionary effects* as well—(2b) explaining (2a). The fact that Karen chooses to accept these contributions, rather than to ask about them, for example, offers very good evidence that she thinks she understands their content.

Incrementally, as discourse unfolds, interlocutors have only partial information about these contributions. As described by Clark (1996), grounding requires interlocutors to manage uncertainty at four levels: (1) the signals that they exchange with one another; (2) the words that are used; (3) the meanings that those words convey; and (4) what commitments interlocutors make to these meanings. The joke in (1) trades on the difficulty of grounding at Levels 1 and 2. Given the endemic semantic ambiguity, vagueness, and other forms of underspecification associated with utterances, interlocutors frequently also face transient uncertainties about their partners' contributions at Levels 3 and 4.

Interlocutors' choices in conversation reflect the specific ambiguities they encounter and the likelihood they assign to them. For example, when interlocutors see their uncertainty about a prior public commitment, or piece of logical form, as problematic, they can seek clarification, as the sales assistant *B* does in (3b)—a simplified version of a dialogue from the British National Corpus (Burnard, 1995) that is annotated with clarification acts (Purver et al., 2003) (we thank Matthew Purver for pointing us to this example):

(3) a. *A*: I would like one of the small reducers.
b. *B*: One going from big to small
or from small to big?
c. *A*: Big to small.
d. *B*: Big to small, ok.

(3b) is an example where specific clarification is sought on the intended meaning of *small reducers*. In (4), from DeVault and Stone (2007), *B* seeks specific clarification on the *illocutionary* content of *A*'s utterance (4b) rather than its locutionary content: was it an *Acceptance* of (4a) or something else, perhaps merely an *Acknowledgement*?

(4) a. *B*: Add the light blue empty circle please.
b. *A*: okay
c. *B*: Okay, so you've added it?
d. *A*: i have added it.

In both cases, *A*'s response to *B*'s clarification request is designed to help *B* resolve the specific ambiguity that *B* has called attention to.

Of course, as Clark (1996) underscores, not all uncertainty is problematic. If the issue is sufficiently unimportant or a misunderstanding is sufficiently unlikely, interlocutors can choose to tolerate the uncertainty and proceed anyway. This is crucial in systems where modules like speech recognition never offer certainty (Paek and Horvitz, 2000). But it could also be what *B* does in (1) or Karen does in (2), for instance.

So overall, grounding moves and anti-grounding moves can be implicit (see (2) for grounding and (1) for anti-grounding) or explicit (see (3cd) and (4ab) for grounding and (3ab)

and (4bc) for anti-grounding). Moreover, a misunderstanding or lack of grounding can be mutually recognised (see (3) and (4)) or not (see (1)). Even when a grounding move is explicit, there can still be uncertainty about both the level of grounding that the agent has reached—-e.g., *B* is uncertain whether *A*'s explicit endorsement in (4b) marks grounding at Level 3 or grounding at Level 4. There can also be uncertainty about the semantic scope of the endorsement—e.g., an utterance like *I agree* doesn't make explicit whether the acceptance is of all the clauses in the prior turn or only the last clause (see Lascarides and Asher (2009) for discussion).

3 Challenges

Our work draws on previous grounding models based on *discourse coherence* and those based on *probabilistic inferences about strategy*. Both of these traditions provide insights into the data of Sections 1 and 2, but neither tells a complete story.

Coherence approaches start from the insight that the relationships between utterances in discourse give evidence about mutual understanding. An early illustration of this type of reasoning is the work of McRoy and Hirst (1995), who recognise and repair misunderstandings in dialogue by identifying utterances that are best explained by assuming that the speaker's public commitments about the coherent organisation of the discourse are in conflict with those of the addressee.

More recent work in the coherence tradition tends to adopt the influential approach of Traum (1994), who posits specific categories of communicative action in dialogue, called *grounding acts*. The prototypical grounding acts model works by modeling assertions as introducing content with a status of *pending*. Subsequent acknowledgement acts may transfer that content out of what's *pending* and into what's *grounded*. Important work in this tradition includes both theoretical analyses (Poesio and Traum, 1998; Ginzburg, 2010) and system-building efforts (Matheson et al., 2000; Traum and Larsson, 2003; Purver, 2004).

Lascarides and Asher (2009) simplify and extend this idea. They analyse dialogue in terms of a single set of relational speech acts, formalised so as to represent what information each act commits its agent to, implicitly or explicitly. The account predicts facts about implicit grounding, illustrated in dialogue (2), without the need to describe the dialogue in terms of a separate layer of inferred grounding acts. We build on their account here.

Such models are good at characterising agreement but not as good at characterising uncertainty or misunderstanding. For example, in cases where interlocutors proceed despite uncertainty, neither a *pending* status nor a *grounded* status seems appropriate. On the one hand, interlocutors accept that there may be errors; on the other, they act as though the likely interpretation was correct. Such models are also limited by their *symbolically-defined* dynamics. Misunderstandings like those in (1) surface in the dialogue as inconsistencies that can potentially be corrected in a vast number of alternative ways—some of which are intuitively likely, others of which are not. Symbolic models need rules to specify which hypotheses are worth exploring—an open problem—while probabilistic models naturally assign each one a posterior probability based on all the available information.

Probabilistic approaches to grounding were inaugurated by Paek and Horvitz (2000), who describe the decision-theoretic choices of a spoken language interface directly in terms of Clark's model of contributing to conversation. Paek and Horvitz characterise their system's information state in terms of the probabilistic evidence it has about the real-world goals that users are trying to achieve with the system. This evidence includes the system's prior expectations about user behaviour, as well as the system's interpretations of user utterances. Paek and Horvitz show that this representation is expressive enough for the system to assess conflicting evidence about user intent, to ask targeted clarification questions, and to adopt an appropriate grounding criterion in trading off whether to seek more information or to act in pursuit of users' likely domain goals.

A range of related research has exploited probabilistic models in dialogue systems (Walker, 2000; Roy et al., 2000; Singh et al., 2002; Bohus and Rudnicky, 2006; DeVault and Stone, 2007; Williams and Young, 2007; Henderson et al., 2008). However, this research continues to focus primarily on inference about user goals, while largely sidestepping the knowledge and inference required to relate utterances to discourse context, as illustrated in dialogue (2). Moreover, because this work is generally carried out in the setting of spoken dialogue systems, researchers usually formalise whether the system understands the user, but draw no inferences about whether the user un-

derstands the system.

The present paper aims to reconcile these two perspectives in a common theoretical framework.

4 Public Commitments

We adopt from Lascarides and Asher (2009) a representation of the logical form (LF) of coherent dialogue.[1] This LF records the content to which each speaker is publicly committed through their contributions to the dialogue. Commitments are relational. Each utterance typically commits its speaker not only to new content, but also to a specific implied connection to prior discourse (maybe an utterance by another speaker), and perhaps indirectly to earlier content as well. The inventory of these *rhetorical relations* maps out the coherent ways dialogue can evolve—examples include *Explanation*, *Narration*, *Answer*, *Acknowledgement* and many others (Lascarides and Asher, 2009). Pragmatic rules for reconstructing implied relations provide a defeasible mechanism for resolving ambiguity and calculating implicatures.

More formally, the LF of a dialogue in Dialogue SDRT (DSDRT) is the LF of each of its turns, where each turn maps each dialogue agent to a Segmented Discourse Representation Structure (SDRS) specifying all his current public commitments. An SDRS is a set of labels (think of labels as naming dialogue segments) and a mapping from those labels to a representation of their content. Because content includes rhetorical relations $R(a,b)$ over labels a and b, this creates a hierarchical structure of dialogue segments. SDRSs are well-formed only if its set of labels has a unique root label—in other words, an SDRS represents just one extended dialogue segment consisting of rhetorically connected sub-segments.

Abstracting for now away from uncertainty, Lascarides and Asher (2009) suggest that by the end of dialogue (2) Mark and Karen are respectively committed to the contents of dialogue segments π_{1M} and π_{2K}, as shown in (2′) (contents of the 'minimal' segments a, b and c are omitted for reasons of space; we label the public commitments of speaker s in turn t with segment π_{ts}):

(2′) Mark: $\pi_{1M} : \mathit{Explanation}(a,b)$
Karen: $\pi_{2K} : \mathit{Explanation}(a,b) \wedge \mathit{Explanation}(b,c)$

[1] LF is a public construct like a game board, not a subjective construct related to mental state. Though controversial, this view is defensible—and it makes probabilistic modeling a lot easier (DeVault and Stone, 2006).

Karen's and Mark's public commitments share labels a and b. This reflects the reality that an agent's dialogue move relates in a coherent way to prior contributions. Assuming that agreement (or grounding at Level 4) is shared public commitment, LF (2′) entails that Mark and Karen agree that (2b) caused (2a). Lascarides and Asher (2009) infer that $\mathit{Explanation}(a,b)$ is a part of Karen's commitment, given her commitment to $\mathit{Explanation}(b,c)$, via default principles that predict or constrain the semantic scope of implicit and explicit endorsements and challenges. The relevant default principle here is that an implicit endorsement of a prior utterance normally involves acceptance of its illocutionary effects as well.

5 Strategy and Uncertainty

The assumption that interlocutors are pursuing reasonable strategies for pushing the conversation forward, given their information state, often allows observers to draw powerful inferences about what that information state is. For example, suppose Mark understands Karen's move correctly in (2), and thus assigns a high probability to the representation of Karen's commitments that we have ascribed in (2′). Mark can reason that since Karen has accepted the meaning that he intended to convey, then she must have understood it. Thus, implicit agreement—and even disagreement, as in (1)—should make it possible to draw conclusions about (implicit) grounding at lower levels.

In other cases, the representation of an agent A's public commitments may feature a segment a uttered by a prior agent B, and yet by the dynamic interpretation of A's SDRS A is not committed to a's content or its negation. In such cases, observers may not be able to tell whether A has identified the content associated with the earlier utterance a. In other words, the LF reveals a *lack of grounding*. For instance, A's public commitments may include a relation $CR(a,b)$ (*CR* for Clarification Request), whose semantics entails that b is associated with a question K_b all of whose possible answers help to resolve the meaning associated with utterance a. Normally, A would make such a move only when A was uncertain about that content. Seeing a *CR* thus allows interlocutors to infer a lack of grounding at Level 3. Some clarificatory utterances, such as *echo questions* or *fragment reprises* have additional constraints on their use, which reveals even more about what an interlocutor did or did not

recognise at Level 1 or Level 2. See Purver (2004) and Ginzburg (2010) for more formal details about clarification requests and their semantics.

In our view, these inferences are ultimately about what it's rational for a speaker to do. People tend to avoid agreeing with something they know they don't understand, or asking about something they know they do. Doing so doesn't move the conversation forward. In other words, these inferences rest on principles of *cognitive modelling* which are different from, and complementary to, the principles of interpretation which characterise the possible logical form of discourse.

In Figure 1, we schematise our approach to these inferences qualitatively in a dynamic Bayesian network (DBN). The model describes the discourse context as a public scoreboard that evolves, step by step, as a consequence of the moves interlocutors make to update it. M_t is the move made at time t. We think of it as a relational speech act; that is, as a bit of logical form with an intended rhetorical connection to the discourse context. In other words, M_t completely resolves anaphoric reference, discourse attachment, and the propositions expressed. The move for (2c), for example, would include the rhetorical connections $Explanation(a,b) \wedge Explanation(b,c)$ and the content of segment c. X_t is the discourse context at time t. We assume that it is a DSDRS, as illustrated in (2′). Finally, E_t is the observable utterance associated with the update at time t. Depending on the modality of conversation, this might be typed text, acoustic form, or the observable correlates of a multimodal communicative act.

The relevant dynamics involve two ingredients. A model of discourse coherence and discourse update, expressed as $X_{t+1} = u(X_t, M_t)$, describes how moves update the current context to yield a new context. This is the familiar update of dynamic semantics—when X_t and M_t are compatible, X_{t+1} is a new context that takes the information from both into account; otherwise, in cases of incoherence, presupposition failure and the like, X_{t+1} is a defective context that specifies the attempts made and the fact that they failed. A model of language, expressed as $P(E_t|X_t)$, describes the relationship between utterance form and meaning; uncertainties here reflect the variance in the way an utterance may be performed and observed.

The cognitive model surfaces in Figure 1 through models of discourse interpretation and discourse planning. The DBN casts the conversation as involving alternating contributions from two interlocutors A and B. We use the variables A_t and B_t to represent the subjective information state of these agents at time t. The models of discourse interpretation yield updates in the interlocutors' mental states as a function of their observations of an utterance produced by their partner. They are formalised as relationships $P(A_{t+1}|A_t, E_t)$ when t is even and $P(B_{t+1}|B_t, E_t)$ when t is odd. The models of discourse planning, meanwhile, describe the moves interlocutors make as a function of their current information state, and who takes the turn to speak. We have $P(M_{t+1}|A_t)$ when t is even and $P(M_{t+1}|B_t)$ when t is odd. Discourse coherence takes on new force in these planning models. Rational agents strive to make coherent moves, and thereby to commit to certain propositions that match their beliefs and interests.

The network as a whole is analogous to a Hidden Markov Model, with the observable state given by a sequence of utterances E_1 through E_n, and the hidden state at each time given by the joint distribution over X_t, A_t and B_t. A probabilistic *observer* of a conversation reasons in this network by observing the utterance sequence and reasoning about the hidden variables. That's the position we're in when we read an example like (1). The posterior distribution over the hidden state would normally permit specific conclusions about X_t. If the model predicts that A follows this aspect of the dialogue state, the model derives a match between A_t and X_t. If the model predicts A doesn't follow, the model would associate A_t with a value or values that don't match X_t. The model can make the same predictions even if it cannot pin down X_t. This situation would be realised by a broader posterior distribution over X_t and by correlations in the joint distribution over A_t and X_t.

The model would be used differently to implement a *participant* in a conversation. A participant in a conversation doesn't need to draw inferences about their own mental state; they actually implement particular interpretation and planning procedures. These procedures, however, would have a rational basis in the probabilities of the model, if the agent takes not only E_t but also the values for their own moves M_t as observed, and uses the model only to draw inferences about their partner.

This point bears on the nature of models such as $P(A_{t+1}|A_t, E_t)$, and $P(M_t|A_t)$. They can implicitly encode arbitrarily complex reasoning. Thus, Figure 1 is best thought of as shorthand for a *net-*

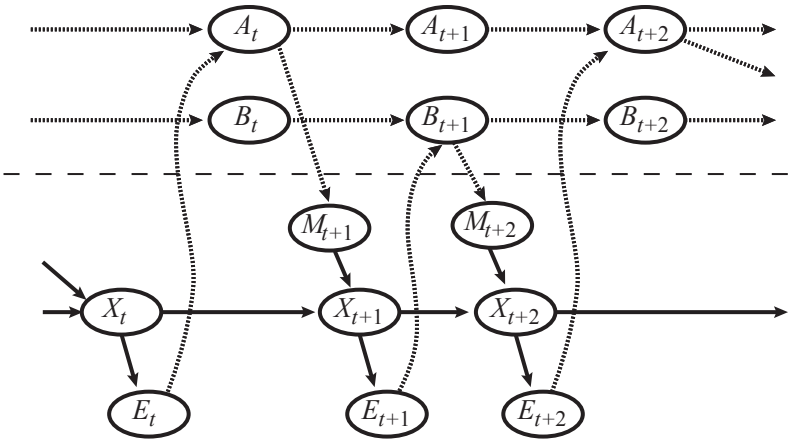

Figure 1: Fragments of the DBN indicating the probabilistic relationships relating one interlocutor A's mental state, the other interlocutor B's mental state, interlocutors' alternating discourse moves M, the evolving discourse context X and the observable correlates of discourse update E (including utterances). Solid dependencies indicate linguistic models; dotted ones, cognitive models.

work of influence diagrams (Gal, 2006). For example, suppose A tracks the hidden dynamics of the conversational record by Bayesian inference. Then A's information state at each time t includes expectations about the current discourse context X_t given the evidence A has accumulated so far in the discourse O_{At} (a combination of observed utterances and planned moves)—this serves as a prior distribution $P_A(X_t|O_{At})$ that's part of A's information state. A also has discourse expectations $P_A(X_{t+1}|X_t)$ and a linguistic model $P_A(E_{t+1}|X_{t+1})$. To describe the discourse interpretation of our Bayesian agent we use a standard DBN definition of filtering, as in (5).

(5) $P_A(X_{t+1}|O_{At+1}) \propto$
$\sum_{X_t} P_A(E_{t+1}|X_{t+1})P_A(X_{t+1}|X_t)P_A(X_t|O_{At})$.

This posterior distribution describes A's state at time $t+1$. This more specific model lets us flesh out how A acts to achieve coherence in planning M_{t+2}. For example, if $P_A(X_{t+1}|O_{At+1})$ assigns a high value to DSDRS K, then $P(M_{t+2}|A_{t+1})$ will be low when $u(K,M_{t+2})$ is incoherent.

Similarly, A may have a substantive model of B's planning and interpretation. Then A's information state will involve a distribution $P_A(B_t)$ over A's model of B, and A will have expectations of the form $P_A(B_{t+1}|E_{t+1},B_t)$ and $P_A(M_{t+1}|B_{t+1})$ describing B's interpretation and planning. These models now underwrite A's discourse expectations, allowing $P_A(X_{t+1}|X_t)$ to be described in terms of $P_A(M_{t+1}|B_t)$. It could be that A has a very simple model of B. Maybe A assumes B understands perfectly, or guesses interpretations at random. However, as familiar game-theoretic considerations remind us, A might instead model B as another Bayesian reasoner. That model may even describe B via a nested model of A! A useful assumption is that agents are uncertain about the exact degree of sophistication of their partner, but assume it is low (Camerer et al., 2004).

6 Worked Examples

We use the examples of Sections 1 and 2 to illustrate the dimensions of variation which our model affords. To make the discussion concrete, we will consider the reasoning of one interlocutor, typically A, using a probabilistic model of the form illustrated in Figure 1. Thus the whole of Figure 1 is understood to encode A's knowledge, with suitable variables (e.g., A_t, E_t and M_{t+1}) observed and the joint distribution over the other variables inferred. We are interested in cases where A speaks, and then A retrospectively assesses B's interpretation of what A has said in light of B's response. We will not assume that A maintains a detailed model of B's planning process. However, we assume that A tracks B's probabilistic representation of the discourse context, and moreover that A and B apply a common, public model of discourse update $u(X_t,M_{t+1})$ and of linguistic expression $P(E_t|X_t)$. For simplicity in treating the examples, we also assume that there are no pending ambiguities in the initial context, so that effectively X_0 and B_0 are observed (by both interlocutors). Obviously, this assumption does not hold in general in the model.

A's inference involves three mathematical constructs. The first is A's assessment of B's interpretation of an initial move M_1 made by A. A is uncertain about the probability B assigns to particular interpretations of the discourse up to time 1, given B's available evidence. This means the model has a continuous random variable z_i for each candidate DSDRS representation K_i for the discourse; z_i gives the probability that B thinks the interpretation of the discourse up to time 1 is K_i. If we take A and B to entertain N interpretations, $\vec{z}$ is a vector in N-dimensional space, subject to the constraint that coordinates sum to one (a point on the $N-1$ simplex). B's state at time 1 thus includes a vector $\vec{z}$ and the model includes a prior probability density over this vector, conditional on available evidence: $p(\vec{z}|X_1, B_0)$ which we represent mathematically as $pr(\vec{z})$. Assuming Bayesian inference by B, it is derived from B_0 via (5) by marginalising in expectation over E_1.

The second key construct describes A's expectations about what B will do next. This is realised in the model's value for the likelihood $P(M_2|B_1)$. Concretely, for each epistemic state $\vec{z}$ for B, the model assigns a likelihood $l(U_j|\vec{z})$ that B chooses move U_j in $\vec{z}$. For each epistemic state $\vec{z}$, the function $l(U|\vec{z})$ defines a point in the $D-1$ simplex, if there are D possible next moves, representing the model's expectation about B's behaviour there.

The final key construct is the model's retrospective assessment of what B's mental state must have been, given the move observed at E_2. That is the posterior $p(\vec{z}|X_1, B_0, E_2)$. We abbreviate this as $po(\vec{z})$; it is another density over the $N-1$ simplex. The model derives this by Bayesian inference:

$$(6)\ \ po(\vec{z}) \propto pr(\vec{z}) \sum_{U_j} l(U_j|\vec{z}) P(E_2|u(X_1, U_j))$$

The equation shows how an interlocutor gets retrospective insight into their partner's mental state by combining evidence from the observed utterance E_2 and discourse coherence, with inference about why the interlocutor might have planned such an utterance, $l(U_j|\vec{z})$, and expectations about what their mental state would have been, $pr(\vec{z})$.

Let's look at (1b). We track B's interpretation via a DSDRS K_1 saying it's windy and another K_2 saying it's Wednesday. We expect understanding, so our prior $pr(\vec{z})$ naturally favors $\vec{z}$ where K_1 has high probability. Now, given the utterance, we can assign high probability to an observed value U_2 for the variable M_2 with the form $Correction(a,b)$ where a is the immediately prior discourse segment and b says it's Thursday. Our posterior distribution $po(\vec{z})$ factors in our prior estimate of B's state, this evidence, and our model $l(U_2|\vec{z})$ of B's choice. Now $u(K_1, U_2)$ is incoherent and $u(K_2, U_2)$ is coherent, so $l(U_2|\vec{z})$ is going to be very low if $\vec{z}$ assigns much probability to K_1. That's how the model recognises the misunderstanding.

Now let's look at (2b). The model of language should predict that any value U_j for M_2 that features $Explanation(b,c)$ as a part will be more likely than an alternative that doesn't. This entails at least a partial commitment to the prior utterance. If we assume that agents tend not to commit in uncertain states—giving a low probability to $l(U_j|\vec{z})$ for such U_j when $\vec{z}$ has high entropy—then (6) sharpens our information about $\vec{z}$. Following Lascarides and Asher (2009), we assume a further constraint on dialogue policy: if you only partially endorse the prior discourse, you tend to say so. So the model predicts further probabilistic disambiguation: among those U_j that feature $Explanation(b,c)$, those that also feature $Explanation(a,b)$ will get a higher posterior probability than those that do not.

Next is (3b). Here the model of language should predict that any likely value U_j for M_2 will feature $CR(a,b)$. Rationality dictates for such U_j that $l(U_j|\vec{z})$ will be high only when $\vec{z}$ has high entropy, and this is reflected in our updated posterior over $\vec{z}$. Indeed, since the linguistic form of the clarification elicits particular information about the prior context, we can use similar reasoning to recognise particular points of likely uncertainty in $\vec{z}$. A similar analysis applies to (4).

7 Discussion and Conclusion

We have proposed a programmatic Bayesian model of dialogue that interfaces linguistic knowledge, principles of discourse coherence and principles of practical rationality. New synergies among these principles, we have argued, can lead naturally to more sophisticated capabilities for recognising and negotiating problematic interactions.

Of course, we must still specify a model in detail. We hope to streamline this open-ended effort by capturing important correlations in dialogue, as found in alignment phenomena for example, through simple generative mechanisms proposed in Pickering and Garrod (2004). We also face difficult computational challenges in fitting our models

to available data and drawing conclusions quickly and accurately from them. We would also like to determine whether existing Bayesian approaches to unsupervised learning, such as Goldwater and Griffiths (2007), can apply to our model. At any rate, until we can demonstrate our ideas through systematic implementation, training and evaluation, our account must remain preliminary.

Acknowledgements

Thanks to Mark Johnson, Matthew Purver, David Schlangen, Chung-chieh Shan and Mike Wunder for helpful discussion. Supported in part by NSF CCF 0541185 and HSD 0624191.

References

D. Bohus and A. Rudnicky. 2006. A K hypothesis + other belief updating model. In *Proceedings of the AAAI Workshop on Statistical and Emprical Approaches for Spoken Dialogue Systems*.

L. Burnard, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

CF. Camerer, T. Ho, and J. Chong. 2004. A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119:861–898.

H. Clark. 1996. *Using Language*. Cambridge University Press.

D. DeVault and M. Stone. 2006. Scorekeeping in an uncertain language game. In *Proceedings of SEMDIAL (BRANDIAL)*, pages 139–146.

D. DeVault and M. Stone. 2007. Managing ambiguities across utterances in dialogue. In *Proceedings of SEMDIAL (DECALOG)*.

Y. Gal. 2006. *Reasoning about Rationality and Beliefs*. Ph.D. thesis, Harvard.

J. Ginzburg. 2010. *The Interactive Stance: Meaning for Conversation*. CSLI Publications.

S. Goldwater and T. Griffiths. 2007. A Fully Bayesian Approach to Unsupervised POS Tagging. *Proceedings of ACL*, pages 744-751.

J. Henderson, P. Merlo, G. Musillo, and I. Titov. 2008. A latent variable model of syncrhonous parsing for syntactic and semantic dependencies. In *Proceedings of CoNLL*, pages 178–182.

A. Lascarides and N. Asher. 2009. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

C. Matheson, M. Poesio, and D. Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL*, pages 2–9.

S. McRoy and G. Hirst. 1995. The repair of speech act misunderstandings by abductive inference. *Computational Linguistics*, 21(4):435–478.

T. Paek and E. Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of UAI*, pages 455–464.

MJ. Pickering and S. Garrod. 2004. Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225.

M. Poesio and D. Traum. 1998. Towards an axiomatisation of dialogue acts. In *Proceedings of SEMDIAL (TWENDIAL)*, pages 207–222.

M. Purver, J. Ginzburg, and P. Healey. 2003. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 235–255. Kluwer Academic Publishers.

M. Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, London.

N. Roy, J. Peneau, and S. Thrun. 2000. Spoken dialogue management for robots. In *Proceedings of ACL*, pages 93–100, Hong Kong.

H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turntaking in conversation. *Language*, 50(4):696–735.

SP. Singh, DJ. Litman, MJ. Kearns, and MA. Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16:105–133.

R. Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics*, pages 315–322. Academic Press.

D. Traum and S. Larsson. 2003. The information state approach to dialogue management. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer Academic Publishers.

D. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

MA. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artifical Intelligence Resesarch*, 12:387–416.

JD Williams and SJ Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):231–422.