Interruptions as Speech Acts

Peter Wallis and Bruce Edmonds

Centre for Policy Modelling
Manchester Metropolitan University
pwallis@acm.org

Abstract. This paper introduces a model of human communication in which 'accounting-for' is the basis of meaning, and argues that interruptions should be handled in the same way as any other speech act. The model has at its core the idea that human languages are inherently *intentional* – we focus on our conversational partner's goals – and that what is needed is *mixed initiative at the level of intent.* It would seem interruptions can reaffirm or contradict the speaker's current intent and the paper finishes with a description of our (very) shallow approach to intention recognition.

1 Introduction

A speech interface for robots and computers has been part of the AI vision from the very beginning but after 50 years of trying it turns out that talking, like walking, is far more complex than the early visionaries anticipated. Over the last few years, large corporations with commensurate research budgets have decided the technology has come of age, but their approach to dialog management however is hardly more sophisticated than chat-bots. These corporations certainly employ the best and brightest but following perhaps the Microsoft foray with the Desktop Assistant (the Paper-Clip) commercial ventures in the area tend to be conservative. Google Assistant and Siri basically perform internet searches and Amazon Echo is primarily a home automation controller. Indeed it is interesting to note that, when the Amazon Echo needs to interrupt the user, it does it via the well understood mechanism of setting an alarm. Technically Alexa could say "Excuse me it has been 10 minutes" but instead the Echo plays a ring tone. The Echo does this because, as we discovered in 2008 when we put robot rabbits in older peoples homes, it is surprisingly hard to get a machine performing as a social actor to initiate a conversation [17]. It seems we modern humans are conditioned to not get overly annoyed when a machine interrupts us with an alarm. People (or machines) talking to us is another matter.

It turns out this kind of conditioning is endemic in the way we communicate. In order to chart a course through this complex network of norms and social relations we have turned to a suite of techniques from the human sciences broadly under the banner of Conversation Analysis[7, 5]. This approach not only provides explanations of what goes wrong in human-machine conversations [16] but also provides (unbelievably) good quantitative results [18]. Critically, it also provides a model that can be implemented.

2 Meaning and action

The conventional wisdom is that natural languages are the definitive symbol system and computers are, in a very literal sense, *universal* symbol manipulation systems. So what could we be possibly missing when it comes to conversational machines? The answer has been of course the notion of agency and situated action. Computers can do something other than manipulate symbols; they can implement arbitrary causal relations between sensors and actuators. Computers might compute anything that is computable, but they can also implement the decision processes of a thermostat. Behaviour-based robotics [1] has certainly made some significant progress over good old fashioned AI (GOFAI) systems that sense, model, plan and act. Applied to language understanding – in particular dialog – the success of situated action suggests that we take seriously Austin's notion of language as action [2]. Austin and Searle has been championed before but the point often gets morphed into something about the action being to *inform* and we are back to all the issues with the conduit metaphor [11]. Conversation Analysis [7, 5] is a qualitative approach which looks at the "work done" by a speaker when making an utterance. Rather than looking in heads for meaning, we need to look at the relationship between the head and the world around it.

As an example of the phenomena of interest in CA, consider this classic example from the literature in a doctor/patient conversation:

Patient: *this treatment, it won't have any effect on us having kids will it?*

Doctor: [3.2 seconds silence]

Patient: *It will?*

Doctor: I'm afraid so..

Although it might seem reasonable to consider words to have meaning that can be looked up independent of context, the same is not true of silence and in the example that is certainly a meaningful silence. Whatever mechanism is at work here, it is *also* hard at work when we figure out the meaning of words.

Looking through the CA literature, human communication is is full of normative behaviours – rules that can be broken, but to break them will be interpreted. These rules are "behind the scenes" in that we do not consciously think about them but we know they are there and shared by our "community of practice". Making an apology is a complex process [10], but then so is saying goodbye [12].

The idea that language use requires folk knowledge may be obvious but the extent to which folk knowledge is core is perhaps borne out by the success amateurs have in developing conversational agents for things like the Loebner Prize [8]. Folk know exactly, in context, what to say. What the untrained do not know is how to abstract from the surface form of an actual apology say, to something that can generalize across different contexts. Indeed fifty years of NLP research suggests experts do not know how to do that abstraction either.

2.1 How language works

In order to systematically analyse such phenomena CA has disavowed conjecture about the mechanism or “rules in the head” that might have general applicability. Instead the focus on what happens in particular instances of communication and what observable behaviour contributes to choices made. The scientific knowledge is in how to study “folk” knowledge and, as with other ethnomethods, the point is to capture the everyday knowledge that people use to do what they do. Conversation Analysis provides a methodology but, having collected our butterflies, engineers need generalizations in order to make something that can hold a conversation. CA is strong on methodology and shy on theory, but Seedhouse [13] gives a summary of “the findings of CA over the last 50 years” providing an implementable generalization of how language works. To summarize, a speaker’s utterance will fall into one of the following categories:

Seen but unnoticed An utterance will go seen but unnoticed if it is the answer to the conversational partner’s (CP’s) question, a greeting in response to a greeting, and so on. If the speaker produces the second part of an “adjacency pair,” then the CP (who produced the first part of the pair) will not “notice” the utterance but will take in this expected response and move on.

Noticed and accounted for If the speaker says an utterance that is not expected by his or her CP — not the normal response — then the CP does not instantly give up, but actually works hard to figure out why the speaker said what was said. As a classic example consider some one walking in to a corner shop:

A: Hello. Do you sell stamps?
B: First class or second class?

Unless it is pointed out, people often do not notice that B’s response is not an answer to the question. B’s response can however be accounted for.

Risks sanction If the utterance makes no sense, and the CP cannot figure out how it relates to what went before, then the CP will start working toward sanctioning the CP. It seems humans have a notion of fairness and feel justified in sanctioning the speaker if they think the speaker is not cooperating in the communication process. This is not a prescriptive rule taught to well brought up children; it is descriptive of what people do. The form of the sanction depends on many things and is highly culturally dependent. This is where notions of power and distance, roles and expectations comes into play.

This process of working through the seen-but-unnoticed, noticed-and-accounted-for, to sanction is not something we think about; it is just what we do and as such it is hard to notice in action. For instance it is quite surprising just how hard we are willing to work at accounting for a speaker’s utterance. What is more, and in contrast to the views of many CA practitioners, it requires thinking about intentions.

2.2 Language as Intentional and Cooperative

We humans read off the intent of others – it may be difficult to recognize the intent of Hitler invading Poland, but seeing two children tugging at a teddy bear the human observer will be quite sure they both *want* it [3]. And we are willing to work hard at it. Consider this (naturally occurring) example from Eggins and Slade [4] talking about sequential relevance:

> **A:** What's that floating in the wine?
> **B:** There aren't any other solutions.
> You will try very hard to find a way of interpreting B's turn as somehow an answer to A's question, even though there is no obvious link between them, apart from their appearance in sequence. Perhaps you will have decided that B took a common solution to a resistant wine cork and poked it through into the bottle, and it was floating in the wine. Whatever explanation you came up with, it is unlikely that you looked at the example and simply said 'it doesn't make sense', so strong is the implication that adjacent turns relate to each other.

This example demonstrates just how hard human communicators are willing to work at *recognizing intent* in the speech acts of others. By contrast consider the example in Figure 1 from Mann [9] which is a good example of just how easy we humans find it to recognize intent. In this conversation it is not until line 11 that Mother uses any words or concepts that might be semantically linked with the Child's opening statement, and indeed the words she speaks are from an

1 Child: I'm hungry
2 Mother: Did you do a good job of your geography homework?
3 Child: Yeah.
4 Child: What's to eat?
5 Mother: Let me read it.
6 Mother: What is the capital of Brazil?
7 Child: Rio de Janeiro.
8 Mother: Think about it.
9 Child: It's Brasilia.
10 Child: Can I eat now?
11 Mother: I'll let you have something later.
12 Mother: What is the capital of Venezuela?
13 Child: Caracas.
14 Mother: Fine.
15 Child: So what can I eat?
16 Mother: You want some cereal?
17 Child: Sure.

Fig. 1. Mann [9] gives this example when introducing his idea of dialog games.

entirely different context. Of course the child knows what mum *wants* and it is only because the child knows what mother wants that her utterances make any sense as part of this conversation. If one is not convinced that mother's desires are a factor, consider a conversation in which at line 2 Mother says "I rode my new bike thirty five kilometres today". Being a member of the community of practice one can put oneself in the shoes of the child and imagine one's confusion. One may indeed "try very hard to find a way of interpreting" riding a bike as relevant to feeding a child — indeed it is expected — but the point is that one's explanation will involve unravelling Mother's reasoning about bikes and feeding children. That is, one will be trying to understand what Mother *intended* by her utterance about bikes. The great apes, according to Tomasello [14] are perfectly capable of recognizing intent, but they do not have the compunction to cooperate in the communicative process. They would presumably look at the Eggins and Slade example, decide it didn't make sense, and move on.

2.3 Sanction

We humans rarely get to the stage of sanctioning but conversational agents are another matter. Consider the conversation in Figure 2 between a human and a chat bot. On a first pass what stands out is the way the conversation simply falls apart. One can point to the place where things start to go wrong, but for a considerable number of turns, the human at least is working toward repairing the interaction with apologies and warnings. At (1) the machine utters something that the judge cannot account for. The judge tries to get the machine to explain, which fails, resulting in an explicit threat of sanction. In the end the threatened "keywords" include swearing even though the judge is well aware that the world is watching.

PROGRAM: G'day mate.
JUDGE02: G'day. How's it going?
PROGRAM: My mother once said ”How is it going is a mystery.” She was a bit nuts.
JUDGE02: Uh-huh.
PROGRAM: Are you still learning to speak, my friend?
JUDGE02: NO, no. But your anecdote took me a bit by surprise.
JUDGE02: How was your morning?
PROGRAM: Oh well, that's a pity! ← (1)
JUDGE02: I'm sorry; what's a pity?
PROGRAM: Apologies accepted.
JUDGE02: I feel compelled to insult you, and we've only been talking a minute.
PROGRAM: We are bloody well talking, aren't we?
JUDGE02: Oh baby. We sure are. So here are some keywords for you: shit mother baseball.
...

Fig. 2. A conversation with the winning entry for the 1997 Loebner Prize competition.

A standard response to this example is to think we just need to ensure the machine does not say anything that cannot be accounted for. Notice however that four lines prior to (1) the human says something the machine (acts as if it) cannot account for and the human's response is quite different. The events are mirror images of each other but the human's handling of the situation is so automatic for us that it is hard to notice the symmetry. Human language use is situated action in an environment. Getting machines to do conversation, the context is just as much part of the process as the code and that context is full of highly socialized people.

3 Implementation

When someone picks up a phone or runs a screen-based conversational agent, they are *already* attending to (engaged with) the agent. Setting this up with a physical agent is discussed elsewhere [17] but once engaged, a conversational partner (CP) will either treat an utterance as **seen-but-unnoticed**, or will **notice-and-account-for** it, or the CP will **risk-sanction**. To notice-and-account-for requires some form of intention recognition. Intention recognition is an open-ended question but the real question is just how much is needed in order to make conversational machines seem just not very bright as opposed to stupid or offensive. The mechanism we currently use is a variant on plan choice in a classic BDI agent architecture.

3.1 BDI Dialog Management

The Belief, Desire and Intentions (BDI) agent architectures [19] were developed to address the problem of situated action while at the same time maintaining the notion of commitment to a plan. BDI architectures have been used for dialog many times before and the key feature being that this approach provides *mixed initiative at the level of intent.*

Most BDI systems do not do planning but rather manage plans obtained from a fixed plan library. What is more, it is not expected that there is planning "all the way down". Indeed plans in the library may contain sets of chatbot-like pattern-action rules that simply "produce behaviour" that an agent might have when it has the relevant goal. There may be several plans that might achieve a particular goal, and thus plan failure does not necessarily mean goal failure – there is a level of **commitment** to the goal that is not seen in many of the more traditional approaches involving planning. Critically for dialog systems, the goal is explicit and can be used in explanations of behaviour.

We have been developing a dialog scripting language based on XML that uses a combination of features of Voice XML [15] and JAM [6]. The core construct is the `say` element that has as its body the text to say and takes as an argument a (reference to a) grammar to pass to the speech recognition infrastructure. There is also a `plan` element that, for the purposes of this paper might consist of a sequence of `say` and `if` elements. A plan takes the name of a goal as an argument so that, when the goal is `posted` the system (may) form an intent to achieve the goal by executing the (body of the) plan. Figure 3.1 shows two plans, one of which tells a knock-knock joke, and the other of which goes through the process of saying goodbye. Note that telling a knock-knock joke requires more than 2 turns and the process might be interrupted. As such the process is a good example of why dialog is situated action.

```
<plan achieves="tell a joke" >
  <say recognize="whosthere" >
    Knock knock.  </say>
  <say recognize="madamwho">
    Madam </say>
  <say>
    Ma damn foot is stuck.
  </say>
</plan>
```

```
<plan achieves="sayBye" trigger="thanks|bye" >
  <say recognize="thanks|bye" resultId="X">
    Thank you for using this service. </say>
  <if cond="X==’bye’">
    <say> Good bye.</say>
    <success/>
  <elseif cond="X==’thanks’"/>
    <say recognize="bye" resultId="Y">
      Goodbye. </say>
    <if cond="Y==’bye’">
      <success/>
    </if>
  </if>
</plan>
```

Fig. 3. Two plans, one to tell a joke, the other says goodbye.

3.2 A walk-through

Consider the plan to tell a joke. The seen-but-unnoticed is handled by the < *say recognise* = ”..” > construct which says the text and waits for a user response. If the user says something that is not recognised by the currently active say statement, then the first assumption is that the user has changed his or her goal and the system looks through the plan library for a **trigger rule** that matches the input. If one is found, the relevant plan is posted. In theory of course the trigger would create a belief that the CP has a new goal, and the system would reason about the goals it has and possibly choose a new goal in response. This does however seem excessive given the types of dialog we believe we can handle with the trigger mechanism.

Looking at the knock-knock joke example, consider what happens when the goal *tell a joke* is posted. The system finds this plan, or another which also tells a joke, and executes it. Upon completion, if the goal is not removed, the next plan to tell a joke is found and so on. At some point the user will get sick of knock-knock jokes and can quit the program *at any time* by saying ‘good bye’ which matches the trigger grammar of the *sayBye* plan. The trigger mechanism provides an elegant solution to interruptions that does not entail explicitly checking for *no_match* conditions at every step.

4 Conclusion

Accounting-for is a crucial part of how humans communicate. To make machines do this requires some form of intention recognition, and this paper describes out simple approach to accounting-for based on trigger grammars. Using this approach, the Conversational Partner can not only interrupt the system while it is talking, he or she can interrupt the system's current intent. Our implementation is lacking in many, many ways, but the framework captures the idea of language as intentional and cooperative and is a basis for our goal of having machines do really natural language processing.

References

1. Arkin, R.C. (ed.): Behavior-Based Robotics. MIT Press, Cambridge, MA (1998)
2. Austin, J.L.: How to do Things with Words. Clarendon Press, Oxford, UK (1955)
3. Dennett, D.C.: The Intentional Stance. The MIT Press, Cambridge, MA (1987)
4. Eggins, S., Slade, D.: Analysing Casual Conversation. Cassell, Wellington House, 125 Strand, London (1997)
5. ten Have, P.: Doing Conversation Analysis: A Practical Guide (Introducing Qualitative Methods). SAGE Publications (1999)
6. Huber, M.J.: JAM: A BDI-theoretic mobile agent architecture. In: Third International Conference on Autonomous Agents (Agents 99). pp. 236–243. ACM Press (1999)
7. Hutchby, I., Wooffitt, R.: Conversation Analysis: principles, practices, and applications. Polity Press (1998)
8. The Loebner Prize (July 2002), http://www.loebner.net/Prizef/loebner-prize.html
9. Mann, W.C.: Dialogue games: Conventions of human interaction. Argumentation 2, 511–532 (1988)
10. Owen, M.: Apologies and Remedial Interchanges. Mouton Publishers (1983)
11. Reddy, M.J.: The conduit metaphor: A case of frame conflict in our language about language. In: Ortony, A. (ed.) Metaphor and Thought. Cambridge University Press (1993)
12. Schegloff, E.A., Sacks, H.: Opening up closings. Semiotica 8(4) (1973)
13. Seedhouse, P.: The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective. Blackwell (September 2004)
14. Tomasello, M.: Origins of Human Communication. The MIT Press, Cambridge, Massachusetts (2008)
15. The voice xml standard (December 2004), http://www.w3.org/TR/voicexml20/
16. Wallis, P.: Revisiting the DARPA communicator data using Conversation Analysis. Interaction Studies 9(3), 434–457 (October 2008)
17. Wallis, P.: A robot in the kitchen. In: ACL Workshop WS12: Companionable Dialogue Systems. Uppsala (2010)
18. Wallis, P., Crockett, K., Little, C.: When things go wrong. In: Ramchurn, S.D., Fisher, J., Rosenfeld, A., Tran-Thanh, L., Gal, K. (eds.) Human-Agent Interaction Design and Models (HAIDM). Paris (2014)
19. Wooldridge, M.: Reasoning about Rational Agents. The MIT Press, Cambridge, MA (2000)